1

L'apprentissage statistique : pourquoi, comment ?

Introduction

Une des tâches essentielles du cerveau consiste à transformer des informations en connaissances : identifier les lettres qui constituent un texte, les assembler en mots et en phrases, en extraire un sens, sont des activités qui nous paraissent naturelles une fois l'apprentissage nécessaire accompli avec succès.

L'objectif de l'apprentissage statistique est d'imiter, à l'aide d'algorithmes exécutés par des ordinateurs, la capacité qu'ont les êtres vivants à *apprendre par l'exemple*. Ainsi, pour apprendre à un enfant la lecture des lettres ou des chiffres, on lui présente des exemples de ceux-ci, écrits dans des styles et avec des polices différents. On ne fournit généralement pas à l'enfant une description analytique et discursive de la forme et de la topologie des caractères : on se contente de lui montrer des exemples. À la fin de l'apprentissage, on attend de l'enfant qu'il soit capable de lire non seulement tous les chiffres et lettres qui lui ont été présentés durant son apprentissage, mais également tous les chiffres et lettres qu'il est susceptible de rencontrer : en d'autres termes, on attend de lui qu'il ait une capacité de *généralisation* à partir des exemples qui lui ont été présentés. De même, à l'issue de l'apprentissage d'un modèle statistique à partir d'exemples, celui-ci doit être capable de *généraliser*, c'est-à-dire de fournir un résultat correct, dans des situations qu'il n'a pas connues pendant l'apprentissage.

Considérons deux exemples simples de tâches qui peuvent être accomplies par apprentissage artificiel :

- Dans les centres de tri postal, la lecture automatique des codes postaux, et des autres éléments de l'adresse des lettres et paquets, est fréquemment effectuée à l'aide de modèles obtenus par apprentissage statistique, à partir d'exemples de chacune des classes de chiffres. Il s'agit là d'un problème de *classification* : chaque chiffre inconnu doit être attribué à une classe parmi les 10 classes de chiffres possibles (ou être attribué à une classe dite « de rejet » si le chiffre est trop mal écrit pour être reconnu par la machine : l'objet postal doit alors être traité manuellement).
- Dans l'industrie pharmaceutique, on cherche à prédire l'activité thérapeutique d'une molécule à partir de sa structure, avant même de synthétiser cette molécule, afin d'éviter qu'une synthèse coûteuse risque de se révéler finalement inutile. Cette prédiction est fréquemment effectuée par des modèles, construits par apprentissage statistique, à partir de bases de données de molécules dont les activités thérapeutiques sont connues.

Ces deux problèmes, quoique très différents, ont une caractéristique commune essentielle : ils ne peuvent pas être résolus par l'application de connaissances existant a priori. Il n'existe pas d'équation mathématique, issue des connaissances des chimistes et des pharmaciens, qui permette de prédire précisément l'activité d'une molécule connaissant sa structure ; de même, il n'existe pas d'équation qui décrive les propriétés topologiques des chiffres manuscrits. C'est dans de telles conditions que le recours à l'appren-

tissage statistique à partir d'exemples se révèle très fructueux. Nous présenterons bien d'autres exemples d'applications dans ce chapitre et les suivants.

Cet ouvrage présente trois grandes familles de modèles statistiques obtenus par apprentissage artificiel – les *réseaux de neurones*, les *machines à vecteur supports* et les *cartes auto-adaptatives* – qui connaissent un grand succès, depuis plusieurs années ; ils font l'objet de très nombreuses applications.

L'objectif de ce chapitre est de présenter les bases de la conception d'un modèle par apprentissage, de manière aussi intuitive que possible, mais avec la rigueur nécessaire pour une mise en œuvre raisonnable et l'obtention de résultats fiables. On présente tout d'abord un exemple très élémentaire de modélisation par apprentissage, qui montre la dualité entre l'approche algorithmique, traditionnelle en apprentissage, d'une part, et l'approche statistique, qui en est devenue indissociable, d'autre part. La notion fondamentale étant celle de modèle, on présente ensuite quelques définitions qui précisent ce que l'on entend par modèle dans cet ouvrage ; on introduit notamment la distinction entre modèles linéaires et modèles non linéaires en les paramètres, ainsi que la distinction entre modèles statiques et modèles dynamiques. La section suivante décrit deux problèmes académiques d'apprentissage, l'un dans le domaine de la classification, l'autre dans celui de la prédiction ; ces exemples simples permettent de mettre en évidence le dilemme biais-variance, qui constitue un problème central pour la pratique de l'apprentissage statistique. On présente ensuite, de manière plus formelle, les éléments de la théorie de l'apprentissage : fonction de perte, erreur de prédiction théorique, classifieur de Bayes, dilemme biais-variance. Il s'agit là essentiellement de résultats asymptotiques, valables dans l'hypothèse où le nombre d'exemples est infini. La cinquième section est plus proche de la pratique, en ce sens que les résultats qui y sont présentés tiennent compte du fait que les données sont en nombre fini : ce sont les bornes sur l'erreur de prédiction, fournies par la théorie de V. Vapnik. Les quatre sections suivantes sont de nature entièrement pratique : elles exposent les différentes tâches à accomplir pour concevoir un modèle par apprentissage – collecte des données, prétraitements, sélection des variables, apprentissage, sélection de modèles. Ces deux dernières tâches font l'objet de deux sections suivies d'un résumé de la stratégie de conception de modèles. On présente ensuite a conception des modèles les plus simples : les modèles linéaires en leurs paramètres. Enfin, la dernière section du chapitre fournit les éléments de statistiques nécessaires à une bonne compréhension de la mise en œuvre des méthodes décrites tout au long de l'ouvrage.

Premier exemple : un problème élémentaire d'apprentissage statistique

Comme indiqué plus haut, l'objectif de l'apprentissage statistique est de réaliser, à partir d'exemples, un modèle prédictif d'une grandeur numérique, de nature quelconque (physique, chimique, biologique, financière, sociologique, etc.).

La démarche de conception d'un modèle par apprentissage nécessite de postuler une fonction, dont les variables (également appelées *facteurs*) sont susceptibles d'avoir une influence sur la grandeur à modéliser ; on choisit cette fonction parce que l'on pense qu'elle est susceptible

- *d'apprendre* les données existantes, c'est-à-dire de les reproduire le mieux possible,
- *de généraliser*, c'est-à-dire de prédire le comportement de la grandeur à modéliser dans des circonstances qui ne font pas partie des données d'apprentissage.

Cette fonction dépend de *paramètres* ajustables : l'apprentissage artificiel consiste en l'ajustement de ces paramètres de telle manière que le modèle ainsi obtenu présente les qualités requises d'apprentissage et de généralisation.

Dans cet ouvrage, toutes les variables seront regroupées en un vecteur noté $\boldsymbol{x}$, et tous les paramètres en un vecteur noté $\boldsymbol{w}$. Un modèle *statique* sera désigné par $g(\boldsymbol{x}, \boldsymbol{w})$: après apprentissage, c'est-à-dire estimation des paramètres $\boldsymbol{w}$, la valeur que prend la fonction, lorsque les variables prennent un ensemble de valeurs $\boldsymbol{x}$, constitue la prédiction effectuée par le modèle. Les modèles *dynamiques* seront définis dans la section suivante, intitulée « Quelques définitions concernant les modèles ».

À titre d'exemple très simple de modèle statique, supposons que l'on ait effectué N mesures (p_1, p_2, …, p_N) du poids d'un objet, avec des balances et dans des lieux différents. Nous cherchons à estimer le poids de cet objet. Nous observons que les résultats des mesures sont tous à peu près identiques, à des fluctuations près qui peuvent être dues à l'imprécision des mesures, aux réglages différents des balances, ou à des variations locales de l'accélération de la pesanteur. On peut donc supposer raisonnablement que la masse de l'objet est constante ; en conséquence, la première étape de conception d'un modèle prédictif consiste à postuler un modèle de la forme

$$g(\boldsymbol{x}, \boldsymbol{w}) = w,$$

où w est un paramètre constant dont la valeur est l'estimation du poids de l'objet. La deuxième étape consiste à estimer la valeur de w à partir des mesures disponibles : c'est ce qui constitue l'apprentissage proprement dit. Une fois l'apprentissage terminé, le modèle fournit une estimation du poids de l'objet, donc une prédiction du résultat de la mesure de celle-ci, quels que soient la balance utilisée et le lieu de la mesure.

Cet exemple contient donc, sous une forme très simplifiée, les étapes que nous avons décrites plus haut :
- On s'est fixé un objectif : prédire la valeur d'une grandeur ; dans cet exemple très simple, cette valeur est constante, mais, en général, la valeur prédite dépend de variables $\boldsymbol{x}$.
- On a postulé un modèle $g(\boldsymbol{x}, \boldsymbol{w})$, où $\boldsymbol{x}$ est le vecteur des variables du modèle, et $\boldsymbol{w}$ est le vecteur des paramètres du modèle ; dans cet exemple, il n'y a pas de variable puisque la grandeur à prédire est constante, et il y a un seul paramètre w. Le modèle postulé est donc simplement la fonction constante $g(\boldsymbol{x}, \boldsymbol{w}) = w$.

Il reste alors à estimer l'unique paramètre du modèle, c'est-à-dire à effectuer l'apprentissage du modèle à partir des données disponibles.

Cet apprentissage peut être considéré sous deux points de vue, qui suggèrent deux méthodes d'estimation différentes ; elles conduisent évidemment au même résultat.

Point de vue algorithmique

Nous cherchons la valeur du paramètre w pour laquelle la prédiction du modèle est aussi proche que possible des mesures. Il faut donc définir une « distance » entre les prédictions et les mesures ; la distance la plus fréquemment utilisée est la *fonction de coût des moindres carrés*

$$J(\boldsymbol{w}) = \sum_{k=1}^{N} \left(p_k - g(\boldsymbol{x}_k, \boldsymbol{w}) \right)^2,$$

c'est-à-dire la somme des carrés des différences entre les prédictions $g(\boldsymbol{x}_k, \boldsymbol{w})$ et les mesures p_k. $\boldsymbol{x}_k$ désigne le vecteur des valeurs que prennent les variables lors de la mesure k. Puisque nous avons postulé un modèle constant, cette fonction de coût s'écrit

$$J(w) = \sum_{k=1}^{N} \left(p_k - w \right)^2.$$

Pour trouver la valeur de w pour laquelle cette fonction est minimale, il suffit d'écrire que sa dérivée est nulle :

$$\frac{dJ(w)}{dw} = 0,$$

ce qui donne :

$$w = \frac{1}{N}\sum_{k=1}^{N} p_k.$$

Le meilleur modèle prédictif, au sens de la « distance » des moindres carrés que nous avons choisie, et compte tenu des données dont nous disposons, sous l'hypothèse que la masse de l'objet est constante, est donc

$$g(\boldsymbol{x},\boldsymbol{w}) = \frac{1}{N}\sum_{k=1}^{N} p_k.$$

Le poids prédit est donc simplement la moyenne des poids mesurés.

Point de vue statistique

Prenons à présent le problème sous l'angle des statistiques. Puisque l'on a de bonnes raisons de penser que le poids p_0 de cet objet est constant, il est naturel, d'un point de vue statistique, de modéliser les résultats de ses mesures comme des réalisations d'une variable aléatoire *P*. Celle-ci est la somme d'une variable aléatoire certaine P_0, d'espérance mathématique p_0, et d'une variable aléatoire *B*, d'espérance mathématique nulle (le lecteur qui n'est pas familier avec ces notions en trouvera les définitions dans la dernière section de ce chapitre) :

$$P = P_0 + B$$

de sorte que l'on a :

$$E_P = p_0$$

où E_P désigne l'espérance mathématique de la variable aléatoire *P*.

La variable aléatoire *B* modélise l'ensemble des perturbations et bruits de mesure. Le « vrai » poids (inconnu) de l'objet étant p_0, l'apprentissage a donc pour objectif de trouver une valeur du paramètre *w* qui soit aussi proche que possible de p_0. Dans cet exemple, l'objectif de l'apprentissage est donc d'estimer l'espérance mathématique de la variable aléatoire *P* connaissant des réalisations p_k ($k = 1$ à N) de celle-ci. Or la moyenne est un estimateur non biaisé de l'espérance mathématique, c'est-à-dire qu'elle tend vers p_0 lorsque le nombre de mesures tend vers l'infini (ce résultat est démontré dans la dernière section de ce chapitre, intitulée « Éléments de statistiques »). La meilleure estimation de p_0 que nous puissions obtenir, à partir des données disponibles, est donc la moyenne des mesures :

$$\frac{1}{N}\sum_{k=1}^{N} p_k.$$

On retrouve donc le modèle prédictif obtenu par l'approche algorithmique : $g(\boldsymbol{x},\boldsymbol{w}) = \frac{1}{N}\sum_{k=1}^{N} p_k$.

Ayant ainsi déterminé le modèle par apprentissage, il est très important d'estimer la confiance que l'on peut avoir en cette prédiction : pour cela, on calcule un *intervalle de confiance* sur la prédiction fournie.

Le calcul de l'intervalle de confiance sur la moyenne d'observations est décrit dans la dernière section de ce chapitre.

Ces deux points de vue, algorithmique et statistique, ont longtemps été séparés. Les tout premiers développements de la théorie de l'apprentissage, apparus dans les années 1980, étaient essentiellement inspirés par le point de vue algorithmique, ce qui n'intéressait guère les statisticiens. Ce n'est que dans les années 1990 qu'une véritable synergie s'est créée entre les deux approches, permettant le développement de méthodologies efficaces et fiables pour la conception de modèles par apprentissage.

Quelques définitions concernant les modèles

Dans tout cet ouvrage, on désignera sous le terme de *modèle* une équation paramétrée (ou un ensemble d'équations paramétrées) permettant de calculer la valeur de la grandeur (ou des grandeurs) à modéliser à partir des valeurs d'autres grandeurs appelées *variables* ou *facteurs*. On distinguera les *modèles statiques* des *modèles dynamiques*, et les *modèles linéaires en leurs paramètres* des *modèles non linéaires en leurs paramètres*.

Modèles statiques

Un modèle statique est une fonction paramétrée notée $g(\boldsymbol{x},\boldsymbol{w})$, où $\boldsymbol{x}$ est le vecteur dont les composantes sont les valeurs des variables, et où $\boldsymbol{w}$ est le vecteur des paramètres du modèle.

Modèles statiques linéaires en leurs paramètres

Un modèle statique est linéaire en ses paramètres s'il est une combinaison linéaire de fonctions non paramétrées des variables ; il est de la forme

$$g(\boldsymbol{x},\boldsymbol{w}) = \sum_{i=1}^{p} w_i f_i(\boldsymbol{x}),$$

où f_i est une fonction connue, non paramétrée, ou à paramètres connus. Ce modèle peut encore s'écrire sous la forme d'un produit scalaire :

$$g(\boldsymbol{x},\boldsymbol{w}) = \boldsymbol{w} \cdot \boldsymbol{f}(\boldsymbol{x}),$$

où $\boldsymbol{f}(\boldsymbol{x})$ est le vecteur dont les composantes sont les fonctions $f_i(\boldsymbol{x})$.

Les polynômes, par exemple, sont des modèles linéaires en leurs paramètres : les fonctions $f_i(\boldsymbol{x})$ sont les monômes des variables $\boldsymbol{x}$. Les polynômes sont néanmoins non linéaires en leurs variables.

On appelle *modèle linéaire* un modèle qui est linéaire en ses paramètres et en ses variables. Les modèles linéaires sont donc de la forme :

$$g(\boldsymbol{x},\boldsymbol{w}) = \sum_{i=1}^{p} w_i x_i = \boldsymbol{w} \cdot \boldsymbol{x} .$$

Un *modèle affine* est un modèle linéaire qui contient une constante additive :

$$g(\boldsymbol{x},\boldsymbol{w}) = w_0 + \sum_{i=1}^{p-1} w_i x_i .$$

Remarque

Un modèle affine peut donc être considéré comme un modèle linéaire dont une des variables est constante, égale à 1. Il est donc inutile, en général, de faire une distinction entre modèles linéaires et modèles affines.

Modèles statiques non linéaires en leurs paramètres

On peut imaginer une grande variété de modèles non linéaires en leurs paramètres. Nous étudierons particulièrement dans cet ouvrage les modèles non linéaires en leurs paramètres qui sont de la forme

$$g(\boldsymbol{x},\boldsymbol{w}) = \sum_{i=1}^{p} w_i f_i(\boldsymbol{x},\boldsymbol{w'})$$

où les fonctions f_i sont des fonctions non linéaires, paramétrées par les composantes du vecteur $\boldsymbol{w'}$. Le vecteur $\boldsymbol{w}$ a donc pour composantes les paramètres w_i ($i = 1$ à p) et les composantes de $\boldsymbol{w'}$. Les *réseaux de neurones*, qui sont largement étudiés dans cet ouvrage, constituent un exemple de modèles non linéaires en leurs paramètres et non linéaires en leurs variables.

Modèles dynamiques

Dans les modèles décrits dans la section précédente, le temps ne joue aucun rôle fonctionnel : si les variables $\boldsymbol{x}$ sont indépendantes du temps, la valeur fournie par le modèle (ou *sortie* du modèle) est indépendante du temps. Les modèles dynamiques, en revanche, ont une forme de *mémoire* : la sortie du modèle à un instant donné dépend de ses sorties passées. En conséquence, elle peut évoluer dans le temps, à partir d'un état initial, même si les variables $\boldsymbol{x}$ sont constantes, voire nulles.

La très grande majorité des applications des modèles statistiques sont réalisées à l'aide d'ordinateurs, ou de circuits électroniques numériques. Dans les deux cas, les mesures des variables sont effectuées à intervalles réguliers, dont la durée est appelée *période d'échantillonnage.* De même, les prédictions du modèle ne sont pas fournies de manière continue, mais à intervalles réguliers, généralement caractérisés par la même période d'échantillonnage que les mesures des variables. De tels systèmes sont dits *à temps discret*, par opposition aux systèmes physiques naturels, qui sont des systèmes *à temps continu*.

Ces derniers sont décrits par des modèles dynamiques à temps continu, qui sont des équations (ou des systèmes d'équations) différentielles du type :

$$\frac{dy}{dt} = g(y,\boldsymbol{x},\boldsymbol{w})$$

où t désigne le temps, y la prédiction effectuée par le modèle, $\boldsymbol{x}$ et $\boldsymbol{w}$ les vecteurs des variables et des paramètres respectivement.

Pour les modèles à temps discret, le temps n'est plus une variable continue :

$t = kT$

où T désigne la période d'échantillonnage et k est un nombre entier positif. La prédiction de la valeur prise par la grandeur à modéliser à l'instant kT, connaissant les prédictions effectuées aux n instants précédents, et les valeurs des variables aux m instants précédents, peut alors être mise sous la forme :

$$y(kT) = g\left[y\left((k-1)T\right), y\left((k-2)T\right), \ldots y\left((k-n)T\right), \boldsymbol{x}\left((k-1)T\right), \boldsymbol{x}\left((k-2)T\right), \ldots \boldsymbol{x}\left((k-n')T\right), \boldsymbol{w}\right]$$

où n et n' sont des entiers positifs ; n est appelé *ordre* du modèle. Cette forme de modèle est assez naturelle, mais nous verrons, dans les sections du chapitre 2 consacrées à la modélisation dynamique « boîte noire », et dans les chapitres 4 et 5, qu'il existe des formes plus générales de modèles dynamiques.

Comme pour les modèles statiques, la fonction $g(y, \boldsymbol{x}, \boldsymbol{w})$ peut être soit linéaire, soit non linéaire, par rapport à ses variables et à ses paramètres. Dans la suite de ce chapitre, nous ne considérerons que des modèles statiques ; les modèles dynamiques seront abordés dans les chapitres 2, 4 et 5.

Deux exemples académiques d'apprentissage supervisé

On considère à présent deux exemples académiques, qui permettent de mettre en évidence les problèmes fondamentaux qui se posent dans le domaine de l'apprentissage statistique. Ces deux exemples entrent dans la catégorie de l'apprentissage *supervisé*, dans lequel un *professeur* détermine la réponse que devrait fournir le modèle : dans un problème de classification, le professeur fournit, pour chaque exemple, une étiquette indiquant à quelle classe appartient l'objet ; dans un problème de prédiction, le professeur associe à chaque exemple une mesure de la grandeur à modéliser. L'apprentissage supervisé n'est pas le seul type d'apprentissage ; le chapitre 7 de cet ouvrage sera consacré à un outil très important de l'apprentissage *non supervisé*, les *cartes topologiques*.

Un exemple de modélisation pour la prédiction

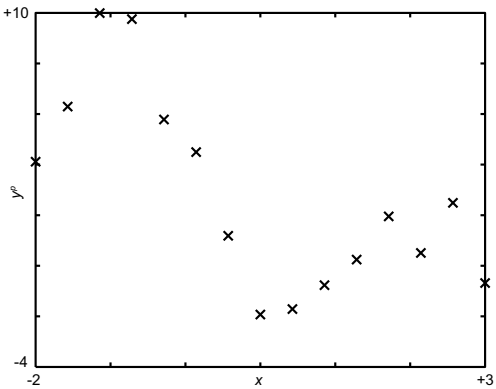


Figure 1-1. Un problème académique de modélisation

Considérons une grandeur y^p, engendrée par un *processus* de nature quelconque – physique, chimique, biologique, sociologique, économique, … – , que l'on cherche à modéliser afin d'en prédire le comportement ; elle dépend d'une seule variable x. Un ensemble d'apprentissage est constitué de $N_A = 15$ mesures y_k^p (k = 1 à N_A), effectuées pour diverses valeurs x_k (k = 1 à N_A) de la variable x. Elles sont représentées par des croix sur la figure 1-1. Nous cherchons à établir un modèle $g(\boldsymbol{x}, \boldsymbol{w})$ qui permette de prédire la valeur de la grandeur à modéliser pour une valeur quelconque de x dans le domaine considéré ($-2 \leq x \leq +3$).

Il s'agit d'un problème académique en ce sens que le processus par lequel ont été créées ces données est connu, ce qui n'est jamais le cas pour un problème réaliste d'apprentissage statistique : on sait que chaque élément k de l'ensemble d'apprentissage a été obtenu en ajoutant à la valeur de $10 \sin(x_k)/x_k$ une réalisation d'une variable aléatoire obéissant à une loi normale (gaussienne de moyenne nulle et d'écart type égal à 1).

Comme indiqué plus haut, il faut d'abord postuler une fonction $g(\boldsymbol{x}, \boldsymbol{w})$. Puisque la grandeur à modéliser ne dépend que de la variable x, le vecteur $\boldsymbol{x}$ se réduit à un scalaire x. En l'absence de toute indication sur la nature du processus générateur des données, une démarche naturelle consiste à postuler des fonctions de complexité croissante, dans une famille de fonctions données. Choisissons la famille des polynômes ; dans cette famille, le modèle polynomial de degré d s'écrit :

$$g(x, \boldsymbol{w}) = w_0 + w_1 x + w_2 x^2 + \ldots + w_d x^d$$

C'est donc un modèle à d+1 paramètres w_0, w_1, …, w_d. Le modèle le plus simple de cette famille est le modèle constant $g(\boldsymbol{x}, \boldsymbol{w}) = w_0$, mis en œuvre dans la section intitulée « Premier exemple ».

Pour effectuer l'apprentissage de ces modèles, on peut utiliser la méthode des moindres carrés, déjà mentionnée. Les détails en seront décrits plus loin, dans la section intitulée « Conception de modèles linéaires par rapport à leur paramètres » ; pour l'instant, il est intéressant d'observer les résultats de ces apprentissages, représentés sur la figure 1-2 pour $d = 1$ (fonction *affine*), $d = 6$ et $d = 10$; le même graphique comporte également une représentation de la fonction 10 sin x / x.

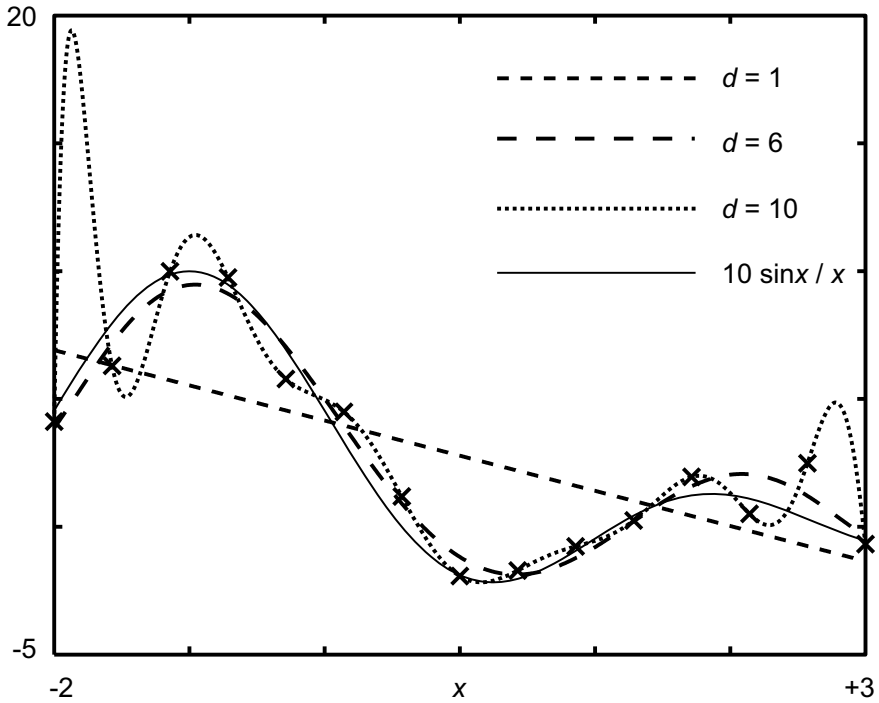


Figure 1-2. Trois modèles polynomiaux

Le modèle affine ($d = 1$) ne rend pas du tout compte des observations car il n'a pas la « souplesse » souhaitable pour s'adapter aux données ; dans le jargon de l'apprentissage statistique, on dira que la *complexité* du modèle est insuffisante. À l'inverse, le modèle polynomial de degré 10 est suffisamment complexe pour passer très précisément par tous les points d'apprentissage ; on observe néanmoins que cette précision sur l'ensemble d'apprentissage est obtenue au détriment des qualités de généralisation du modèle : c'est le phénomène de *surajustement*. En effet, au voisinage de $x = -2$ comme au voisinage de $x = +3$, ce modèle fournit des prédictions très éloignées de la « réalité » représentée en trait plein. En revanche, le modèle polynomial de degré 6 présente un bon compromis : la courbe ne passe pas exactement par tous les points – ce qui est normal puisque ces points résultent en partie d'un tirage aléatoire – mais elle est assez proche de la « vraie » fonction 10 sin$x$ / x.

Afin de rendre ces considérations plus quantitatives, on a constitué, outre l'ensemble d'apprentissage, un deuxième ensemble de données, dit *ensemble de test*, indépendant du précédent, mais dont les N_T

éléments sont issus de la même distribution de probabilité. On définit *l'erreur quadratique moyenne sur l'ensemble d'apprentissage* (EQMA) et *l'erreur quadratique moyenne sur l'ensemble de test* (EQMT) :

$$EQMA = \sqrt{\frac{1}{N_A}\sum_{k=1}^{N_A}\left(y_k^p - g(\boldsymbol{x}_k, \boldsymbol{w})\right)^2} \qquad EQMT = \sqrt{\frac{1}{N_T}\sum_{k=1}^{N_T}\left(y_k^p - g(\boldsymbol{x}_k, \boldsymbol{w})\right)^2} .$$

L'ensemble de test, comprenant $N_T = 1000$ éléments, est représenté sur la figure 1-3. De plus, 100 ensembles d'apprentissage de $N_A = 15$ éléments chacun ont été constitués.

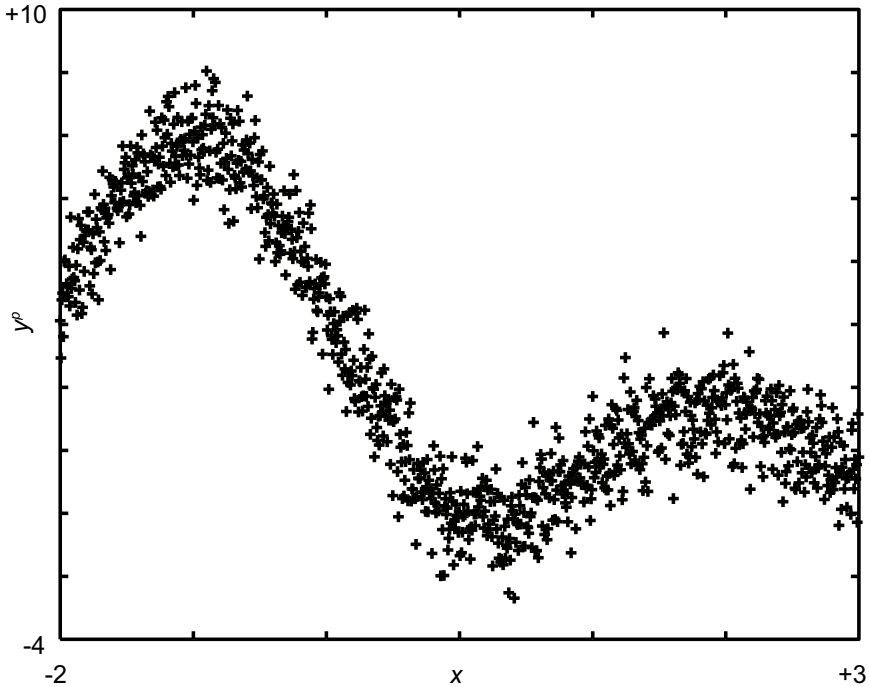


Figure 1-3. Ensemble de test

100 modèles ont été créés à partir de ces ensembles d'apprentissage, et, pour chacun de ces modèles, l'EQMA et l'EQMT ont été calculées. La figure 1-4 montre l'évolution des moyennes des EQMA et EQMT, en fonction de la complexité (degré) du modèle polynomial postulé.

Remarque 1

Le fait de présenter des moyennes des EQMA et EQMT, sur 100 modèles obtenus à partir de 100 ensembles d'apprentissage différents, permet d'éviter l'observation de phénomènes liés à une réalisation particulière du bruit présent dans les observations d'un ensemble d'apprentissage donné. Dans la pratique, on ne dispose évidemment que d'un seul ensemble d'apprentissage.

Remarque 2

Dans la pratique, si l'on disposait d'un ensemble de 1 000 exemples, on utiliserait beaucoup plus que 15 exemples pour effectuer l'apprentissage. Par exemple, on utiliserait 500 exemples pour l'apprentissage et 500 pour tester le modèle. Dans cette section, nous nous plaçons volontairement dans un cadre académique, pour mettre en évidence les phénomènes importants. La méthodologie à adopter pour la conception de modèles est présentée dans la section de ce chapitre intitulée « La conception de modèle en pratique », et elle est largement développée dans le chapitre 2.

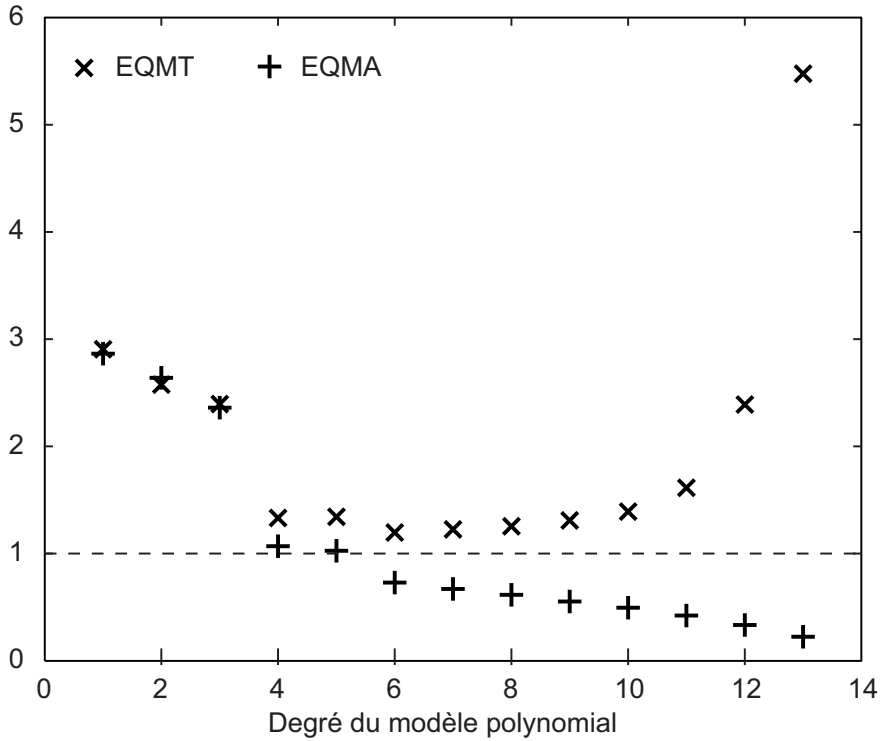


Figure 1-4. Erreurs quadratiques moyennes sur l'ensemble d'apprentissage et sur l'ensemble de test

On observe que l'erreur d'apprentissage (EQMA) diminue lorsque la complexité du modèle augmente : le modèle apprend de mieux en mieux les données d'apprentissage. En revanche, l'erreur sur l'ensemble de test (EQMT) passe par un optimum ($d = 6$) puis augmente : l'augmentation de la complexité du modèle au-delà de $d = 6$ se traduit par une dégradation de ses capacités de généralisation.

Remarque

Les brusques variations de l'EQMA et de l'EQMT observées lorsque l'on passe du degré 3 au degré 4 sont dues à la nature particulière de l'exemple étudié : en effet, dans le domaine de variation de x considéré, la fonction $\sin x / x$ présente deux points d'inflexion (points où la dérivée seconde de la fonction est nulle). Or un polynôme de degré d a au plus $d - 2$ points d'inflexion : pour que le modèle polynomial puisse reproduire les deux points d'inflexion de la fonction génératrice des données, il faut donc qu'il soit au moins de degré 4.

On observe également que l'EQMT reste toujours supérieure à l'écart-type du bruit (qui vaut 1 dans cet exemple), et que l'EQMT du modèle qui a la meilleure généralisation est voisine de l'écart-type du bruit.

Ainsi, le meilleur modèle réalise un compromis entre la précision de l'apprentissage et la qualité de la généralisation. Si le modèle postulé est trop peu complexe, l'apprentissage et la généralisation sont peu précis ; si le modèle est trop complexe, l'apprentissage est satisfaisant, mais la généralisation ne l'est pas. Ce compromis entre la qualité de l'apprentissage et celle de la généralisation, gouverné par la complexité du modèle, est connu sous le terme de *dilemme biais-variance* : un modèle qui a un *biais* faible apprend très bien les points d'apprentissage, mais il peut avoir une *variance* élevée car il peut être fortement tributaire de détails de l'ensemble d'apprentissage (modèle surajusté). En revanche, un modèle peut avoir un *biais* élevé

(il n'apprend pas parfaitement les éléments de l'ensemble d'apprentissage) mais une variance faible (il ne dépend pas des détails de l'ensemble d'apprentissage). Le phénomène observé dans cet exemple est absolument général, comme nous le démontrerons dans la section intitulée « Dilemme biais-variance ».

Dans la section intitulée « Éléments de théorie de l'apprentissage », on donnera une expression quantitative de la notion de complexité. On montrera notamment que, *pour les modèles polynomiaux,* la complexité n'est rien d'autre que le nombre de paramètres du modèle, soit $d + 1$ pour un polynôme de degré d ; on montrera également que le dilemme biais-variance est gouverné par le rapport du nombre de paramètres au nombre d'exemples disponibles.

Retrouvons à présent le même phénomène sur un second exemple académique, qui est cette fois un problème de classification.

Un exemple de classification

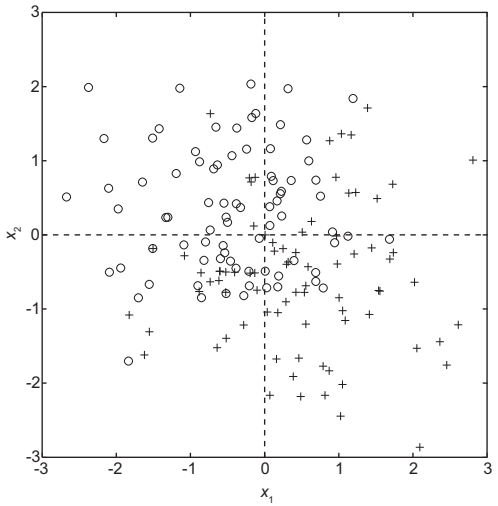


Figure 1-5. Ensemble d'apprentissage pour un problème académique de classification

Rappelons qu'un problème de classification consiste à affecter un objet inconnu à une classe parmi plusieurs. Considérons un problème à deux classes A et B. On souhaite que soit attribuée à tout élément de la classe A une *étiquette* $y^p = +1$, et à tout élément de B une étiquette $y^p = -1$. On dispose d'un ensemble d'apprentissage, constitué d'exemples de chacune des classes, dont la classe est connue : des étiquettes exactes leur ont été affectées. Dans le problème considéré ici, chaque « objet » est décrit par un vecteur $\boldsymbol{x}$ à deux composantes : on peut donc le représenter par un point dans le plan des variables (x_1, x_2). La figure 1-5 représente un ensemble d'apprentissage comprenant 80 exemples par classe. Les exemples de la classe A sont représentés par des croix, ceux de la classe B par des cercles. On cherche la *frontière* entre ces classes, c'est-à-dire une ligne, dans ce plan, qui sépare les exemples avec un nombre d'erreurs minimal : on souhaite que la plupart des exemples d'une classe (voire tous) soient d'un côté de la frontière, et que la plupart des exemples de l'autre classe (voire tous) soient de l'autre côté ; on dit qu'il y a une erreur de classification lorsqu'un exemple est situé « du mauvais côté » de la frontière.

Comme dans l'exemple de modélisation, on présente ici un problème académique : le processus générateur des données est connu, ce qui n'est pas le cas pour un problème réel. Les exemples de l'ensemble d'apprentissage ont été tirés de deux distributions gaussiennes isotropes d'écart-type égal à 1, dont les centres sont respectivement le point $(x_1 = + 0{,}5 ; x_2 = - 0{,}5)$ pour la classe A, et $(x_1 = - 0{,}5 ; x_2 = + 0{,}5)$ pour la classe B. On démontrera, dans la section intitulée « Classifieur de Bayes », que la diagonale du carré, qui est l'axe de symétrie du problème, est la frontière pour laquelle le risque d'erreur de classification est minimal. On voit que cette frontière théorique idéale ne sépare pas parfaitement bien tous les exemples d'apprentissage : le taux d'erreur sur l'ensemble d'apprentissage n'est pas nul si l'on choisit cette frontière, mais nous montrerons que le taux d'erreur sur l'ensemble de tous les objets, engendrés par le même processus générateur, mais n'appartenant pas à l'ensemble d'apprentissage, est minimal.

Le classifieur de Bayes présente donc une généralisation optimale ; malheureusement, on ne peut le déterminer que si les distributions des exemples sont connues, ce qui n'est généralement pas le cas dans un problème réel. On peut seulement s'efforcer de trouver un classifieur qui en soit proche. C'est ce qui va être tenté par les deux méthodes décrites ci-dessous.

La méthode des k plus proches voisins

Une approche naïve consiste à considérer que des points voisins ont une grande chance d'appartenir à une même classe. Alors, étant donné un objet inconnu décrit par le vecteur $\boldsymbol{x}$, on peut décider que cet objet appartient à la classe de l'exemple d'apprentissage qui est le plus proche de l'extrémité de $\boldsymbol{x}$. De manière plus générale, on peut décider de considérer les k plus proches voisins de l'objet inconnu, et d'affecter celui-ci à la classe à laquelle appartient la majorité des k exemples les plus proches (on prend de préférence k impair). Cette approche, appelée *méthode des k plus proches voisins*, revient à postuler une fonction $g(\boldsymbol{x},k)=\frac{1}{k}\sum_{i=1}^{k} y_i^p$, où la somme porte sur les k exemples les plus proches de $\boldsymbol{x}$, et à mettre en œuvre la règle suivante : l'objet décrit par $\boldsymbol{x}$ est affecté à la classe A si $\operatorname{sgn}\left(g\left(\boldsymbol{x},k\right)\right)=+1$, et il est affecté à la classe B dans le cas contraire[1]. On construit ainsi un modèle constant par morceaux, égal à la moyenne des étiquettes des k exemples les plus proches. Le seul paramètre du modèle est donc k, le nombre de plus proches voisins pris en considération dans la moyenne.

Pour visualiser les résultats, le calcul est effectué pour 10 000 points disposés régulièrement sur une grille de 100×100 points. La figure 1-6 montre les résultats obtenus pour $k = 1$, $k = 7$, $k = 21$ et $k = 159$ (cette dernière valeur est la valeur maximale de k puisque l'ensemble d'apprentissage comporte en tout 160 exemples) ; les points affectés à la classe A par le classifieur sont représentés en gris foncé, ceux qui sont affectés à la classe B en gris clair.

Pour $k = 1$, on observe que la frontière est très irrégulière, et définit des « îlots » de l'une des classes dans l'autre classe. Ce phénomène s'explique facilement : comme chaque point de l'ensemble d'apprentissage est son propre plus proche voisin, il est forcément bien classé. La frontière dépend donc complètement de l'ensemble d'apprentissage choisi : un autre tirage aléatoire de points dans les *mêmes* distributions gaussiennes aurait produit une frontière très différente. C'est un modèle qui a un biais faible (tous les exemples de l'ensemble d'apprentissage étant bien appris, le taux d'erreur sur l'ensemble d'apprentissage est nul) et une variance élevée (la frontière varie beaucoup si l'on change l'ensemble d'apprentissage). La capacité de généralisation est donc certainement très faible, le modèle étant complètement surajusté à l'ensemble d'apprentissage disponible. La croix en traits épais ($x_1 = -2$, $x_2 = -2{,}5$), qui n'appartient pas à l'ensemble d'apprentissage, est mal classée.

Lorsque l'on augmente k, la frontière devient plus régulière, et plus proche de la frontière optimale ($k = 7$, $k = 21$). La croix en traits épais est correctement classée dans l'ensemble des croix. Pour $k = 159$, on observe en revanche que la frontière devient très régulière, mais qu'elle est très éloignée de la solution optimale (la diagonale du carré). La croix en traits épais est à nouveau mal classée.

On passe ainsi de modèles de faible biais et grande variance (faibles valeurs de k) à des modèles de faible variance mais de biais élevé (grandes valeurs de k). Comme dans l'exemple précédent, on voit apparaître la nécessité de trouver un compromis satisfaisant entre le biais et la variance ; ce compromis dépend la valeur de $1 / k$.

1. La fonction $\operatorname{sgn}(u)$ est définie de la manière suivante : $\operatorname{sgn}(u) = +1$ si $u > 0$, $\operatorname{sgn}(u) = -1$ si $u \leq 0$

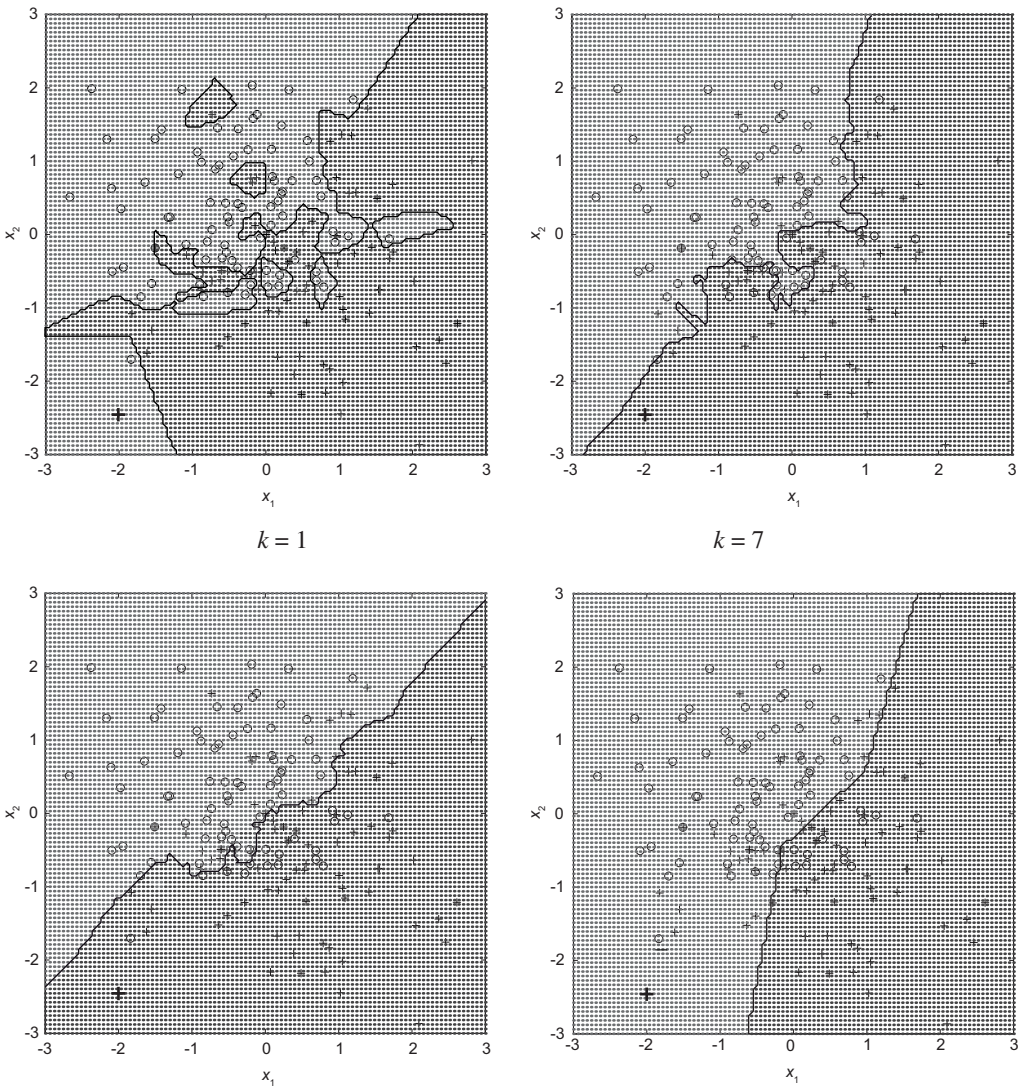


Figure 1-6. Classification par la méthode des k plus proches voisins

Pour caractériser quantitativement ce phénomène, on peut procéder comme pour l'exemple précédent : on constitue un ensemble de test de 1000 points, et 100 ensembles d'apprentissage de tailles identiques (100 exemples par classe), tirés des mêmes distributions de probabilités. Pour différentes valeurs de *k*, on construit un modèle à partir de chaque ensemble d'apprentissage par la méthode des plus proches voisins, soit 100 modèles pour chaque valeur de *k*. Pour chaque modèle, on calcule le taux d'erreur de classification (rapport du nombre d'exemples mal classés au nombre total d'exemples) sur l'ensemble d'apprentissage et sur l'ensemble de test ; on calcule enfin la moyenne de ces taux d'erreur sur les 100 ensembles d'apprentissage considérés. La figure 1-7 présente les taux moyens d'erreur de classification sur l'ensemble d'apprentissage (+), et l'erreur sur l'ensemble de test (x), pour *k* variant de 3 à 199. Pour les faibles complexités (*k* grand), le taux d'erreur sur les ensembles d'apprentissage et de test sont grands, et du même ordre de grandeur ; pour les complexités élevées (*k* petit), le taux d'erreur sur l'ensemble d'apprentissage tend vers zéro, tandis que le taux d'erreur sur l'ensemble de test croît. Ce comportement est donc tout à fait analogue à celui qui a été observé pour la prédiction (figure 1-4). Le taux d'erreur sur l'ensemble de test passe par un minimum, appelé « limite de Bayes », qui, dans le cas particulier de deux distributions gaussiennes, peut être calculé si l'on connaît les moyennes et écarts-types de ces distributions (voir la section « Classifieur de Bayes ») ; avec les valeurs numériques considérées ici, ce taux théorique est de 23,9 %, ce qui est bien le résultat observé dans cette expérience numérique (la valeur du taux d'erreur théorique est établie dans la section de ce chapitre intitulée « Classification : règle de Bayes et classifieur de Bayes »).

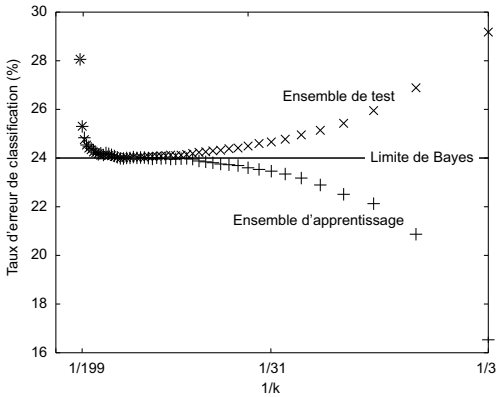


Figure 1-7. Erreurs d'apprentissage et de test pour la méthode des k plus proches voisins

Ainsi, le dilemme biais-variance, illustré dans l'exemple de modélisation, se retrouve ici sous une forme différente : l'augmentation du nombre de plus proches voisins, donc la diminution de la « complexité », entraîne une augmentation du nombre d'erreurs de classification dans l'ensemble d'apprentissage, mais une diminution du nombre d'erreurs en-dehors de l'ensemble d'apprentissage, donc une meilleure généralisation.

Le tableau 1-1 résume les aspects du dilemme biais-variance, pour la classification par la méthode des plus proches voisins d'une part, et pour la prédiction d'autre part.

	Classification (*k* plus proches voisins)	**Prédiction (modèles linéaires)**
Dilemme biais-variance gouverné par	$\frac{\text{Nombre d'exemples}}{\text{Nombre de plus proches voisins}}$	$\frac{\text{Nombre de paramètres}}{\text{Nombre d'exemples}}$
Limite inférieure de l'erreur de généralisation	Limite de Bayes	Variance du bruit

Tableau 1-1. Dilemme biais-variance pour la classification par la méthode des plus proches voisins et pour la prédiction par des modèles linéaires ou polynomiaux

Classification linéaire ou polynomiale

Rappelons que la méthode des *k* plus proches voisins consiste à calculer, pour tout objet décrit par $\boldsymbol{x}$, la fonction

$$g(\boldsymbol{x}) = \frac{1}{k} \sum_{\substack{k \text{ plus proches} \\ \text{voisins de } \boldsymbol{x}}} y_k^p$$

et à utiliser la règle de décision suivante : si $\text{sgn}\left(g\left(\boldsymbol{x}\right)\right) = +1$ l'objet décrit par $\boldsymbol{x}$ est affecté à la classe *A*, si $\text{sgn}\left(g\left(\boldsymbol{x}\right)\right) = -1$ il est affecté à la classe *B*.

Cette approche peut être généralisée de la manière suivante : on cherche à estimer, par apprentissage, les paramètres d'une fonction $g(\boldsymbol{x},\boldsymbol{w})$ telle que $\text{sgn}\left(g(\boldsymbol{x},\boldsymbol{w})\right) = +1$ pour tous les objets de la classe *A* et $\text{sgn}\left(g\left(\boldsymbol{x},\boldsymbol{w}\right)\right) = -1$ pour tous les objets de la classe *B*. La fonction $\gamma(\boldsymbol{x},\boldsymbol{w}) = \dfrac{1+\text{sgn}\left[g(\boldsymbol{x},\boldsymbol{w})\right]}{2}$, qui vaut +1 pour tous les éléments de *A* et 0 pour tous les éléments de *B*, est appelée *fonction indicatrice*.

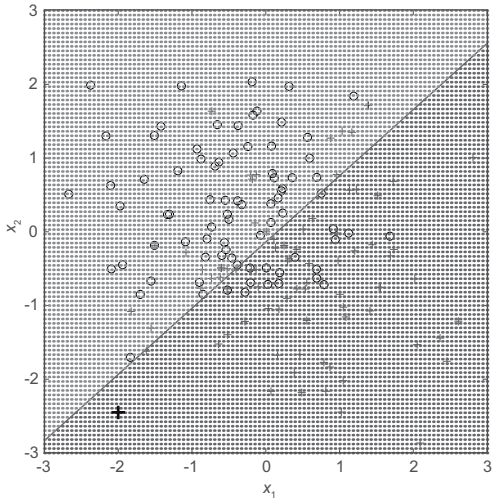


Figure 1-8. Séparation linéaire

Au lieu de postuler une fonction constante par morceaux comme on le fait dans la méthode des *k* plus proches voisins, postulons à présent une fonction polynomiale. La plus simple d'entre elles est la fonction affine $g(\boldsymbol{x},\boldsymbol{w}) = w_0 + w_1x_1 + w_2x_2$, que l'on peut encore écrire $g(\boldsymbol{x},\boldsymbol{w}) = \boldsymbol{x} \cdot \boldsymbol{w}$, où le symbole $\cdot$ représente le produit scalaire ; $\boldsymbol{x}$ est le vecteur de composantes $\{1,\ x_1,\ x_2\}$ et $\boldsymbol{w}$ est le vecteur de composantes $\{w_0, w_1, w_2\}$. Pour chaque exemple *k* de l'ensemble d'apprentissage, on écrit que $g\left(\boldsymbol{x}^k,\boldsymbol{w}\right) = y_k^p$, où $y_k^p = +1$ pour tous les exemples de la classe *A* et $y_k^p = -1$ pour tous les exemples de la classe *B*. On met alors en œuvre la méthode des moindres carrés, décrite plus loin dans la section « Apprentissage de modèles linéaires », pour estimer le vecteur des paramètres $\boldsymbol{w}$. Pour l'ensemble d'apprentissage représenté sur la figure 1-5, le résultat obtenu est représenté sur la figure 1-8. On observe que la frontière ainsi définie est proche de la première diagonale du carré, laquelle garantit la meilleure généralisation.

Comme dans le cas de la modélisation que nous avons étudié plus haut, le dilemme biais-variance est gouverné par le rapport du nombre de paramètres du modèle (1 + degré du polynôme) au nombre d'exemples disponibles. La figure 1-9 montre l'évolution du taux d'erreur de classification, sur l'ensemble d'apprentissage et sur l'ensemble de test, à complexité donnée (3 paramètres), en fonction du nombre d'exemples.

Lorsque le nombre d'exemples est faible, le taux d'erreur sur l'ensemble d'apprentissage est très petit (biais faible) et le taux d'erreur sur l'ensemble de test est très grand (variance importante). En revanche, lorsque le nombre d'exemples augmente, les deux taux d'erreur convergent vers le taux d'erreur de Bayes (qui, rappelons-le, peut être calculé analytiquement dans ce cas, et vaut 23,9 %).

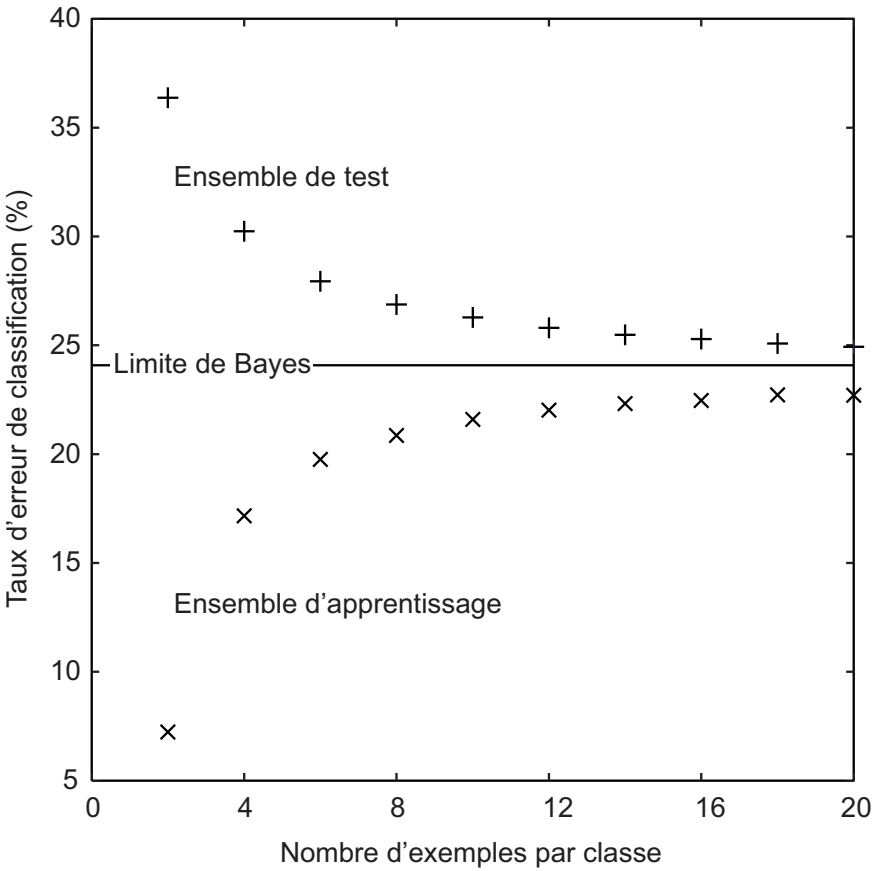


Figure 1-9. Taux d'erreur en fonction du nombre d'exemples, à complexité fixée

Conclusion

Dans cette section, deux problèmes académiques simples d'apprentissage supervisé ont été présentés : un exemple de prédiction et un exemple de classification. Ces deux exemples ont permis de mettre en évidence un problème central de l'apprentissage artificiel : le dilemme biais-variance, c'est-à-dire la nécessité de trouver le meilleur compromis possible entre la capacité du modèle à apprendre les exemples d'apprentissage et sa capacité à généraliser à des situations non apprises. Ces observations empiriques vont à présent être justifiées de manière plus rigoureuse.

Éléments de théorie de l'apprentissage

Cette section présente quelques résultats théoriques fondamentaux concernant l'apprentissage supervisé, pour la prédiction et la classification. On présentera tout d'abord un formalisme général pour la modélisation par apprentissage. On introduira ensuite le classifieur de Bayes, et l'on en démontrera les propriétés. Enfin, on prouvera que le dilemme biais-variance est un phénomène général.

Fonction de perte, erreur de prédiction théorique

Puisque l'apprentissage cherche à reproduire les données, il faut définir une « distance » entre les prédictions du modèle et les données : on définit donc une fonction dite « fonction de perte »

$$\pi\left[y^p, g(\boldsymbol{x},\boldsymbol{w})\right] \geq 0,$$

où y^p est la valeur souhaitée et $g(\boldsymbol{x}, \boldsymbol{w})$ est la valeur prédite par le modèle, dont les paramètres sont les composantes du vecteur $\boldsymbol{w}$, étant donné le vecteur de variables $\boldsymbol{x}$. Pour une tâche de prédiction, y^p est la valeur mesurée de la grandeur à prédire ; pour une tâche de classification à deux classes, y^p vaut +1 pour un objet d'une classe et −1 (ou 0) pour un objet de l'autre classe.

Exemples

Une distance naturelle, très fréquemment utilisée, est l'erreur quadratique de modélisation :

$$\pi\left[y^p, g(\boldsymbol{x},\boldsymbol{w})\right] = \left[y^p - g(\boldsymbol{x},\boldsymbol{w})\right]^2.$$

Il arrive aussi que l'on utilise la valeur absolue de l'erreur :

$$\pi\left[y^p, g(\boldsymbol{x},\boldsymbol{w})\right] = \left|y^p - g(\boldsymbol{x},\boldsymbol{w})\right|.$$

Comment décrire mathématiquement la « qualité » du modèle ? Comme dans la première section de ce chapitre, on peut modéliser les résultats des mesures y^p comme des réalisations d'une variable aléatoire Y^p, et les vecteurs des variables $\boldsymbol{x}$ comme des réalisations d'un vecteur aléatoire $\boldsymbol{X}$. Alors les valeurs de la fonction de perte π deviennent elles-mêmes des réalisations d'une variable aléatoire Π, fonction de Y^p et de $\boldsymbol{X}$, et il est naturel de caractériser la performance du modèle par l'espérance mathématique de Π, ou erreur de prédiction théorique, que nous noterons P^2 (cette quantité est toujours positive, d'après la définition de π) :

$$P^2 = E_\Pi = \int\int \pi\left(y^p, g(\boldsymbol{x},\boldsymbol{w})\right) p_{Y^p,\boldsymbol{X}} dy^p d\boldsymbol{x}$$

où $p_{Y^p,\boldsymbol{X}}$ est la probabilité conjointe de la variable aléatoire Y^p et du vecteur aléatoire $\boldsymbol{X}$; les intégrales portent sur toutes les valeurs possibles de la grandeur à modéliser et des variables qui la gouvernent. Cette erreur de prédiction est bien une erreur *théorique* : en pratique, on ne dispose que d'un ensemble de données de taille finie, et les distributions de probabilités sont inconnues. On ne peut donc jamais calculer cette erreur, mais seulement l'estimer à partir de l'ensemble de données dont on dispose.

Notons que, pour les modèles dont les paramètres $\boldsymbol{w}$ sont déterminés par apprentissage, ces derniers dépendent aussi des réalisations de Y^p présentes dans l'ensemble d'apprentissage : les paramètres du modèle peuvent donc aussi être considérés comme des réalisations de variables aléatoires. Nous utiliserons cette remarque dans la section intitulée « Dilemme biais-variance ».

L'erreur de prédiction théorique peut alors s'écrire :

$$P^2 == E_{\boldsymbol{X}}\left[E_{Y^p|\boldsymbol{X}}(\Pi)\right]$$

où $E_{Y^p|\boldsymbol{X}}(\Pi)$ désigne l'espérance mathématique de la variable aléatoire $\Pi\left(Y^p|\boldsymbol{X}\right)$, c'est-à-dire l'espérance mathématique de la fonction de perte pour les prédictions effectuées par le modèle *pour un vecteur de variables $\boldsymbol{x}$ donné*.

Démonstration

La probabilité conjointe peut s'écrire $p_{Y^p,\boldsymbol{X}} = p_{Y^p}\left(y^p|\boldsymbol{x}\right)p_{\boldsymbol{X}}$. L'erreur de prédiction théorique s'écrit donc

$$P^2 = \int\int \pi\left(y^p, g(\boldsymbol{x},\boldsymbol{w})\right)\, p_{Y^p}\left(y^p|\boldsymbol{x}\right) p_{\boldsymbol{x}}\, dy^p\, d\boldsymbol{x}$$
$$= \int\left[\int \pi\left(y^p, g(\boldsymbol{x},\boldsymbol{w})\right)\, p_{Y^p}\left(y^p|\boldsymbol{x}\right) dy^p\right] p_{\boldsymbol{x}}\, d\boldsymbol{x}$$
$$= E_{\boldsymbol{x}}\left[E_{Y^p|X}(\Pi)\right].$$

Considérons un exemple caractérisé par le vecteur de variables $\boldsymbol{x}$. En ce point, le meilleur modèle est le modèle pour lequel l'erreur de prédiction théorique est minimum. Appliquons cette propriété successivement à deux tâches : la prédiction et la classification.

Prédiction

Comme indiqué plus haut, la fonction de perte la plus fréquemment utilisée pour la prédiction est

$$\pi\left[y^p, g(\boldsymbol{x},\boldsymbol{w})\right] = \left[y^p - g(\boldsymbol{x},\boldsymbol{w})\right]^2$$

Alors le meilleur modèle possible est la *fonction de régression* de la grandeur à modéliser :

$$f(\boldsymbol{x}) = E_{Y^p|X}$$

Démonstration

Rappelons que l'espérance mathématique de la fonction de perte est donnée par :

$$E_{Y^p|X}(\Pi) = \int\left(y^P - g(x,\boldsymbol{w})\right)^2 p_{Y^p}\left(y^p|\boldsymbol{x}\right) dy^p.$$

Son minimum est obtenu pour le modèle $f(x)$ tel que

$$0 = \left(\frac{dE_{Y^p|X}}{dg(\boldsymbol{x},\boldsymbol{w})}\right)_{g(\boldsymbol{x},\boldsymbol{w})=f(\boldsymbol{x})}$$
$$= \left(\frac{d\int\left(y^P - g(x,\boldsymbol{w})\right)^2 p_{Y^p}\left(y^p|\boldsymbol{x}\right) dy^p}{dg(\boldsymbol{x},\boldsymbol{w})}\right)_{g(\boldsymbol{x},\boldsymbol{w})=f(\boldsymbol{x})}$$
$$= 2\int\left(y^P - f(x)\right) p_{Y^p}\left(y^p|\boldsymbol{x}\right) dy^p$$
$$= 2\int y^P p_{Y^p}\left(y^p|\boldsymbol{x}\right) dy^p - 2f(\boldsymbol{x})\int p_{Y^p}\left(y^p|\boldsymbol{x}\right) dy^p.$$

La première intégrale n'est autre que l'espérance mathématique de Y^p étant donné $\boldsymbol{x}$; la seconde est égale à 1 par définition de la densité de probabilité. On obtient ainsi : $E_{Y^p|X} = f(\boldsymbol{x})$.

La distribution de probabilité des observations étant inconnue, la fonction de régression est inconnue. Pour connaître sa valeur en $\boldsymbol{x}$, il faudrait réaliser une infinité de mesures de la grandeur y^p pour une valeur donnée des variables $\boldsymbol{x}$ et faire la moyenne des résultats de ces mesures, ce qui n'est évidemment pas réaliste.

Classification : règle de Bayes et classifieur de Bayes

Considérons à présent un problème de classification à deux classes A et B. Affectons l'étiquette $y^p = +1$ à tous les exemples de la classe A et l'étiquette $y^p = -1$ à tous les exemples de la classe B. Comme nous l'avons fait plus haut, nous cherchons une fonction $g(\boldsymbol{x}, \boldsymbol{w})$ qui permettra d'affecter à la classe A tous les éléments pour lesquels $\text{sgn}[g(\boldsymbol{x}, \boldsymbol{w})] = +1$, et à la classe B tous les éléments pour lesquels $\text{sgn}[g(\boldsymbol{x}, \boldsymbol{w})] = -1$.

Cette fonction doit être telle que l'erreur de prédiction théorique soit minimale (on trouvera dans le chapitre 6 un traitement beaucoup plus détaillé de ce problème).

Règle de décision de Bayes

Pour la prédiction, considérée dans la section précédente, on a mis en œuvre, pour définir l'erreur théorique, la fonction de perte des moindres carrés. Pour la classification, on ne cherche pas à approcher les valeurs des résultats de mesures, mais à classer correctement des objets. On utilise donc une autre fonction de perte, mieux adaptée à ce problème :

$$\pi\left[y^p, \operatorname{sgn}\left(g(\boldsymbol{x},\boldsymbol{w})\right)\right] = 0 \text{ si } y^p = \operatorname{sgn}\left(g(\boldsymbol{x},\boldsymbol{w})\right)$$
$$\pi\left[y^p, \operatorname{sgn}\left(g(\boldsymbol{x},\boldsymbol{w})\right)\right] = 1 \text{ si } y^p \neq \operatorname{sgn}\left(g(\boldsymbol{x},\boldsymbol{w})\right)$$

Ainsi, la fonction de perte vaut 1 si le classifieur commet une erreur de classement pour l'objet décrit par $\boldsymbol{x}$, et 0 sinon. Contrairement au cas de la prédiction, cette fonction est à valeurs discrètes. L'espérance mathématique de la variable aléatoire discrète Π n'est autre que la probabilité pour que le classifieur considéré commette une erreur de classification pour un objet décrit par $\boldsymbol{x}$; en effet :

$$\begin{aligned} E_{\Pi}(\boldsymbol{x}) &= 1 \times \Pr_{\Pi}(1|\boldsymbol{x}) + 0 \times \Pr_{\Pi}(0|\boldsymbol{x}) \\ &= \Pr_{\Pi}(1|\boldsymbol{x}). \end{aligned}$$

Cette quantité est inconnue : pour l'estimer, il faudrait disposer d'une infinité d'objets décrits par $\boldsymbol{x}$, dont les classes sont connues, et compter la fraction de ces objets qui est mal classée par le classifieur considéré.

La variable aléatoire Π est fonction de Y^p. Son espérance mathématique peut donc s'écrire :

$$E_{\Pi}(\boldsymbol{x}) = \pi\left(+1, \operatorname{sgn}\left(g(\boldsymbol{x},\boldsymbol{w})\right)\right)\Pr_{Y^p}(+1|\boldsymbol{x}) + \pi\left(-1, \operatorname{sgn}\left(g(\boldsymbol{x},\boldsymbol{w})\right)\right)\Pr_{Y^p}(-1|\boldsymbol{x}).$$

La probabilité d'appartenance d'un objet à une classe C connaissant le vecteur de variables $\boldsymbol{x}$ qui décrit cet objet, notée $\Pr_{Y^p}(C|\boldsymbol{x})$, est appelée *probabilité a posteriori* de la classe C pour l'objet décrit par $\boldsymbol{x}$.

On remarque que $E_{\Pi}(\boldsymbol{x})$ ne peut prendre que deux valeurs :

$$E_{\Pi}(\boldsymbol{x}) = \Pr_{Y^p}(+1|\boldsymbol{x}) \text{ si } \operatorname{sgn}\left(g(\boldsymbol{x},\boldsymbol{w})\right) = -1,$$
$$E_{\Pi}(\boldsymbol{x}) = \Pr_{Y^p}(-1|\boldsymbol{x}) \text{ si } \operatorname{sgn}\left(g(\boldsymbol{x},\boldsymbol{w})\right) = +1.$$

Supposons que la probabilité a posteriori de la classe A au point $\boldsymbol{x}$ soit supérieure à celle de la classe B :

$$\Pr_{Y^p}(+1|\boldsymbol{x}) > \Pr_{Y^p}(-1|\boldsymbol{x}).$$

Rappelons que l'on cherche la fonction $g(\boldsymbol{x},\boldsymbol{w})$ pour laquelle la probabilité d'erreur de classification au point $\boldsymbol{x}$, c'est-à-dire $E_{\Pi}(\boldsymbol{x})$, soit minimum. La fonction $g(\boldsymbol{x},\boldsymbol{w})$ pour laquelle $E_{\Pi}(\boldsymbol{x})$ est minimum est donc telle que $\operatorname{sgn}\left(g(\boldsymbol{x},\boldsymbol{w})\right) = +1$, puisque, dans ce cas, $E_{\Pi}(\boldsymbol{x}) = \Pr_{Y^p}(-1|\boldsymbol{x})$, qui est la plus petite des deux valeurs possibles.

À l'inverse, si $\Pr_{Y^p}(-1|\boldsymbol{x}) > \Pr_{Y^p}(+1|\boldsymbol{x})$, la fonction $g(\boldsymbol{x},\boldsymbol{w})$ qui garantit le plus petit taux d'erreur en $\boldsymbol{x}$ est telle que $\operatorname{sgn}\left(g(\boldsymbol{x},\boldsymbol{w})\right) = -1$.

En résumé, *le meilleur classifieur possible est celui qui, pour tout* $\boldsymbol{x}$*, affecte l'objet décrit par* $\boldsymbol{x}$ *à la classe dont la probabilité a posteriori est la plus grande en ce point.*

Cette règle de décision (dite *règle de Bayes*) garantit que le nombre d'erreurs de classification est minimal ; pour pouvoir la mettre en œuvre, il faut calculer (ou estimer) les probabilités a posteriori des classes.

Classifieur de Bayes

Le classifieur de Bayes utilise, pour le calcul des probabilités a posteriori, la *formule de Bayes* : étant donné un problème à c classes C_i (i = 1 à c), la probabilité a posteriori de la classe C_i est donnée par la relation

$$\Pr\left(C_i|\boldsymbol{x}\right) = \frac{p_X\left(\boldsymbol{x}|C_i\right)\Pr_{C_i}}{\sum_{j=1}^{c} p_X\left(\boldsymbol{x}|C_j\right)\Pr_{C_j}}$$

où $p_X\left(\boldsymbol{x}|C_j\right)$ est la densité de probabilité du vecteur $\boldsymbol{x}$ des variables observées pour les objets de la classe C_j (ou *vraisemblance* du vecteur $\boldsymbol{x}$ dans la classe C_j), et $\Pr_{C_j}$ est la *probabilité a priori* de la classe C_j, c'est-à-dire la probabilité pour qu'un objet tiré au hasard appartienne à la classe C_j.

Si toutes les classes ont la même probabilité a priori $1/c$, la règle de Bayes revient à classer l'objet inconnu $\boldsymbol{x}$ dans la classe pour laquelle $\boldsymbol{x}$ a la plus grande vraisemblance : c'est une application de la méthode du *maximum de vraisemblance*.

Ainsi, si l'on connaît analytiquement les vraisemblances, et si l'on connaît les probabilités a priori des classes, on peut calculer *exactement* les probabilités a posteriori.

Exemple : cas de deux classes gaussiennes de mêmes variances

Reprenons le cas considéré plus haut, dans la section intitulée « un exemple de classification » : deux classes A et B dans un espace à deux dimensions, telles que les vraisemblances des variables sont gaussiennes, de même variance σ, de centres $\boldsymbol{x}_A$ (x_{1A}, x_{2A}) et $\boldsymbol{x}_B$ (x_{1B}, x_{2B}) :

$$p_X\left(\boldsymbol{x}|A\right) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\frac{\left(x_1 - x_{1A}\right)^2}{2\sigma^2}\right]\exp\left[-\frac{\left(x_2 - x_{2A}\right)^2}{2\sigma^2}\right]$$

$$p_X\left(\boldsymbol{x}|B\right) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\frac{\left(x_1 - x_{1B}\right)^2}{2\sigma^2}\right]\exp\left[-\frac{\left(x_2 - x_{2B}\right)^2}{2\sigma^2}\right].$$

Supposons que les probabilités a priori des classes soient les mêmes, égales à 0,5.

> Dans l'exemple considéré plus haut, chaque classe était représentée par le même nombre d'exemples. Si la probabilité a priori des classes est estimée par la fréquence des exemples, c'est-à-dire le rapport du nombre d'exemples d'une classe au nombre total d'exemples, on est dans le cas où les deux probabilités a priori sont égales à 0,5.

Alors la formule de Bayes permet de calculer les probabilités a posteriori :

$$\Pr(A|\boldsymbol{x}) = \frac{0,5\frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\frac{(x_1-x_{1A})^2}{2\sigma^2}\right]\exp\left[-\frac{(x_2-x_{2A})^2}{2\sigma^2}\right]}{0,5\frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\frac{(x_1-x_{1A})^2}{2\sigma^2}\right]\exp\left[-\frac{(x_2-x_{2A})^2}{2\sigma^2}\right]+0,5\frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\frac{(x_1-x_{1B})^2}{2\sigma^2}\right]\exp\left[-\frac{(x_2-x_{2B})^2}{2\sigma^2}\right]}$$

$$\Pr(B|\boldsymbol{x}) = \frac{0,5\frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\frac{(x_1-x_{1B})^2}{2\sigma^2}\right]\exp\left[-\frac{(x_2-x_{2B})^2}{2\sigma^2}\right]}{0,5\frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\frac{(x_1-x_{1A})^2}{2\sigma^2}\right]\exp\left[-\frac{(x_2-x_{2A})^2}{2\sigma^2}\right]+0,5\frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\frac{(x_1-x_{1B})^2}{2\sigma^2}\right]\exp\left[-\frac{(x_2-x_{2B})^2}{2\sigma^2}\right]}.$$

La règle de classification de Bayes affecte l'objet décrit par $\boldsymbol{x}$ à la classe dont la probabilité a posteriori est la plus grande (ou, puisque les probabilités a priori sont égales, à la classe pour laquelle la vraisemblance de $\boldsymbol{x}$ est la plus grande).

La frontière entre les classes est donc le lieu des points, dans l'espace des vecteurs $\boldsymbol{x}$, où les vraisemblances sont égales : c'est le lieu des points tels que

$$\exp\left[-\frac{(x_1-x_{1A})^2}{2\sigma^2}\right]\exp\left[-\frac{(x_2-x_{2A})^2}{2\sigma^2}\right] = \exp\left[-\frac{(x_1-x_{1B})^2}{2\sigma^2}\right]\exp\left[-\frac{(x_2-x_{2B})^2}{2\sigma^2}\right].$$

soit encore

$$(x_1-x_{1A})^2+(x_2-x_{2A})^2 = (x_1-x_{1B})^2+(x_2-x_{2B})^2.$$

La frontière optimale entre les classes est donc le lieu des points équidistants des centres des distributions : c'est la médiatrice du segment de droite qui joint ces centres.

Dans l'exemple considéré plus haut, les centres des gaussiennes étaient symétriques par rapport à la diagonale du carré représenté sur la figure 1-6 et la figure 1-8, donc la meilleure frontière possible entre les classes était la diagonale de ce carré. Le résultat le plus proche du résultat théorique était le séparateur linéaire de la figure 1-8 ; en effet, on avait postulé un modèle linéaire, et celui-ci était « vrai » au sens statistique du terme, c'est-à-dire que la solution optimale du problème appartenait à la famille des fonctions dans laquelle nous cherchions une solution par apprentissage. On était donc dans les meilleures conditions possibles pour trouver une bonne solution par apprentissage.

Connaissant la surface de séparation fournie par le classifieur de Bayes, et sachant que les classes ont le même nombre d'éléments, il est facile de trouver le taux d'erreur de ce classifieur : c'est la probabilité de trouver un élément de la classe *A* (classe des +) dans le demi-plan supérieur gauche (ou, par symétrie, la probabilité de trouver un élément de *B* (classe des o) dans le demi-plan complémentaire) :

$$\frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{+\infty}\exp\left[-\frac{(x_1-x_{1A})^2}{2\sigma^2}\right]\int_{x_2>x_1}\exp\left[-\frac{(x_2-x_{2A})^2}{2\sigma^2}\right]dx_1dx_2,$$

avec $\sigma = 1$ dans l'exemple considéré.

Cette expression se calcule très simplement en effectuant une rotation des axes de 45° dans le sens trigonométrique, suivie d'une translation, de manière que la frontière entre les classes devienne verticale et que le centre de la classe A soit à l'origine (figure 1-10). Le taux d'erreur est alors la probabilité cumulée d'une variable normale entre $-\infty$ et $-\sqrt{2}/2$. On trouve facilement cette dernière valeur à l'aide d'un logiciel de statistiques, ou sur le Web (par exemple http://www.danielsoper.com/statcalc/calc02_do.aspx) : elle vaut environ 24 %, comme indiqué plus haut.

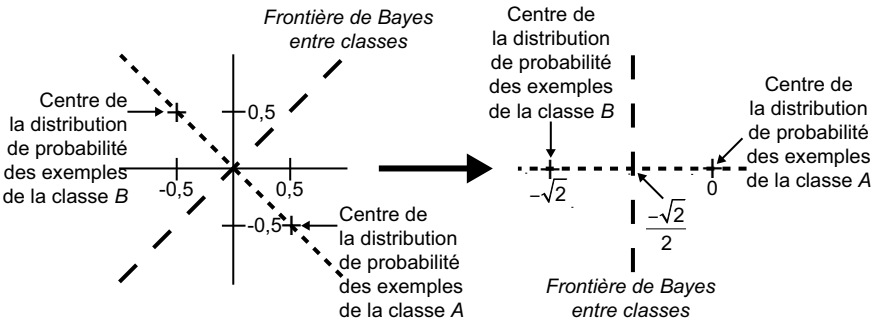


Figure 1-10. Rotation et translation des axes

Dilemme biais-variance

Les deux exemples académiques considérés en début de chapitre ont permis de mettre en évidence le dilemme biais-variance. Muni des éléments théoriques de la section précédente, on peut à présent formaliser ce problème.

Considérons le cas de la prédiction par un modèle dont les paramètres sont déterminés par apprentissage ; comme indiqué plus haut, la fonction de perte la plus fréquemment utilisée dans ce cas est le carré de l'erreur de modélisation :

$$\pi\left[y^p, g(\boldsymbol{x},\boldsymbol{w})\right] = \left[y^p - g(\boldsymbol{x},\boldsymbol{w})\right]^2$$

et l'erreur de prédiction théorique est

$$P^2 = E_{\boldsymbol{X}}\left[E_{Y^p|X}\left[\left[y^p - g(\boldsymbol{x},\boldsymbol{w})\right]^2\right]\right].$$

Cherchons l'erreur de prédiction en un point $\boldsymbol{x}$ de l'espace des variables

$$P^2(\boldsymbol{x}) = E_{Y^p|\boldsymbol{X}}\left[\left[y^p - g(\boldsymbol{x},\boldsymbol{w})\right]^2\right],$$

en supposant que les observations y^p effectuées en ce point $\boldsymbol{x}$ sont des réalisations de la variable aléatoire $Y^p = f(\boldsymbol{x}) + \varepsilon$

où ε est une variable aléatoire d'espérance mathématique nulle et de variance σ^2, et où $f(\boldsymbol{x})$ est une fonction certaine ; l'espérance mathématique de Y^p est donc $f(\boldsymbol{x})$, la fonction de régression de y^p, dont on a vu plus haut que c'est le meilleur modèle possible au sens de la fonction de perte choisie.

Supposons enfin que le modèle soit obtenu par apprentissage : les paramètres $\boldsymbol{w}$ du modèle doivent donc être considérés comme des réalisations d'un vecteur aléatoire $\boldsymbol{W}$ qui dépend des réalisations de Y^p

présentes dans l'ensemble d'apprentissage ; de même, les prédictions $g(\boldsymbol{x}, \boldsymbol{w})$ peuvent être considérées comme des réalisations d'une variable aléatoire $G(\boldsymbol{x}, \boldsymbol{W})$ qui dépendent de Y^p. Pour rendre les équations plus lisibles, on remplace ici la notation var_X par var (X) et E_X par $E(X)$.

L'erreur de prédiction théorique au point $\boldsymbol{x}$ est alors donnée par :

$$P^2(\boldsymbol{x}) = \sigma^2 + \text{var}\left[G(\boldsymbol{x}, \boldsymbol{W})\right] + \left[E\left[f(\boldsymbol{x}) - G(\boldsymbol{x}, \boldsymbol{W})\right]\right]^2,$$

où le phénomène aléatoire est la constitution de l'ensemble d'apprentissage.

Démonstration

Rappelons que, pour une variable aléatoire Z, on a la relation

$$E_{Z^2} = \text{var}_Z + \left[E_Z\right]^2.$$

Le modèle étant construit par apprentissage, ses paramètres, donc les prédictions du modèle, sont eux-mêmes des réalisations de variables aléatoires $\boldsymbol{W}$ et $G(\boldsymbol{x}, \boldsymbol{W})$ par l'intermédiaire de Y^p. On peut donc écrire :

$$\begin{aligned} P^2(\boldsymbol{x}) &= = E\left[\left[Y^p - G(\boldsymbol{x}, \boldsymbol{W})\right]^2\right] = \text{var}\left[Y^p - G(\boldsymbol{x}, \boldsymbol{w})\right] + \left[E\left[Y^p - G(\boldsymbol{x}, \boldsymbol{W})\right]\right]^2 \\ &= \text{var}\left[Y^p - f(\boldsymbol{x}) + f(\boldsymbol{x}) - G(\boldsymbol{x}, \boldsymbol{W})\right] + \left[E\left[Y^p - f(\boldsymbol{x}) + f(\boldsymbol{x}) - G(\boldsymbol{x}, \boldsymbol{W})\right]\right]^2 \\ &= \text{var}\left[\varepsilon + f(\boldsymbol{x}) - G(\boldsymbol{x}, \boldsymbol{W})\right] + \left[E\left[\varepsilon + f(\boldsymbol{x}) - G(\boldsymbol{x}, \boldsymbol{W})\right]\right]^2. \end{aligned}$$

La fonction $f(\boldsymbol{x})$ étant certaine (elle ne dépend pas de $\boldsymbol{W}$, donc de l'ensemble d'apprentissage), sa variance est nulle. D'autre part, l'espérance mathématique de ε est nulle : on a donc finalement :

$$P^2(\boldsymbol{x}) = \sigma^2 + \text{var}\left[G(\boldsymbol{x}, \boldsymbol{W})\right] + \left[E\left[f(\boldsymbol{x}) - G(\boldsymbol{x}, \boldsymbol{W})\right]\right]^2.$$

Le premier terme de la somme est la variance du bruit de mesure. Le deuxième est la variance de la prédiction du modèle au point $\boldsymbol{x}$, qui représente la sensibilité du modèle à l'ensemble d'apprentissage. Le troisième est le biais du modèle, c'est-à-dire le carré de l'espérance mathématique de l'écart entre les prédictions fournies par le modèle et celles qui sont fournies par le meilleur modèle possible (la fonction de régression $f(\boldsymbol{x})$).

Cette relation très importante appelle plusieurs commentaires :

- La qualité d'un modèle ne peut être évaluée que par comparaison entre son erreur de prédiction et la variance du bruit sur les mesures. Un modèle qui fournit des prédictions en désaccord de 10 % avec les mesures est un excellent modèle si les mesures ont elles-mêmes une précision de 10 % ; mais si la précision sur les mesures est de 1 %, le modèle est très mauvais : il faut chercher à l'améliorer. Si la précision sur les mesures est de 20 %, la performance de 10% annoncée pour le modèle est très suspecte : son estimation doit être remise en cause. Les trois termes de la somme étant positifs, l'erreur de prédiction théorique ne peut être inférieure à la variance des observations en $\boldsymbol{x}$, c'est-à-dire à la variance du bruit qui affecte les mesures ; *en d'autres termes, on ne peut pas espérer qu'un modèle, conçu par apprentissage, fournisse des prédictions plus précises que les mesures à partir desquelles il a été construit*. C'est ce qui a été observé sur la figure 1-4, où le minimum de la racine carrée de l'erreur de prédiction théorique, estimée par l'EQMT, était de l'ordre de l'écart-type du bruit.
- On retrouve par cette relation le fait que le meilleur modèle est la fonction de régression : en effet, si $g(\boldsymbol{x}, \boldsymbol{w}) = f(\boldsymbol{x})$, la variance est nulle puisque le modèle ne dépend pas de $\boldsymbol{w}$, et le biais est nul ; l'erreur de prédiction est donc la plus petite possible, égale à la variance du bruit.

- Si le modèle ne dépend pas de paramètres ajustables, la variance est nulle, mais le biais peut être très grand puisque le modèle ne dépend pas des données. Par exemple, si $g(\boldsymbol{x}, \boldsymbol{w}) = 0$, la variance est nulle et le biais vaut $\left[f(\boldsymbol{x}) \right]^2$.

Dans les exemples académiques de prédiction et de classification que nous avons présentés, nous avons observé que le biais et la variance varient en sens inverse en fonction de la complexité du modèle : un modèle trop complexe par rapport aux données dont on dispose possède une variance élevée et un biais faible, alors qu'un modèle de complexité insuffisante a une variance faible mais un biais élevé. Comme l'erreur de généralisation fait intervenir la somme de ces deux termes, elle passe par un optimum qui est au moins égal à la variance du bruit. C'est exactement ce que nous avons observé sur la figure 1-4 : l'erreur quadratique moyenne sur l'ensemble de test, qui est une estimation de l'erreur de généralisation, passe par un minimum pour un polynôme de degré 6, qui présente donc la complexité optimale compte tenu des données d'apprentissage dont on dispose.

La relation qui vient d'être établie fournit l'erreur de prédiction théorique en un point $\boldsymbol{x}$. L'erreur de prédiction théorique est

$$P^2 = E_X\left[P^2(\boldsymbol{x}) \right] = \int P^2(\boldsymbol{x}) p_X d\boldsymbol{x}$$
$$= \sigma^2 + E_X\left[\mathrm{var}\left[G(\boldsymbol{x}, \boldsymbol{W}) \right] \right] + E_X\left[E\left[f(\boldsymbol{x}) - G(\boldsymbol{x}, \boldsymbol{W}) \right] \right]^2.$$

Remarque

L'espérance mathématique E_X n'a pas le même sens que l'espérance mathématique E : la première porte sur toutes les conditions expérimentales possibles, tandis que la seconde porte sur toutes les réalisations possibles de l'ensemble d'apprentissage.

Pour vérifier numériquement cette relation, reprenons l'exemple de la modélisation par apprentissage à partir de données qui ont été créées artificiellement en ajoutant à la fonction $10 \sin(x)/x$ un bruit pseudo-aléatoire de variance égale à 1, en $N_A = 15$ points x_k. Pour estimer le biais et la variance en un point $\boldsymbol{x}$, 100 ensembles d'apprentissage différents ont été créés, en tirant au hasard, dans une distribution normale centrée, 100 valeurs de y^p pour chaque valeur de x_k ; on a fait l'apprentissage de 100 modèles différents $g(\boldsymbol{x}, \boldsymbol{w}_i)$, $i = 1$ à 100, c'est-à-dire que 100 vecteurs de paramètres ont été estimés par la méthode des moindres carrés (qui sera décrite plus loin). Un ensemble de test de 1 000 points a été créé, et, en chaque point de cet ensemble, le biais et la variance du modèle de paramètres $\boldsymbol{w}_i$ ont été estimés :

- estimation du biais du modèle $g(\boldsymbol{x}, \boldsymbol{w}_i)$ au point $x_{\boldsymbol{k}}^{\boldsymbol{test}}$: $\dfrac{1}{100}\displaystyle\sum_{i=1}^{100}\left(10\frac{\sin x_k^{test}}{x_k^{test}} - g\left(x_k^{test}, \boldsymbol{w}_i\right) \right)^2$

- estimation de la variance du modèle $g(\boldsymbol{x}, \boldsymbol{w}_i)$ au point $\boldsymbol{x}_k^{test}$:

$$\frac{1}{99}\sum_{i=1}^{100}\left(g\left(x_k^{test}, \boldsymbol{w}_i\right) - \frac{1}{100}\sum_{j=1}^{100} g\left(x_k^{test}, w_j\right) \right)^2.$$

L'erreur de prédiction $P^2\left(x_k^{test}\right)$ est estimée par :

$$\frac{1}{100}\sum_{i=1}^{100}\left(y_k^{test} - g\left(x_k^{test}, \boldsymbol{w}_i\right) \right)^2.$$

Finalement, les espérances mathématiques de ces trois quantités sont estimées par la moyenne de chacune d'elles sur les 1 000 points de test.

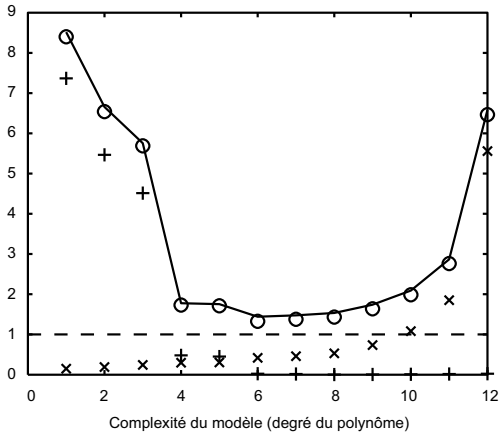


Figure 1-11. Dilemme biais-variance pour la régression
x *: estimation de l'espérance mathématique de la variance ;*
+ : estimation de l'espérance mathématique du biais
o *: variance du bruit + variance de la prédiction + biais de la prédiction ;*
trait plein : estimation de l'espérance mathématique de l'erreur de prédiction;
tirets : variance du bruit

La figure 1-11 montre, en fonction de la complexité du modèle, les estimations du biais du modèle, de la variance du modèle, ainsi que la valeur de la variance du bruit. La somme de ces trois quantités (représentée par des cercles) est en excellent accord avec l'estimation de l'erreur de prédiction (courbe en trait plein). On observe clairement que le biais et la variance varient en sens opposés, et que la somme passe par un minimum pour les polynômes de degré 6.

Les résultats ci-dessus ont été établis pour la prédiction. Pour la classification, ils prennent une forme analogue, comme illustré numériquement sur la figure 1-7. De manière générale, on peut résumer la problématique du dilemme biais-variance comme représenté sur la figure 1-12 : le meilleur modèle, au sens statistique du terme, constitue un compromis entre l'ignorance (modèles incapables d'apprendre) et la stupidité (modèles surajustés, qui apprennent très bien et sont incapables de généraliser).

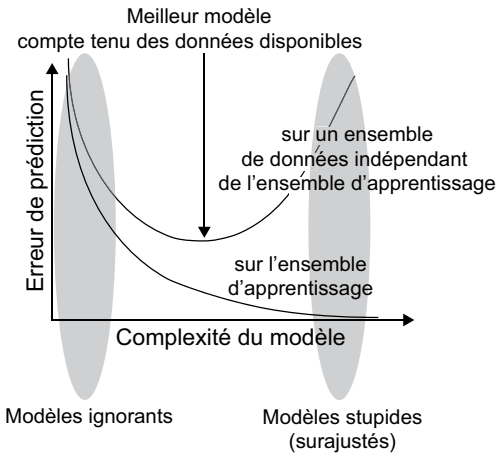


Figure 1-12. Représentation symbolique du dilemme biais-variance

De la théorie à la pratique

Les résultats qui ont été présentés dans la section précédente sont des résultats *asymptotiques*, c'est-à-dire qu'ils sont exacts si l'on dispose d'une quantité infinie de données. Ils sont très utiles, car ils expliquent les grandes lignes des phénomènes que l'on observe, et mettent en évidence les problèmes qu'il faut résoudre. Néanmoins, dans une situation réelle de mise en œuvre des méthodes d'apprentissage artificiel, on dispose toujours d'une quantité finie de données, insuffisante pour estimer de manière très précise les

intégrales nécessaires au calcul des espérances mathématiques ou des variances ; de plus, les distributions de probabilités auxquelles obéissent les données sont également inconnues. Dans cette section, on présente des résultats théoriques sur lesquels il est possible de s'appuyer pour trouver des méthodes pratiques de conception de modèles par apprentissage. Le lecteur qui ne cherche pas à approfondir la théorie de l'apprentissage peut sans dommage omettre de lire cette section et passer directement à la section intitulée « La conception de modèles en pratique ».

Remplacer des intégrales par des sommes

Rappelons que l'objectif de la modélisation par apprentissage est de trouver des fonctions paramétrées qui sont susceptibles de rendre compte des données disponibles, et de fournir des prédictions aussi précises que possible concernant des données dont on ne dispose pas lors de l'apprentissage. L'objectif théorique est donc de trouver le vecteur de paramètres $\boldsymbol{w}$ pour lequel l'erreur de prédiction *théorique*

$$P^2 = E_{\Pi} = \int\int \pi\left(y^p, g(\boldsymbol{x},\boldsymbol{w})\right) p_{Y^p,\boldsymbol{X}}\, dy^p d\boldsymbol{x}$$

est minimale. L'intégrale n'étant pas calculable, il convient donc de l'estimer à l'aide des données disponibles. On estime donc l'erreur de prédiction théorique par l'erreur de prédiction *empirique* (également appelée *risque empirique*)

$$P*^2 = \frac{1}{N}\sum_{k=1}^{N} \pi\left(y_k^p, g(\boldsymbol{x}_k,\boldsymbol{w})\right)$$

où $\pi\left(y_k^p, g(\boldsymbol{x}_k,\boldsymbol{w})\right)$ est la fonction de perte choisie.

L'apport fondamental de la théorie de l'apprentissage, par rapport aux statistiques classiques, réside dans l'étude de la manière dont l'erreur empirique converge (ou ne converge pas) vers l'erreur théorique. Ainsi, en statistique, on montre que la moyenne est un estimateur non biaisé de l'espérance mathématique ; la théorie de l'apprentissage, pour sa part, s'intéresse à la façon dont la moyenne converge vers l'espérance mathématique lorsque le nombre d'exemples augmente. Ainsi on peut évaluer le nombre d'exemples nécessaires pour estimer l'espérance mathématique avec une précision donnée, ou bien évaluer l'erreur que l'on commet en estimant l'espérance mathématique par la moyenne, pour un nombre d'exemples donné.

Comme indiqué plus haut, la fonction de perte la plus utilisée dans le cas de la prédiction est le carré de l'erreur, et l'erreur de prédiction empirique est donnée par

$$P*^2 = \frac{1}{N}\sum_{k=1}^{N} \left(y_k^p - g(\boldsymbol{x}_k,\boldsymbol{w})\right)^2$$

où la somme porte sur un ensemble de données convenablement choisies parmi les données disponibles.

La première tâche consiste à estimer les paramètres $\boldsymbol{w}$, c'est-à-dire à effectuer *l'apprentissage* proprement dit. Pour cela, on choisit, parmi les données disponibles, un *ensemble d'apprentissage*, de cardinal N_A, et l'on cherche, à l'aide d'algorithmes appropriés, le vecteur $\boldsymbol{w}$ pour lequel la *fonction de coût*

$$J = \sum_{k=1}^{N_A} \pi\left(y_k^p - g(\boldsymbol{x}_k,\boldsymbol{w})\right)$$

est minimale. Rappelons que, dans le cas où π est le carré de l'erreur, la fonction

$$J = \sum_{k=1}^{N_A} \left(y_k^p - g(\boldsymbol{x}_k, \boldsymbol{w}) \right)^2$$

est appelée *fonction de coût des moindres carrés*.

Supposons donc que l'on ait trouvé le minimum de la fonction de coût choisie ; la valeur de ce minimum est-elle représentative de la qualité des prédictions que fournira le modèle, muni des paramètres ainsi déterminés, pour des valeurs de $\boldsymbol{x}$ qui ne font pas partie de l'ensemble d'apprentissage ? Les exemples précédents montrent que la réponse est généralement négative. Ainsi, la figure 1-4 montre que l'erreur quadratique moyenne sur l'ensemble d'apprentissage (EQMA), qui vaut $\sqrt{J}$, est très inférieure à l'erreur quadratique moyenne sur l'ensemble de test pour des modèles trop complexes (de degré supérieur ou égal à 7). De même, la figure 1-9 montre que l'erreur sur l'ensemble d'apprentissage est très optimiste, c'est-à-dire très inférieure à l'erreur sur l'ensemble de test, lorsque le nombre d'exemples est petit. D'autre part, l'erreur sur l'ensemble de test elle-même n'est qu'une estimation, à l'aide d'un nombre fini d'exemples, de l'erreur de prédiction théorique. On peut donc en tirer deux enseignements :

- d'une part, il ne faut généralement pas estimer la performance d'un modèle à partir des résultats de l'apprentissage ;
- d'autre part, il faut estimer le mieux possible l'erreur de prédiction.

Les deux sections suivantes décrivent, d'une part, des éléments théoriques qui permettent de borner l'erreur que l'on commet en estimant les capacités de généralisation à partir des estimations obtenues à l'aide de données en nombre fini, et, d'autre part, des éléments méthodologiques qui permettent de définir les « bonnes pratiques » pour la conception de modèles par apprentissage.

Bornes sur l'erreur de généralisation

Les résultats théoriques présentés dans la section « Dilemme biais-variance » sont des résultats *asymptotiques,* qui sont *exacts* dans la limite où les exemples sont en nombre infini. Dans le cas, plus réaliste, où les exemples sont en nombre fini, on ne peut plus établir de résultats exacts ; en revanche, on peut obtenir des résultats *en probabilité.* Le cadre théorique le plus fréquemment utilisé est celui de la théorie de l'apprentissage établie par V. Vapnik [VAPNIK 1998].

Le résultat le plus remarquable de cette théorie consiste en une expression quantitative de la notion de *complexité* du modèle : étant donnée une famille de fonction $g(\boldsymbol{x}, \boldsymbol{w})$, la complexité de cette famille peut être caractérisée par une grandeur, appelée *dimension de Vapnik-Chervonenkis*. Le fait qu'il suffise d'une seule grandeur pour définir la complexité d'une famille de fonctions quelconque est très remarquable ; il faut néanmoins admettre que le calcul de la dimension de Vapnik-Chervonenkis pour une famille de fonctions n'est pas toujours simple.

Pour la famille des polynômes de degré d, la dimension de Vapnik-Chervonenkis est égale au nombre de paramètres du modèle, soit d+1.

En classification, la dimension de Vapnik-Chervonenkis admet une interprétation géométrique simple : c'est le nombre maximal de points qui peuvent être séparés sans erreur par une fonction indicatrice appartenant à la famille considérée. On trouvera dans le chapitre 6 une justification originale et bien développée de la dimension de Vapnik-Chervonenkis, dans le cadre de la classification.

Exemple

Considérons la famille des fonctions affines à deux variables x_1 et x_2. Il est facile de prouver que la dimension de Vapnik-Chervonenkis de cette famille de fonctions est égale à 3 : la figure 1-13 montre que les points appartenant à toutes les configurations possibles de 3 points appartenant à deux classes, en dimension 2, peuvent être séparés par une fonction affine. En revanche, la figure 1-14 montre une configuration de 4 points qui ne sont pas séparables par une fonction de cette famille. Cette configuration admet néanmoins un séparateur quadratique (une hyperbole), ce qui prouve que la dimension de Vapnik-Chervonenkis des fonctions affines de deux variables est égale à 3, et que celle des fonctions quadratiques de deux variables est supérieure à 3 ; comme indiqué plus haut, elle est égale au nombre de paramètres, soit 6 pour les polynômes du second degré à deux variables.

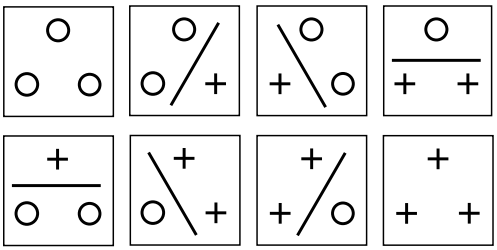


Figure 1-13. Toute configuration de 3 points dans le plan, appartenant à deux classes, admet un séparateur affine.

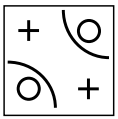


Figure 1-14. Une configuration de 4 points qui n'admet pas de séparateur affine, mais qui admet un séparateur quadratique.

La dimension de Vapnik-Chervonenkis est généralement une fonction croissante du nombre de paramètres. Mais ce n'est pas toujours le cas. Ainsi, la fonction $\mathrm{sgn}(\sin wx)$ a un seul paramètre, mais peut séparer un nombre quelconque de points : il suffit de choisir une longueur d'onde $2\pi/w$ suffisamment petite. Sa dimension de Vapnik-Chervonenkis est infinie (figure 1-15).

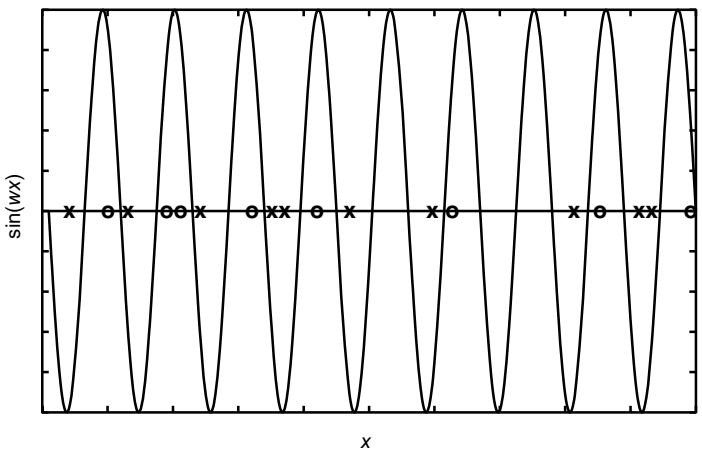


Figure 1-15. La dimension de Vapnik-Chervonenkis de la fonction sin(wx) est infinie.

Cette définition de la complexité permet d'établir des bornes sur l'erreur commise en remplaçant l'erreur de prédiction théorique P^2 par une erreur empirique P^{*2} estimée sur l'ensemble d'apprentissage. Ainsi, supposons que l'on effectue l'apprentissage d'un classifieur en cherchant la fonction indicatrice $\gamma(\boldsymbol{x},\boldsymbol{w}) = \dfrac{1+\mathrm{sgn}\left[g(\boldsymbol{x},\boldsymbol{w})\right]}{2}$ (de valeur 0 ou 1, comme indiqué plus haut) qui minimise une erreur empirique $P^{*2}(\boldsymbol{w})$ sur un ensemble d'apprentissage de cardinal N_A. Soit h la dimension de Vapnik-

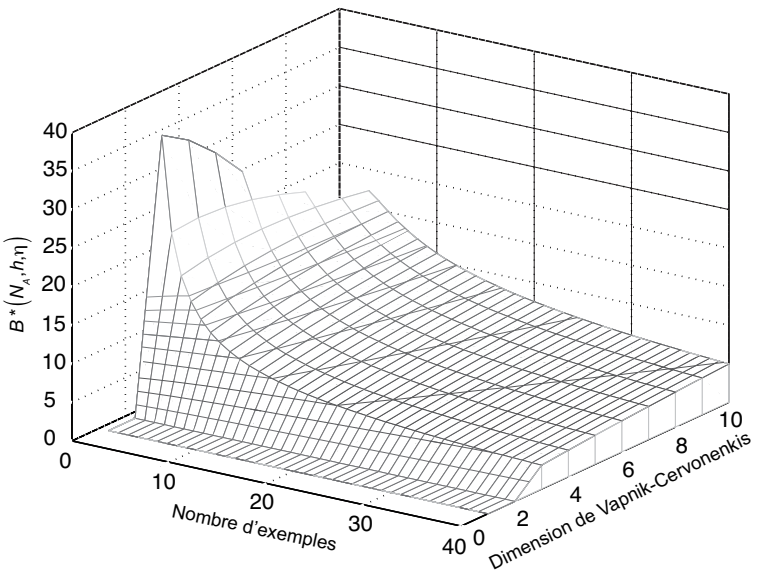
Chervonenkis de $g(\boldsymbol{x}, \boldsymbol{w})$. On a le résultat suivant : si $N_A > h$, alors, avec une probabilité au moins égale à $1 - \eta$, *pour toute fonction de cette famille*, la différence entre l'erreur de généralisation (inconnue) $P^2(\boldsymbol{w})$ commise par cette fonction et l'erreur empirique $P^{*2}(\boldsymbol{w})$ calculée sur les données d'apprentissage est bornée supérieurement par la quantité

$$B\left(N_A, h, \eta\right) = \frac{E\left(N_A, h, \eta\right)}{2}\left(1 + \sqrt{1 + \frac{4P^{*2}\left(\boldsymbol{w}\right)}{E\left(N_A, h, \eta\right)}}\right),$$

où $E\left(N_A, h, \eta\right) = 4\dfrac{h\left[\ln\left(2\dfrac{N_A}{h}\right) + 1\right] - \ln\left(\dfrac{\eta}{4}\right)}{N_A}$.

De plus, pour la fonction $g\left(\boldsymbol{x}, \boldsymbol{w}^*\right)$ pour laquelle l'erreur empirique est minimale (c'est-à-dire pour le modèle de la famille considérée qui est trouvé par apprentissage), avec une probabilité au moins égale à $1\text{-}2\eta$, la différence entre l'erreur de généralisation $P^2(\boldsymbol{w}^*)$ commise *par cette fonction* et la plus petite erreur de généralisation qui puisse être commise par un modèle de cette famille est bornée supérieurement par :

$$B^*\left(N_A, h, \eta\right) = \sqrt{\frac{-\ln\eta}{2N_A}} + \frac{E\left(N_A, h, \eta\right)}{2}\left(1 + \sqrt{1 + \frac{4}{E\left(N_A, h, \eta\right)}}\right).$$



La figure 1-16 montre l'évolution de $B^*(N_A, h, \eta)$ en fonction du nombre d'exemples et de la dimension de Vapnik-Chervonenkis ($\eta = 10^{-2}$). On observe que cette borne croît lorsque le nombre d'exemples diminue, ce qui confirme le fait, mis en évidence dans les exemples présentés plus haut, que la qualité du modèle est d'autant meilleure que le nombre d'exemples est grand devant la complexité du modèle.

Figure 1-16. Exemple de borne théorique

Dans la pratique, la mise en œuvre de ces bornes est peu utile, car elles sont généralement très pessimistes ; elles peuvent éventuellement être utilisées pour comparer des modèles entre eux. Néanmoins, l'approche possède le très grand mérite de mettre en évidence des comportements universels de familles de fonctions, indépendamment de la distribution des exemples, pour des nombres d'exemples

finis, et de fournir des guides pour la conception de modèles utiles dans des applications difficiles. Ainsi, les machines à vecteurs supports, décrites dans le chapitre 6, permettent un contrôle sur la dimension de Vapnik-Chervonenkis.

Minimisation du risque structurel

Les considérations développées dans les sections précédentes conduisent naturellement à un élément important de la méthodologie de conception de modèle, dite méthode de *minimisation du risque structurel*. Elle consiste à :
- postuler des modèles de complexité croissante, par exemple des polynômes de degré croissant ;
- trouver le ou les modèles pour lesquels l'erreur de prédiction empirique est minimale pour chaque complexité, éventuellement en pénalisant la variance par des méthodes de *régularisation* qui seront décrites dans le chapitre 2 ;
- choisir le meilleur modèle.

Les méthodes de conception de modèle qui seront décrites dans cet ouvrage entrent dans ce cadre.

Conception de modèles en pratique

Les exemples qui ont été exposés, et les considérations théoriques qui ont été décrites, illustrent les grandes lignes de la méthodologie de conception de modèles qu'il convient de suivre de manière rigoureuse pour obtenir, par apprentissage, des modèles précis et fiables, donc utiles. Dans cette section, nous récapitulons les étapes de conception d'un tel modèle.

Collecte et prétraitement des données

La première étape est évidemment la collecte des données. Deux situations peuvent se présenter :
- le modèle doit être conçu à partir d'une base de données préexistante, que l'on ne peut pas enrichir ;
- le concepteur du modèle peut spécifier les expériences qui doivent être effectuées pour améliorer le modèle.

Une fois les données disponibles, il convient de les traiter de manière à rendre la modélisation aussi efficace que possible.

Les données sont préexistantes

Là encore, il faut distinguer deux cas :
- les données sont peu nombreuses ; il faut alors s'efforcer de tirer le meilleur parti de ces données, en construisant des modèles aussi parcimonieux que possible en nombre de paramètres ;
- les données sont très nombreuses : on peut alors mettre en œuvre des méthodes dites de *planification expérimentale* ou d'*apprentissage actif*, afin de ne retenir que les exemples qui apportent une réelle information au modèle. La description détaillée de ces méthodes sort du cadre de cet ouvrage, mais des éléments en seront décrits dans les chapitres qui suivent.

Les données peuvent être spécifiées par le concepteur

Dans un tel cas, il est très souhaitable de mettre en œuvre des méthodes de planification expérimentale, surtout si les expériences sont longues ou coûteuses. Les plans d'expérience permettent en effet de limiter

le nombre d'expériences, en n'effectuant que celles qui sont réellement utiles pour la conception du modèle.

Prétraitement des données

Une fois les données disponibles, il faut effectuer un prétraitement qui permette de rendre la modélisation aussi efficace que possible. Ces prétraitements dépendent de la tâche à effectuer et des particularités des données que l'on manipule. *Dans tous les cas*, le prétraitement minimal consiste à normaliser et à centrer les données, de manière à éviter, par exemple, que certaines variables aient de très grandes valeurs numériques par rapport à d'autres, ce qui rendrait les algorithmes d'apprentissage inefficaces. Le prétraitement le plus simple consiste donc à effectuer le changement de variables suivant, pour les variables $\boldsymbol{x}$ comme pour la grandeur à modéliser y^p :

$$u' = \frac{u - \langle u \rangle}{s_u},$$

où $\langle u \rangle$ désigne la moyenne de la grandeur u considérée

$$\langle u \rangle = \frac{1}{N} \sum_{k=1}^{N} u_k,$$

et s_u est l'estimateur de l'écart-type de u :

$$s_u = \sqrt{\frac{1}{N-1} \sum_{k=1}^{N} \left(u - \langle u \rangle\right)^2}.$$

Ainsi, toutes les grandeurs sont de moyenne nulle et d'écart-type unité.

Dans toute la suite, on supposera *toujours* que les grandeurs considérées ont été préalablement normalisées et centrées.

Sélection des variables

Lorsqu'on modélise un processus physique ou chimique bien connu, on détermine généralement, par une analyse préalable du problème, les variables qui ont une influence sur le phénomène étudié ; dans ce cas, une étape de sélection des variables n'est pas toujours nécessaire. En revanche, ce n'est pas le cas lorsqu'on cherche à modéliser un processus économique, social ou financier, ou encore un processus physico-chimique complexe ou mal connu : les experts du domaine peuvent donner des indications sur les facteurs qu'ils estiment pertinents, mais il s'agit souvent de jugements subjectifs qu'il faut mettre à l'épreuve des faits. On est alors conduit à retenir un grand nombre de variables candidates (appelées également *facteurs* ou *descripteurs*), potentiellement pertinentes. Néanmoins, la complexité du modèle croît avec le nombre de variables : par exemple, la dimension de Vapnik-Chervonenkis de polynômes de degré d vaut $\frac{(n+d)!}{n!d!}$, où n est le nombre de variables ; elle croît donc très rapidement avec n. Conserver un contrôle sur le nombre de variables est donc un élément important dans une stratégie de modélisation qui cherche à maîtriser la complexité des modèles. Nous décrirons plus en détail, dans ce chapitre, le problème de la sélection de variables et nous proposerons une méthode efficace pour le résoudre.

Les résultats de la sélection de variables sont susceptibles de remettre en cause des idées reçues concernant le phénomène à modéliser, ou, au contraire, de conforter des conjectures ou des intuitions concernant l'influence des variables candidates sur la grandeur à modéliser.

On peut également souhaiter diminuer le nombre de variables en réduisant la dimension de l'espace de représentation de la grandeur que l'on cherche à modéliser. Les principales méthodes utilisées dans ce but sont l'Analyse en Composantes Principales (ACP), l'Analyse en Composantes Indépendantes (ACI, ou ICA pour Independent Component Analysis) ou encore l'Analyse en Composantes Curvilignes (ACC). L'ACP et l'ACC sont décrites dans le chapitre 3 de cet ouvrage.

Apprentissage des modèles

Les méthodes d'apprentissage de différentes familles de modèles seront décrites en détail dans les différents chapitres de cet ouvrage. Comme nous l'avons déjà vu, elles consistent toutes à optimiser des fonctions bien choisies par des méthodes appropriées. L'apprentissage des modèles linéaires en leurs paramètres est décrit dans ce chapitre, dans la section « Conception de modèles linéaires par rapport à leurs paramètres (régression linéaire) ».

Sélection de modèles

Comme indiqué plus haut, la méthode de minimisation du risque structurel conduit à concevoir des modèles de complexités différentes et à choisir celui qui est susceptible d'avoir les meilleures propriétés de généralisation.

Nous avons vu qu'il est impossible, en général, d'estimer la capacité de généralisation d'un modèle à partir des résultats de l'apprentissage ; une telle procédure conduirait systématiquement à sélectionner un modèle de biais faible et de variance élevée, donc surajusté. Pour sélectionner le meilleur modèle parmi des modèles de complexités différentes, il convient donc de les comparer sur la base des prédictions qu'ils effectuent sur des données qui n'ont pas servi à l'apprentissage. Nous décrivons ci-dessous, dans la section intitulée « Sélection de modèles », les méthodes les plus couramment utilisées.

Sélection de modèles

Comme indiqué plus haut, la sélection de modèles est une étape cruciale dans la conception d'un modèle par apprentissage. Nous décrivons ici les trois méthodes les plus fréquemment mises en œuvre.

Validation simple (hold-out)

Lorsque l'on dispose d'un grand nombre de données, la méthode la plus simple consiste à diviser les données en trois ensembles (figure 1-17) :

- Un ensemble d'apprentissage, de taille N_A, utilisé pour l'apprentissage du modèle ; à l'issue de l'apprentissage, on calcule l'EQMA du modèle obtenu

$$EQMA = \sqrt{\frac{1}{N_A}\sum_{k=1}^{N_A}\left(y_k^p - g\left(\boldsymbol{x}_k, \boldsymbol{w}\right)\right)^2}$$

  où la somme porte sur les éléments de l'ensemble d'apprentissage.

- Un ensemble de validation de taille N_V, disjoint de l'ensemble d'apprentissage, mais issu de la même distribution de probabilité, qui est utilisé pour comparer les performances des modèles du point de vue de leur aptitude à généraliser. On calcule, pour chaque modèle, son Erreur Quadratique Moyenne de Validation (EQMV)

$$EQMV = \sqrt{\frac{1}{N_V}\sum_{k=1}^{N_V}\left(y_k^p - g\left(\boldsymbol{x}_k, \boldsymbol{w}\right)\right)^2}$$

  où la somme porte sur les éléments de la base de validation.
- Un ensemble de test de taille N_T, disjoint des deux précédents, qui sert à évaluer la performance du modèle sélectionné en calculant l'Erreur Quadratique Moyenne de Test (EQMT)

$$EQMT = \sqrt{\frac{1}{N_T}\sum_{k=1}^{N_T}\left(y_k^p - g\left(\boldsymbol{x}_k, \boldsymbol{w}\right)\right)^2}$$

  où la somme porte sur les éléments de la base de test ; ces données ne doivent évidemment pas être utilisées pendant toute la phase de sélection de modèle.

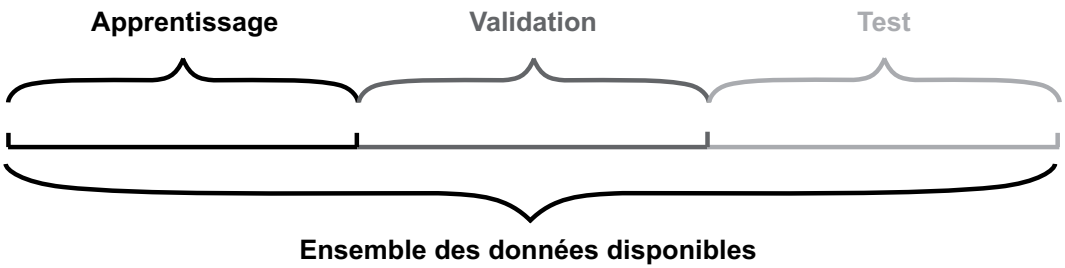


Figure 1-17. Validation simple

Parmi l'ensemble des modèles dont on a effectué l'apprentissage, on choisit évidemment celui dont l'EQMV est la plus petite ; si plusieurs modèles de complexités différentes peuvent prétendre à être choisis car leurs EQMV sont petites, et du même ordre de grandeur, on choisit celui dont la complexité est la plus faible.

Une fois déterminée la famille de fonctions de complexité optimale, on effectue un dernier apprentissage avec l'ensemble des données utilisées préalablement pour l'apprentissage et la validation ; la performance du modèle ainsi obtenu est estimée sur les données réservées pour le test.

Validation croisée (« cross-validation »)

Si l'on ne dispose pas de données abondantes, la validation simple risque de conduire à choisir des modèles surajustés à l'ensemble de validation. On utilise alors la validation croisée. Pour une famille de fonctions :

- séparer les données disponibles en un ensemble d'apprentissage-validation et un ensemble de test ;
- subdiviser le premier ensemble en D sous-ensembles disjoints (typiquement $D = 5$) ;
- itérer D fois, de telle manière que chaque exemple soit présent une et une seule fois dans un sous-ensemble de validation (figure 1-18) ;
  effectuer l'apprentissage sur D-1 sous-ensembles ; calculer la somme des carrés des erreurs sur le sous-ensemble des données restantes ;

$$S_i = \sum_{\substack{k \in \text{sous-ensemble} \\ \text{de validation } i}} \left(y_k^p - g\left(\boldsymbol{x}_k, \boldsymbol{w}_i\right)\right)^2$$

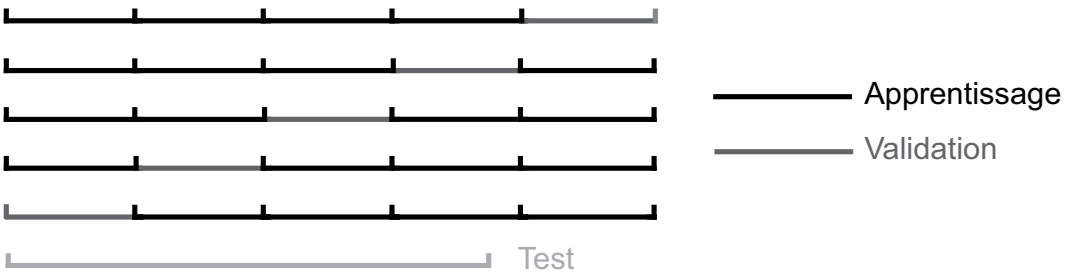


Figure 1-18. Validation croisée

- calculer le *score de validation croisée*

$$\sqrt{\frac{1}{N}\sum_{i=1}^{D} S_i}\ ;$$

- sélectionner le modèle dont le score de validation croisée est le plus faible ; si plusieurs modèles de complexités différentes peuvent prétendre à être choisis car leurs EQMV sont petites, et du même ordre de grandeur, choisir celui dont la complexité est la plus faible.

Une fois déterminée la famille de fonctions de complexité optimale, on effectue l'apprentissage sur l'ensemble des données utilisées préalablement pour la validation croisée, et la performance du modèle ainsi obtenu est estimée sur les données réservées pour le test.

Leave-one-out

Le leave-one-out (également appelé *jackknife*) est la limite de la validation croisée, dans laquelle le nombre de partitions D de l'ensemble d'apprentissage-validation est égal au nombre de ses éléments N. Chaque sous-ensemble de validation est donc constitué d'un seul exemple. Pour une famille de fonctions de complexité donnée, il faut donc réaliser autant d'apprentissages qu'il y a d'exemples dans la base d'apprentissage-validation. Pour chaque exemple k exclu de l'ensemble d'apprentissage, on calcule l'erreur de prédiction

$$r_k^{-k} = y_k^p - g(\boldsymbol{x}, \boldsymbol{w}^{-k})$$

où $g(\boldsymbol{x}, \boldsymbol{w}^{-k})$ désigne le modèle, de paramètres $\boldsymbol{w}^{-k}$, obtenu lorsque l'exemple k est exclu de l'ensemble d'apprentissage.

Une fois la procédure effectuée, on calcule le score de leave-one-out

$$E_t = \sqrt{\frac{1}{N}\sum_{k=1}^{N}\left(r_k^{-k}\right)^2}\ .$$

Comme dans les cas précédents, on choisit le modèle qui a le plus petit score de leave-one-out ; si plusieurs modèles de complexités différentes peuvent prétendre à être choisis car leurs scores de leave-

one-out sont petits, et du même ordre de grandeur, on choisit celui dont la complexité est la plus faible. L'apprentissage final est effectué avec l'ensemble des données disponibles.

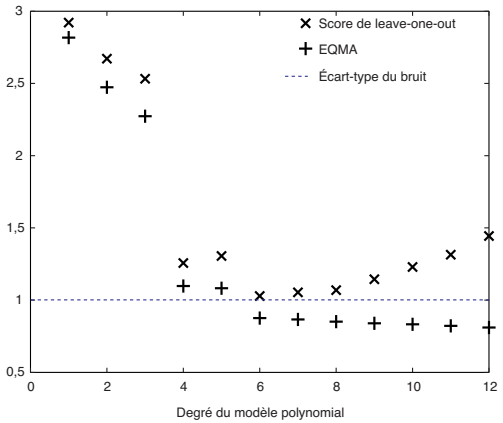


Figure 1-19. EQMA et score de leave-one-out moyens sur 100 bases d'apprentissage comprenant chacune 30 exemples

La figure 1-19 montre le score de leave-one-out et l'EQMA en fonction du degré du modèle polynomial, pour l'exemple étudié plus haut dans la section intitulée « Un exemple de modélisation pour la prédiction ». Les résultats sont remarquablement voisins de ceux qui sont représentés sur la figure 1-4 ; mais, à la différence de ces derniers, l'erreur de généralisation n'est pas estimée sur un ensemble de test de 1 000 exemples (il est tout à fait exceptionnel de disposer de données de test aussi abondantes), mais avec les seuls 30 points disponibles. La procédure conduit à la sélection d'un polynôme de degré 6 ; il faut noter que les résultats présentés ici sont une moyenne sur 100 ensembles d'apprentissage.

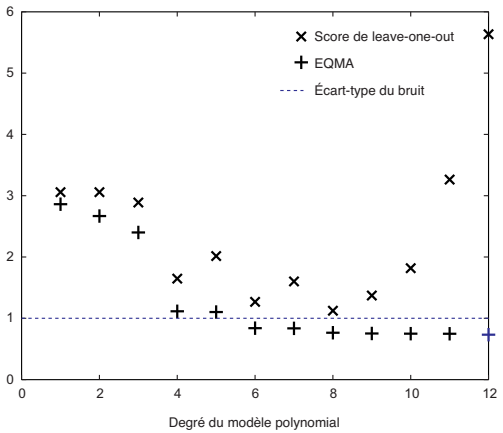


Figure 1-20. EQMA et score de leave-one-out pour un seul ensemble d'apprentissage

Dans la pratique, on ne dispose que d'un ensemble d'apprentissage, ce qui introduit une plus grande variabilité dans les résultats ; ainsi, dans le cas représenté sur la figure 1-20, les modèles de degré 6 et 8 peuvent prétendre à être choisis ; compte tenu du fait que les scores de leave-one-out sont très voisins, on choisit le modèle de degré 6.

Cette technique est donc gourmande en temps de calcul, en raison du grand nombre d'apprentissages nécessaires. Le calcul du PRESS, décrit dans la section « Conception de modèles linéaires » de ce chapitre, et la méthode du *leave-one-out virtuel*, qui sera décrite dans le chapitre 2, constituent des alternatives beaucoup plus économiques en temps de calcul.

Sélection de variables

Comme souligné plus haut, la sélection de variables constitue un élément important dans une stratégie de conception d'un modèle par apprentissage ; elle contribue en effet à la diminution de la complexité d'un modèle. L'ouvrage [GUYON 2006] fournit une excellente vue d'ensemble des approches modernes de la sélection de variables.

La sélection de variables nécessite toujours :
- de définir un critère de pertinence des variables pour la prédiction de la grandeur à modéliser ;
- de ranger les variables candidates par ordre de pertinence ;
- de définir un seuil qui permette de décider que l'on conserve ou que l'on rejette une variable ou un groupe de variables.

Dans cette section, nous poserons d'abord le problème de la définition d'un critère de pertinence sous son angle statistique, puis nous décrirons une méthode efficace de sélection de variables. Enfin, nous décrirons une stratégie générale à appliquer pour la sélection de variables.

Cadre théorique

Cette section pose les bases théoriques nécessaires à une appréhension générale du problème de sélection de variables. Le lecteur peu soucieux de ce cadre théorique peut sans dommage omettre la lecture de cette section et passer directement à la section intitulée « Méthode de la variable sonde ».

La présentation qui suit est inspirée de l'introduction de [GUYON 2006].

L'objectif de la sélection de variables est de discerner, dans un ensemble de variables candidates $\{x_1, x_2, \ldots x_n\}$, qui constituent le vecteur de variables que nous avons noté $\boldsymbol{x}$ dans les sections précédentes, celles qui sont pertinentes pour la modélisation de la grandeur y^p. Comme précédemment, ces variables peuvent être modélisées comme des réalisations des composantes $X_1, X_2, \ldots, X_n$ d'un vecteur aléatoire $\boldsymbol{X}$. On désigne par $\boldsymbol{X}^{-i}$ le vecteur dont les composantes sont celles de $\boldsymbol{X}$ à l'exception de la variable x_i. Enfin, on désigne par $\boldsymbol{S}^{-i}$ un vecteur aléatoire dont les composantes sont un sous-ensemble des composantes de $\boldsymbol{X}^{-i}$ ($\boldsymbol{S}^{-i}$ peut être identique à $\boldsymbol{X}^{-i}$). En résumé, le vecteur $\boldsymbol{X}$ modélise toutes les variables candidates, le vecteur $\boldsymbol{X}^{-i}$ modélise le vecteur des variables candidates dont on a supprimé la variable i, et le vecteur $\boldsymbol{S}^{-i}$ modélise le vecteur des variables candidates dont on a supprimé au moins la variable i, et éventuellement d'autres variables.

Il va de soi que la variable i est *certainement non pertinente* pour prédire la grandeur y^p si et seulement si les variables x_i et y^p varient indépendamment l'une de l'autre lorsque toutes les autres variables sont fixées, ce qui peut s'écrire :

$$p_{X_i,Y^p}\left(X_i, Y^p \middle| \boldsymbol{S}^{-i}\right) = p_{X_i}\left(X_i \middle| \boldsymbol{S}^{-i}\right) p_{Y^p}\left(Y^p \middle| \boldsymbol{S}^{-i}\right).$$

Une variable qui est pertinente n'obéit donc pas à cette relation. Pour savoir si une variable est peu pertinente ou très pertinente, il est donc naturel de chercher à savoir si le membre de gauche de cette égalité est peu différent, ou très différent, du membre de droite. S'agissant de distributions de probabilités, une « différence » s'exprime généralement par la *distance de Kullback-Leibler* entre les distributions. La distance de Kullback-Leibler entre deux distributions de probabilités p_U et p_V est définie par la relation [KULLBACK 1959] :

$$\int_{-\infty}^{+\infty} p_V \ln\left(\frac{p_U}{p_V}\right) du\ dv.$$

Elle s'écrit donc ici :

$$I\left(X_i, Y^p \middle| \boldsymbol{S}^{-i}\right) = \int_{-\infty}^{+\infty} p_{X_i,Y^p}\left(X_i, Y^p \middle| \boldsymbol{S}^{-i}\right) \ln\left(\frac{p_{X_i,Y^p}\left(X_i, Y^p \middle| \boldsymbol{S}^{-i}\right)}{p_{X_i}\left(X_i \middle| \boldsymbol{S}^{-i}\right) p_{Y^p}\left(Y^p \middle| \boldsymbol{S}^{-i}\right)}\right) dx_i dy^p.$$

Cette quantité n'est autre que *l'information mutuelle* entre X_i et Y^p, étant données toutes les autres variables. Plus elle est grande, plus la variable x_i est pertinente pour la prédiction de y^p, toutes les autres variables étant connues.

Puisque l'on cherche un indice de pertinence qui soit indépendant des autres variables candidates, il est naturel de proposer comme indice de pertinence, pour la variable i, la moyenne de l'information mutuelle :

$$r(i) = \sum_{S^{-i}} \Pr\left(S^{-i}\right) I\left(X_i, Y^p \middle| S^{-i}\right).$$

On peut alors fixer un seuil ε et décider de rejeter toutes les variables telles que

$$r(i) < \varepsilon.$$

Il faut néanmoins remarquer que les intégrales qui interviennent dans l'expression de l'indice de pertinence ne sont pas calculables, puisque l'on ne dispose que d'un nombre fini N de réalisations de x_i et de y^p. Ce critère de sélection n'est donc pas applicable en pratique ; en revanche, on peut, au moins en principe, estimer la probabilité pour que l'indice de pertinence soit supérieur à un seuil ε, et décider que la variable candidate doit être rejetée si la probabilité pour que son indice de pertinence soit supérieur au seuil est inférieure à une quantité δ :

$$\Pr\left(r(i,N) > \varepsilon\right) < \delta$$

où $r(i, N)$ désigne l'indice de pertinence estimé pour la variable i à partir d'un échantillon de N exemples.

Les méthodes qui nécessitent l'estimation de densités de probabilité sont généralement de mise en œuvre délicate, notamment lorsque les exemples sont en nombre limité. Nous décrivons ci-dessous une méthode simple et robuste qui est fondée sur l'estimation de corrélations.

Méthode de la variable sonde

Rappelons l'objectif de toute procédure de sélection de variables : classer les variables candidates en deux groupes, les variables que l'on conserve car on les considère pertinentes, et celles que l'on rejette. Supposons que l'on ait défini un indice de pertinence $r(i, N)$ pour la variable i, à partir d'un échantillon de N observations. La variable i étant modélisée comme une variable aléatoire, son indice de pertinence est lui-même une variable aléatoire. La figure 1-21 représente symboliquement les distributions de probabilité de l'indice de pertinence pour les variables pertinentes et pour les variables non pertinentes ; ces distributions sont évidemment inconnues, puisque l'on ne sait pas quelles variables sont pertinentes. Néanmoins, on peut penser que, si l'indice de pertinence est bien choisi, sa distribution, pour les variables pertinentes, possède un pic situé à des valeurs plus élevées que le pic de sa distribution pour les variables non pertinentes. Dans la pratique, les deux distributions ne sont pas parfaitement séparées : si l'on choisit un seuil ε comme indiqué sur la figure, il existe une probabilité non nulle de « faux positif » (probabilité de conserver une variable alors qu'elle n'est pas pertinente), et une probabilité non nulle de « faux négatif » (probabilité de rejeter une variable alors qu'elle est pertinente). Il faut donc choisir judicieusement ce seuil compte tenu des données dont on dispose.

À la fin de la section précédente, un critère de rejet a été proposé : rejeter la variable i si

$$\Pr\left(r(i,N) > \varepsilon\right) < \delta.$$

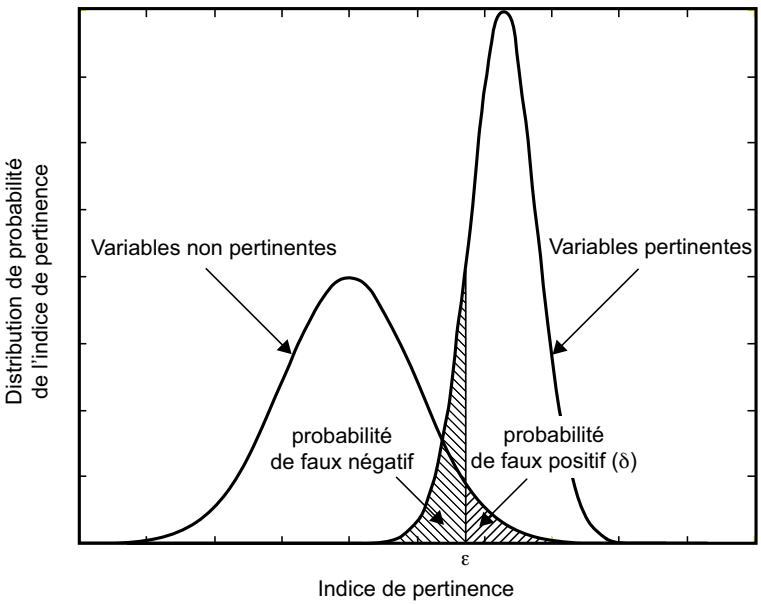


Figure 1-21. Distributions de probabilité de l'indice de pertinence pour des variables pertinentes et pour des variables non pertinentes

Cette condition se traduit sur la figure 1-21 par le fait que l'on choisit ε de telle manière que l'aire hachurée en traits gras soit inférieure à la probabilité δ que l'on s'est fixée.

Si l'on connaissait la distribution de l'indice de pertinence pour les variables non pertinentes, le seul paramètre que le concepteur du modèle aurait à choisir serait donc cette probabilité δ. L'intérêt de la méthode de la variable sonde est qu'elle permet d'estimer la densité de probabilité de l'indice de pertinence des variables non pertinentes. Muni de cette connaissance, on procède de la manière suivante : si les données sont peu nombreuses, on choisit δ « petit », donc ε « grand », car il convient d'être très sélectif afin de limiter le nombre de faux positifs. En revanche, si les données sont nombreuses, on peut se permettre de diminuer le seuil ε, donc de sélectionner un plus grand nombre de variables, au risque de conserver des variables non pertinentes.

Définition de l'indice de pertinence

Comme indiqué dans la section précédente (« cadre théorique »), un indice de pertinence peut naturellement être défini à partir de la notion d'information mutuelle, mais il est très difficile à estimer pratiquement, notamment dans le cas où de nombreuses variables sont candidates. Il est plus simple de définir un indice de pertinence à partir du coefficient de corrélation entre les variables candidates et la grandeur à modéliser, que celle-ci soit binaire (classification) ou réelle (régression).

Dans ce but, on se place dans le cadre de modèles linéaires en leurs paramètres

$$g(\boldsymbol{x},\boldsymbol{w}) = \sum_{i=1}^{p} w_i f_i(\boldsymbol{x}).$$

Dans cette expression, $f_i(\boldsymbol{x})$ peut être soit la variable x_i elle-même, qui est alors appelée « variable primaire », soit une fonction non paramétrée des variables, alors appelée « variable secondaire ». Pour simplifier, on désignera dans la suite par z_i la variable candidate de numéro i, qu'il s'agisse d'une variable primaire ou d'une variable secondaire :

$$g(\boldsymbol{x},\boldsymbol{w}) = \sum_{i=1}^{p} w_i z_i.$$

La figure 1-22 illustre la notion de variables primaire et secondaire, à l'aide d'un graphisme qui sera largement utilisé dans la suite de l'ouvrage. Les cercles représentent des fonctions ; le cercle contenant un signe Σ représente une fonction sommation. Les carrés ne réalisent aucune fonction : ils symbolisent simplement les variables du modèle. Le modèle représenté à gauche est un modèle linéaire en ses paramètres et en ses variables : les variables primaires et secondaires sont identiques. Le modèle de droite est un modèle linéaire en ses paramètres mais non linéaire en ses variables ; les variables secondaires sont obtenues à partir des variables primaires par des transformations non linéaires non paramétrées. Ainsi, le modèle de droite pourrait représenter un polynôme, les fonctions φ_i étant des monômes des variables primaires.

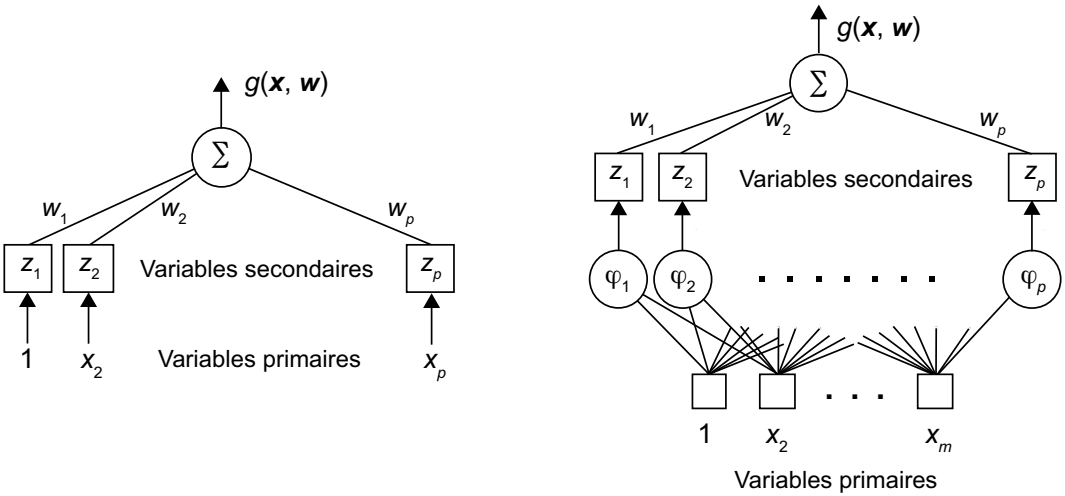


Figure 1-22. Modèles linéaires en leurs paramètres

Le carré du coefficient de corrélation entre deux variables aléatoires U et V centrées (de moyenne nulle), dont on connaît N réalisations, est estimé par la quantité

$$r_{U,V}^2 = \frac{\sum_{k=1}^{N} (u_k v_k)^2}{\sum_{k=1}^{N} u_k^2 \sum_{k=1}^{N} v_k^2}.$$

Cette quantité a une interprétation géométrique simple. Considérons *l'espace des observations*, de dimension N. Dans cet espace, la grandeur u est représentée par un vecteur $\boldsymbol{u}$, dont chaque composante est une observation u_k de u. Le carré du coefficient de corrélation est alors le carré du cosinus de l'angle θ_{uv} entre les vecteurs $\boldsymbol{u}$ et $\boldsymbol{v}$ dans cet espace :

$$r_{U,V}^2 = \cos^2 \theta_{uv} = \frac{(\boldsymbol{u} \cdot \boldsymbol{v})^2}{(\boldsymbol{u} \cdot \boldsymbol{u})(\boldsymbol{v} \cdot \boldsymbol{v})}$$

où le symbole · représente le produit scalaire dans l'espace des observations. Le coefficient de corrélation est donc compris entre zéro (observations non corrélées, vecteurs représentatifs orthogonaux dans l'espace des observations) et 1 (observations complètement corrélées, vecteurs représentatifs colinéaires).

Ainsi, le coefficient de corrélation entre la grandeur à modéliser y^p et la variable candidate z_i est donné par :

$$r^2_{Y^p,Z_i} = \frac{\left(\boldsymbol{y}_k^p \cdot \boldsymbol{z}_i\right)^2}{\left(\boldsymbol{y}_k^p \cdot \boldsymbol{y}_k^p\right)\left(\boldsymbol{z}_i \cdot \boldsymbol{z}_i\right)}$$

où $\boldsymbol{y}_k^p$ et $\boldsymbol{z}_i$ sont les vecteurs représentatifs, dans l'espace des observations, de la grandeur à modéliser et de la variable candidate de numéro *i* (primaire ou secondaire) respectivement.

Attention
Ne pas confondre *z* et z_i. Le vecteur z, qui intervient par exemple dans la notation du modèle $g(z, w)$, désigne le vecteur des variables du modèle : il est de dimension *p*. En revanche, le vecteur z_i représente la variable numéro *i* du modèle dans l'espace des observations : il est de dimension *N*, où *N* désigne le nombre d'observations.

À partir de ce coefficient de corrélation, l'indice de pertinence des variables candidates est défini comme le *rang de la variable candidate dans un classement établi par orthogonalisation de Gram-Schmidt* [CHEN 1989]. La procédure est la suivante :

- calculer les coefficients de corrélation entre y_k^p et les *p* variables candidates, et choisir la variable candidate z_i la plus corrélée à y_k^p ;
- projeter le vecteur $\boldsymbol{y}_k^p$ et toutes les variables non sélectionnées sur le sous-espace orthogonal à la variable z_i ;
- itérer dans ce sous-espace.

Les variables sont donc sélectionnées les unes après les autres. À chaque orthogonalisation, la contribution de la dernière variable sélectionnée au vecteur $\boldsymbol{y}_k^p$ est supprimée ; on obtient donc bien un classement des variables par ordre de pertinence décroissante. Il est alors naturel de considérer que le rang d'une variable dans ce classement est le reflet de la pertinence de cette variable par rapport à la modélisation que l'on cherche à effectuer.

La figure 1-23 illustre le processus dans un cas très simple où l'on aurait trois exemples ($N = 3$) et deux variables primaires ou secondaires candidates ($p = 2$), représentées par les vecteurs z_1 et z_2 dans l'espace des observations. La première étape a pour effet de sélectionner la variable z_1, car l'angle entre z_1 et $\boldsymbol{y}^P$ est plus petit que l'angle entre z_2 et $\boldsymbol{y}^p$. La deuxième étape consiste à projeter orthogonalement $\boldsymbol{y}^P$ et la variable non sélectionnée z_2 sur le sous-espace orthogonal à z_1. Toutes les variables candidates étant classées, le processus s'arrête alors. S'il y avait plus de deux variables candidates, le même processus serait itéré dans le sous-espace orthogonal à z_1.

Remarque 1
En pratique, il est préférable d'utiliser une variante de l'algorithme de Gram-Schmidt, appelée algorithme de Gram-Schmidt modifié, qui est plus stable numériquement [BJÖRCK 1967].

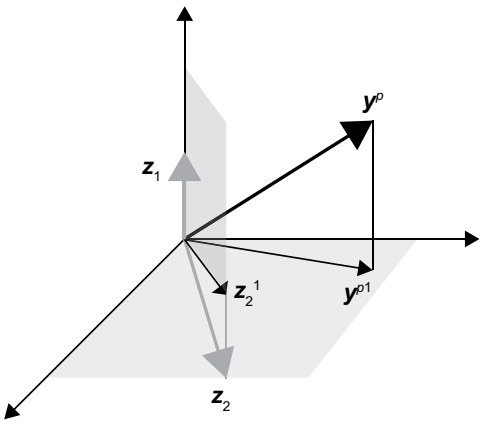


Figure 1-23. Orthogonalisation de Gram-Schmidt pour le classement de deux variables candidates dans un espace des observations de dimension trois

Remarque 2
L'algorithme d'orthogonalisation de Gram-Schmidt décrit ci-dessus est un cas particulier d'un algorithme d'apprentissage de modèles linéaires, décrit plus loin dans la section « Moindres carrés par orthogonalisation de Gram-Schmidt »

Exemple important
Pour illustrer l'importance de considérer les variables secondaires, et de ne pas se limiter aux variables primaires, considérons un problème simple de classification, illustré sur la figure 1-24.
On dispose de quatre exemples, appartenant à deux classes : la classe A, représentée par des croix, à laquelle on affecte l'étiquette $y^p = +1$, et la classe B, représentée par des cercles, à laquelle on affecte l'étiquette $y^p = -1$. Considérons comme variables candidates les variables primaires $z_1 = x_1$, $z_2 = x_2$, ainsi que la variable secondaire $z_3 = x_1 x_2$. Dans l'espace des observations, de dimension 4, les vecteurs représentatifs des variables candidates sont (les numéros des observations sont indiqués sur la figure 1-24)

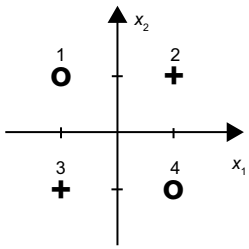


Figure 1-24. Exemple illustrant l'importance des variables secondaires

$$\boldsymbol{z}_1 = \begin{pmatrix} -1 \\ +1 \\ -1 \\ +1 \end{pmatrix} ; \boldsymbol{z}_2 = \begin{pmatrix} +1 \\ +1 \\ -1 \\ -1 \end{pmatrix} ; \boldsymbol{z}_3 = \begin{pmatrix} -1 \\ +1 \\ +1 \\ -1 \end{pmatrix}$$

et le vecteur représentatif de la grandeur à modéliser est

$$\boldsymbol{y}^p = \begin{pmatrix} -1 \\ +1 \\ +1 \\ -1 \end{pmatrix}.$$

Aucune des deux variables primaires, prise séparément, n'est pertinente pour la prédiction de y^p, puisque $\left(\boldsymbol{z}_1 \cdot \boldsymbol{y}^p\right)^2 = 0$ et $\left(\boldsymbol{z}_2 \cdot \boldsymbol{y}^p\right)^2 = 0$. En revanche, le coefficient de corrélation entre z_3 et y^p vaut 1. Par conséquent, la variable secondaire x_1x_2 détermine entièrement le modèle, alors que les variables primaires sont complètement inopérantes pour résoudre ce problème de classification (connu sous le nom de « problème du OU exclusif » ou « problème du XOR ») avec des modèles linéaires en leurs paramètres. Le modèle $g(x, w) = x_1x_2$ sépare complètement les exemples disponibles puisque $\text{sgn}(g(x, w)) = +1$ pour les exemples de la classe A et $\text{sgn}(g(x, w)) = -1$ pour ceux de la classe B. Il faut néanmoins remarquer que le problème peut être résolu avec comme variables x_1 et x_2 si l'on met en œuvre des modèles non linéaires en leurs paramètres, des réseaux de neurones par exemple.

Détermination du seuil de sélection des variables

Disposant d'une méthode de classement des variables candidates, il reste donc à déterminer le rang au-delà duquel les variables candidates doivent être rejetées. Comme indiqué dans la section « Cadre théorique », le problème serait simple si les distributions du rang des variables pertinentes et du rang des variables non pertinentes étaient connues. Ce n'est évidemment pas le cas, mais il est néanmoins possible d'estimer la distribution du rang des variables non pertinentes en créant artificiellement des variables non pertinentes ou « variables sondes ».

Présentation intuitive

Intuitivement, on pourrait envisager la procédure suivante :
- créer une variable sonde dont les « valeurs observées » seraient aléatoires, sans relation avec la grandeur à modéliser : cette variable est donc, par construction, non pertinente ;
- lors du classement par orthogonalisation de Gram-Schmidt, faire participer cette variable au même titre que les autres variables ;
- arrêter le classement des variables lorsque la variable sonde apparaît dans la procédure de classement : toutes les variables non encore classées sont alors rejetées, puisqu'elles sont moins pertinentes que la variable sonde qui, par construction, n'est pas pertinente.

Cette procédure est risquée : en effet, la décision de rejet est fondée sur le classement d'un seul vecteur représentatif de la variable sonde, donc d'une seule réalisation de ce vecteur aléatoire. Si l'on procédait à un autre tirage des valeurs de la variable sonde, on obtiendrait très probablement un autre rang, dans le classement, pour cette variable : on prendrait donc une autre décision de rejet. En d'autres termes, le rang de la variable sonde est lui-même une variable aléatoire, dont la distribution de probabilité est une estimation de la distribution de probabilité du rang des variables non pertinentes.

Présentation rigoureuse

Cette dernière remarque renvoie à la condition de rejet établie dans la section « Cadre théorique » : une variable candidate *i* est rejetée si

$$\Pr\left(r(i,N) > \varepsilon\right) < \delta$$

où $r(i, N)$ est l'indice de pertinence de la variable i, estimé à partir de N observations. Dans le cadre de la méthode de la variable sonde, l'indice de pertinence est le rang $\rho(i, N)$ de la variable candidate i ; la variable i est donc d'autant plus pertinente que son rang est petit. L'équation précédente s'écrit alors :

$$\Pr\left(\rho(i,N) < \rho_0\right) < \delta$$

où ρ_0 est le rang au-delà duquel les variables candidates doivent être rejetées. Or on souhaite que toutes les réalisations de la variable sonde soient rejetées ; l'application de la relation précédente aux variables sondes s'écrit donc :

$$\Pr\left(\rho_S < \rho_0\right) < \delta$$

où ρ_S désigne le rang d'une réalisation de la variable sonde. Ainsi, étant donnée une valeur de δ fixée, le seuil de rejet ρ_0 est le rang tel qu'une réalisation de la variable sonde soit classée au-dessus de ce rang avec une probabilité inférieure à δ, ou encore qu'une réalisation de la variable sonde ait une probabilité $1 - \delta$ d'être classée dans un rang au-delà de ρ_0. Cette situation est résumée sur la figure 1-25, où sont présentées la distribution hypothétique (puisque inconnue) du rang des variables pertinentes, et la distribution du rang de la variable sonde, qui constitue une estimation du rang des variables non pertinentes. Si l'on est prêt à admettre un risque de 10 % ($\delta = 0{,}1$) pour qu'une variable soit conservée alors qu'elle est aussi bien ou moins bien classée qu'une réalisation de la variable sonde (« risque de première espèce »), on lit, sur le graphe de la probabilité cumulée, qu'il faut rejeter toute variable de rang supérieur à 15. On peut noter que cette procé-

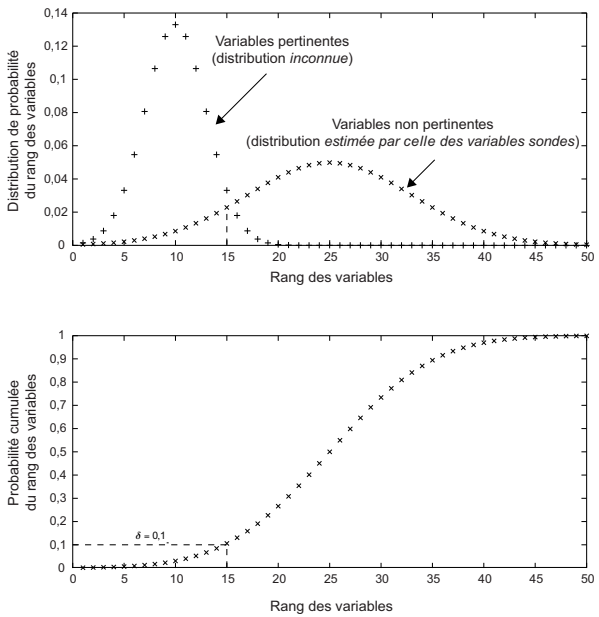


Figure 1-25. Choix du seuil de rejet des variables candidates

dure ne contrôle pas le risque de rejeter d'éventuelles variables pertinentes qui seraient classées au-delà du rang 15 (« risque de deuxième espèce ») ; on verra, dans la section intitulée « Limitations de la méthode », qu'il est néanmoins possible d'estimer ce risque, sans toutefois le contrôler.

En pratique, deux techniques sont utilisables pour engendrer les réalisations de la variable sonde :
- mélanger aléatoirement les observations des variables candidates ;
- tirer des nombres aléatoires dans une distribution de moyenne nulle et de variance 1, puisque les variables candidates ont été préalablement normalisées et centrées, comme indiqué plus haut dans la section « Prétraitement des données ».

Si les variables candidates obéissent à une distribution gaussienne, on peut légitimement considérer que la variable sonde est gaussienne. Alors, la probabilité cumulée du rang de la variable sonde peut être calculée analytiquement [STOPPIGLIA 2003], de sorte qu'il est inutile d'engendrer des réalisations de la variable sonde. On procède de la manière suivante : à chaque étape du classement par la méthode de Gram-Schmidt, on calcule la probabilité cumulée du rang de la variable sonde, et, lorsque celle-ci atteint la valeur δ choisie, on arrête le processus.

Si les variables n'obéissent pas à une distribution gaussienne, on estime la probabilité cumulée du rang de la variable sonde. Pour cela, on engendre un grand nombre de réalisations de la variable sonde, et l'on procède à l'orthogonalisation de Gram-Schmidt. Chaque fois qu'une réalisation de la variable sonde est rencontrée, on en prend note et l'on enlève cette variable du classement : on obtient ainsi une estimation empirique de la probabilité cumulée du rang de la variable sonde. Comme dans le cas précédent, on arrête le processus lorsque l'estimation de la probabilité cumulée atteint la valeur δ fixée à l'avance.

La figure 1-26 illustre cette approche à l'aide d'un exemple académique proposé dans [LAGARDE DE 1983] et repris dans [STOPPIGLIA 2003]. À partir d'un ensemble de 15 observations, on cherche à établir un modèle linéaire (en ses paramètres et en ses variables) avec 10 variables candidates, dont 5 seulement sont pertinentes : les coefficients des autres variables, dans la fonction linéaire génératrice des données, sont nuls. S'agissant d'un problème académique, les exemples ont été engendrés en ajoutant à une fonction linéaire un bruit gaussien centré ; les variables obéissent à une loi normale. L'objectif est de sélectionner les variables pertinentes. La figure 1-26 présente deux courbes : la probabilité cumulée du rang de la variable sonde *calculée* en supposant que la variable sonde obéit à une loi gaussienne, et la probabilité cumulée *estimée*, par la procédure décrite plus haut, à partir de 100 réalisations de la variable sonde, tirées d'une distribution gaussienne. On observe que, dans les deux cas, le choix d'un risque $\delta = 0{,}1$ conduit à sélectionner les 5 variables candidates les mieux classées, qui sont effectivement les 5 variables pertinentes à partir desquelles les données ont été engendrées.

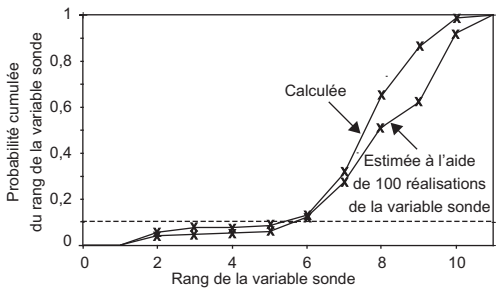


Figure 1-26. Probabilités cumulées, calculées et estimées

Limitations de la méthode

La principale limitation de la méthode de la variable sonde résulte de l'utilisation de l'algorithme de Gram-Schmidt, qui exige que le nombre de variables sélectionnées soit supérieur au nombre d'exemples. Il convient de noter que cette limitation porte sur le nombre de variables *sélectionnées*, et non sur le nombre de variables *candidates* : à l'aide de la méthode de la variable sonde, on peut traiter des problèmes où le nombre de variables candidates est très supérieur au nombre d'exemples.

D'autre part, la méthode contrôle directement le risque de faux positif, c'est-à-dire le risque de conserver une variable alors qu'elle n'est pas pertinente. Elle ne contrôle pas directement le risque de faux négatif, c'est-à-dire le risque de rejeter une variable alors qu'elle est pertinente. Néanmoins, il est possible de conserver également un contrôle sur ce phénomène en estimant le « taux de fausse découverte » (*false discovery rate* ou FDR), comme décrit dans [DREYFUS 2006].

Exemples académiques (classification)

Exemple 1

100 bases de données ont été construites de la manière suivante : pour chaque base, une fonction $g(\boldsymbol{x}, \boldsymbol{w})$ de deux variables a été choisie aléatoirement, 1 200 exemples ont été créés aléatoirement à partir de cette fonction en affectant à la classe A les exemples pour lesquels $\text{sgn}(g(\boldsymbol{x}, \boldsymbol{w})) = +1$. 10 % de ces exemples ont été affectés de manière erronée, de sorte qu'il y a 10 % d'erreur sur la base d'apprentissage. 800 exemples ont été utilisés pour l'apprentissage et 400 pour le test. Enfin, 238 variables non pertinentes ont été ajoutées à l'ensemble des variables, de sorte qu'il y a en tout 240 variables candidates, parmi lesquelles deux seulement sont pertinentes. La méthode décrite ci-dessus a été appliquée aux 240 variables candidates, et un classifieur a été réalisé à l'aide des deux premières variables sélectionnées. À titre de comparaison, un classifieur a été réalisé avec les deux « vraies » variables. Pour les 100 bases de données, la procédure a toujours trouvé au moins une des deux vraies variables, et a trouvé les deux vraies variables dans 74% des cas. Le tableau 1-2 résume les résultats moyens obtenus sur les 100 bases d'apprentissage.

Taux moyen d'erreurs de classification avec les variables sélectionnées	Taux moyen d'erreurs de classification avec les « vraies » variables	Hypothèse nulle : différence entre les taux d'erreurs moyens < 0,125
10,4% (écart-type 1,1%)	10,1% (écart-type 0,7%)	Acceptée

Tableau 1-2

On observe que le taux d'erreur de classification moyen (en moyenne sur les 100 bases de données), obtenu par un classifieur construit avec les descripteurs sélectionnés, est très voisin du taux d'erreur de classification obtenu par un classifieur établi avec les vraies variables. Un test d'hypothèse (voir la dernière section de ce chapitre) accepte l'hypothèse que la différence entre les taux d'erreurs moyens est inférieur à 0,125, c'est à dire à une erreur sur 800 ; en d'autres termes, la différence observée entre les taux d'erreurs des deux classifieurs n'est pas significative, puisque chaque base de données comprend 800 exemples d'apprentissage. Cela signifie que, lorsque la méthode n'a trouvé qu'une des deux vraies variables, l'autre variable sélectionnée permettait de discriminer les exemples de manière aussi précise que la vraie variable qui n'a pas été découverte. Les résultats sont semblables sur les bases de test.

À titre de comparaison, les taux d'erreurs sont d'environ 45 % si les deux variables sont choisies aléatoirement, et de 30 % si une des vraies variables est utilisée, l'autre variable étant choisie aléatoirement. Si l'on utilise un risque de 1% ($\delta = 0{,}1$), les trois premières variables du classement sont sélectionnées, ce qui ne dégrade pas les résultats de manière significative [STOPPIGLIA 2003].

Exemple 2

On construit 100 bases de données de 100 exemples tirés de distributions gaussiennes à deux variables x_1 et x_2, les centres étant dans les positions du problème du XOR (figure 1-24) ; 50 variables aléatoires non pertinentes sont ajoutées à l'ensemble des variables candidates. On utilise cette fois, outre les variables primaires, les monômes du second degré de celles-ci, ce qui produit en tout 1 326 variables candidates dont 52 variables indépendantes. Comme indiqué plus haut, la seule variable pertinente pour résoudre ce problème est le produit $x_1 x_2$; avec un risque de 1%, c'est effectivement la seule variable sélectionnée.

Variable sonde et test de Fisher

La méthode de la variable sonde est apparentée à l'utilisation de tests d'hypothèse pour la sélection de variables. Le lecteur qui n'est pas familier avec les tests d'hypothèses trouvera les concepts et définitions nécessaires dans la dernière section de ce chapitre.

Test de Fisher pour la sélection de variables

Comme précédemment, nous nous plaçons dans le cadre des modèles linéaires par rapport à leurs paramètres

$$g(\boldsymbol{x},\boldsymbol{w}) = \sum_{i=1}^{p} w_i z_i = \boldsymbol{w} \cdot \boldsymbol{z}$$

où les z_i sont les variables primaires ou secondaires.

On suppose que les mesures de la grandeur à modéliser peuvent être considérées comme les réalisations d'une variable aléatoire Y^p telle que $Y^p = \boldsymbol{w}^p \cdot \boldsymbol{z} + \boldsymbol{\Omega}$, où $\boldsymbol{z}$ est le vecteur des variables du modèle (de dimension inconnue), où $\boldsymbol{w^p}$ est le vecteur (non aléatoire mais inconnu) des paramètres du modèle, et où $\boldsymbol{\Omega}$ est une variable aléatoire gaussienne inconnue d'espérance mathématique nulle. On a donc :

$$E_{Y^p} = \boldsymbol{w}^p \cdot \boldsymbol{z}.$$

Nous cherchons à construire un modèle g, à partir d'un ensemble de N mesures $\{ y_k^p, k = 1 \text{ à } N\}$ qui constituent un ensemble de réalisations de la variable aléatoire Y^p ; nous désignons par $\boldsymbol{y}^p$ le vecteur, de dimension N, dont les composantes sont les $y_k{}^p$. Ce modèle dépend de l'ensemble des mesures utilisées pour sa construction : il est donc lui-même une réalisation d'une variable aléatoire G.

Supposons que l'on ait déterminé un ensemble de Q variables qui contient certainement les variables mesurables pertinentes pour la grandeur à modéliser. Un modèle contenant toutes les variables mesurables pertinentes est appelé *modèle complet*. On cherche alors un modèle de la forme

$$G_Q = \boldsymbol{W}^Q \cdot \boldsymbol{z}^Q$$

où $\boldsymbol{z}^Q$ est le vecteur des variables du modèle (de dimension Q+1 puisque, outre les variables pertinentes, le vecteur des variables contient une composante constante égale à 1) et où $\boldsymbol{W}$ est un vecteur aléatoire qui dépend de la réalisation du vecteur $\boldsymbol{Y}^p$ utilisée pour la construction du modèle. Rappelons que l'on dit que ce modèle complet est *vrai*, pour indiquer qu'il existe certainement une réalisation $\boldsymbol{w}^p$ du vecteur aléatoire $\boldsymbol{W}$ telle que $g_Q = E_{Y^p}$.

Supposons que l'apprentissage soit effectué par minimisation de la fonction de coût des moindres carrés

$$J(\boldsymbol{w}) = \sum_{k=1}^{N} \left(y_k^p - g_Q(\boldsymbol{z}_k, \boldsymbol{w}) \right)^2 = \left\| \left(\boldsymbol{y}^p - \boldsymbol{g}_Q(\boldsymbol{z}, \boldsymbol{w}) \right) \right\|^2,$$

où $\boldsymbol{w}$ désigne une réalisation du vecteur des paramètres $\boldsymbol{W}$, $\boldsymbol{z}^k$ est le vecteur des Q+1 variables pour l'exemple k, et où $\boldsymbol{g}_Q(\boldsymbol{z}, \boldsymbol{w})$ est le vecteur des valeurs des réalisations de G_Q pour les N mesures effectuées.

Soit $\boldsymbol{w}_{mc}^Q$ le vecteur des paramètres pour lequel la fonction de coût J est minimum. Le modèle obtenu est donc de la forme $g_Q = \boldsymbol{w}_{mc}^Q \cdot \boldsymbol{z}$, et l'on peut définir le vecteur $\boldsymbol{g}_Q = \boldsymbol{Z}\boldsymbol{w}_{mc}^Q$, où :

- $\boldsymbol{g}_Q$ est le vecteur dont les N composantes sont les prédictions du modèle pour chacune des N mesures effectuées ;

- $\boldsymbol{Z}$ est une matrice (dite *matrice des observations*) dont la colonne i (i = 1 à Q+1) est le vecteur $\boldsymbol{z}_i$ dont les composantes sont les N mesures de la variable numéro i : la matrice $\boldsymbol{Z}$ a donc N lignes et Q+1 colonnes :

$$\boldsymbol{Z} = \begin{pmatrix} z_{11} & \cdots & z_{1,\,Q+1} \\ z_{21} & \ddots & z_{2,\,Q+1} \\ \vdots & \ddots & \vdots \\ z_{N,\,1} & \cdots & z_{N,\,Q+1} \end{pmatrix}$$

où z_{ij} désigne la mesure numéro i de la variable candidate numéro j.

On se pose la question suivante : les Q variables du modèle complet sont-elles toutes pertinentes ? Pour répondre à cette question, on remarque que, si une variable n'est pas pertinente, le paramètre correspondant du modèle complet doit être égal à zéro. On appelle *sous-modèle* du modèle complet un modèle obtenu en mettant à zéro un ou plusieurs paramètres du modèle complet. Pour répondre à la question posée, il faut donc comparer le modèle complet à tous ses sous-modèles. Considérons un de ceux-ci, par exemple le modèle dont le vecteur $\boldsymbol{w}$ a ses q dernières composantes (numérotées de Q-q+2 à Q+1) égales à zéro : $\boldsymbol{g}_{Q-q} = \boldsymbol{Z}\boldsymbol{w}_{mc}^{Q-q}$, où $\boldsymbol{w}_{mc}^{Q-q}$ est le vecteur de paramètres obtenus en minimisant la fonction de coût des moindres carrés $J(\boldsymbol{w}) = \left\| \left(\boldsymbol{y}^p - \boldsymbol{g}_{Q-q}(z, \boldsymbol{w}) \right) \right\|^2$ sous la contrainte que les q dernières composantes du vecteur des paramètres soient nulles. On veut tester l'hypothèse nulle H_0 : les q derniers paramètres du vecteur aléatoire $\boldsymbol{W}$ sont nuls. Si cette hypothèse est vraie, la variable aléatoire

$$U = \frac{N-Q-1}{q} \frac{\left\| \boldsymbol{Y}^p - \boldsymbol{G}_{Q-q} \right\|^2 - \left\| \boldsymbol{Y}^p - \boldsymbol{G}_Q \right\|^2}{\left\| \boldsymbol{Y}^p - \boldsymbol{G}_Q \right\|^2} = \frac{N-Q-1}{q} \frac{\left\| \boldsymbol{G}_Q - \boldsymbol{G}_{Q\text{-}q} \right\|^2}{\left\| \boldsymbol{Y}^p - \boldsymbol{G}_Q \right\|^2}$$

est une variable de Fisher à q et N-Q-1 degrés de liberté.

> En effet, la quantité $\left\| \boldsymbol{Y}^p - \boldsymbol{G}_Q \right\|^2$ est la somme des carrés des composantes du vecteur $\boldsymbol{Y}^p - \boldsymbol{G}_Q$, dont on verra, dans la section consacrée à l'apprentissage des modèles linéaires par rapport à leurs paramètres, qu'il est orthogonal au sous-espace déterminé par les Q + 1 colonnes de la matrice Z. C'est donc la somme de $N - (Q + 1)$ carrés de variables aléatoires indépendantes gaussiennes : elle suit une distribution de Pearson à $N - Q - 1$ degrés de liberté. De même, le vecteur $\boldsymbol{G}_Q - \boldsymbol{G}_{Q-q}$ est dans un espace à q dimensions, donc le carré de sa norme est une somme des carrés de q variables aléatoires indépendantes : $\left\| \boldsymbol{G}_Q - \boldsymbol{G}_{Q-q} \right\|^2$ est donc une variable de Pearson à q degrés de liberté. Le rapport U de ces deux variables est donc une variable de Fisher, comme indiqué dans la section « Éléments de statistiques ».

Supposons que l'on dispose d'une très grande quantité de mesures ; si l'hypothèse nulle est vraie, le numérateur de U est très petit car le procédé de minimisation de la fonction de coût donne des valeurs nulles aux q paramètres « inutiles » du modèle complet, donc $\boldsymbol{g}_Q$ et $\boldsymbol{g}_{Q\text{-}q}$ sont très voisins. Si l'hypothèse nulle est fausse, les deux modèles ne peuvent pas être très voisins, même si le nombre de mesures est très grand, puisque le sous-modèle est trop pauvre pour rendre compte des données expérimentales. On comprend ainsi que la valeur de la réalisation de U doit être petite si l'hypothèse nulle est vraie.

Le test de Fisher consiste donc à choisir un risque α, et à trouver, en inversant la distribution de probabilité cumulée de Fisher, la valeur u_α telle que $\Pr(u < u_\alpha) = \alpha$. On calcule alors la quantité u (réalisation de la variable U avec les mesures disponibles) :

$$u = \frac{N-Q-1}{q} \frac{\left\| \boldsymbol{y}^p - \boldsymbol{g}_{Q-q}\left(\boldsymbol{z}, \boldsymbol{w}_{mc}^{Q-q}\right)\right\|^2 - \left\| \boldsymbol{y}^p - \boldsymbol{g}_{Q}\left(\boldsymbol{z}, \boldsymbol{w}_{mc}^{Q}\right)\right\|^2}{\left\| \boldsymbol{y}^p - \boldsymbol{g}_{Q}\left(\boldsymbol{z}, \boldsymbol{w}_{mc}^{Q}\right)\right\|^2}$$

et l'on accepte l'hypothèse nulle si et seulement si $u < u_{\alpha}$.

Test de Fisher et méthode de la variable sonde

On trouvera dans [STOPPIGLIA 2003] la démonstration du résultat suivant : si le modèle examiné à l'itération k du procédé d'orthogonalisation de Gram-Schmidt est un modèle *complet*, c'est-à-dire s'il contient toutes les variables pertinentes, et si le modèle complet est *vrai*, c'est-à-dire si la fonction de régression appartient à la famille des fonctions dans laquelle on recherche le modèle, alors l'opération de sélection effectuée à l'itération k est équivalente à un test de Fisher entre les modèles obtenus aux itérations k et k–1.

La méthode de la variable sonde présente donc deux avantages par rapport au test de Fisher : d'une part, elle donne une interprétation claire et intuitive du critère de sélection ; d'autre part, elle est applicable, que l'on dispose ou non d'un modèle complet, et que ce modèle soit vrai ou ne le soit pas.

Résumé : stratégies de conception

Dans cette section, nous montrons comment les différentes tâches à accomplir doivent être articulées entre elles pour concevoir un modèle par apprentissage (sélection de variables, apprentissage, sélection de modèles). On suppose que les étapes de collecte des données et de prétraitement de celles-ci ont été effectuées.

Une première stratégie peut être résumée de la façon suivante :
- Effectuer la sélection de variables sur l'ensemble des données disponibles.
- Effectuer l'apprentissage et la sélection de modèles de complexités différentes par validation croisée ou leave-one-out.
- Effectuer l'apprentissage du meilleur modèle avec toutes les données d'apprentissage et de validation.
- Tester le modèle sur un ensemble de tests.

Cette stratégie est simple et relativement peu coûteuse, mais elle n'est pas complètement rigoureuse dans la mesure où toutes les données disponibles sont utilisées pour la sélection de variables.

Pour être plus rigoureux, il convient de procéder de la façon suivante :
- Séparer les données en sous-ensembles d'apprentissage et de validation.
- Pour chaque sous-ensemble d'apprentissage
  - effectuer la sélection de variables, noter le nombre de variables sélectionnées,
  - effectuer l'apprentissage de modèles de complexités différentes et calculer les erreurs de validation.
- Calculer les scores de validation croisée et choisir le meilleur modèle ; soit n_0 le nombre de variables de ce modèle.
- Avec toutes les données utilisées pour l'apprentissage et la validation
  - effectuer le classement de variables par la méthode de Gram-Schmidt et choisir les n_0 variables les mieux classées,
  - avec ces variables, effectuer l'apprentissage du modèle qui a la meilleure complexité.
- Tester le modèle sur l'ensemble de test.

Si l'on n'est pas sûr que la valeur de δ choisie pour effectuer cette procédure est optimale, on peut ajouter une boucle extérieure portant sur différentes valeurs de δ.

Cette stratégie est applicable à toute méthode de sélection de variables fondée sur un classement des variables par ordre de pertinence.

Rappelons qu'il existe un grand nombre de méthodes de sélection de variables. La méthode de la variable sonde, décrite ici, a été présentée car elle est simple et robuste ; elle a été validée sur une grande variété d'applications ; néanmoins, il n'y a pas de méthode miracle, et dans certains cas, d'autres méthodes peuvent se révéler plus efficaces. Une synthèse très complète des méthodes modernes de sélection de variables est présentée dans l'ouvrage [GUYON 2006].

Conception de modèles linéaires par rapport à leurs paramètres (régression linéaire)

On a rappelé au début de ce chapitre le lien étroit qui existe entre apprentissage artificiel et statistiques. Avant même l'introduction du terme d'apprentissage, les statisticiens avaient largement développé la conception de modèles linéaires en leurs paramètres, ou *régression linéaire*. Il est donc important, dès ce chapitre introductif, de rappeler les méthodes de conception de modèles linéaires. De nombreux ouvrages sont entièrement consacrés à ce sujet (par exemple [SEBER 1977], [DRAPER 1998])

Rappelons qu'un modèle est dit « linéaire en ses paramètres », ou simplement « linéaire » s'il est de la forme :

$$g(\boldsymbol{x},\boldsymbol{w}) = \sum_{i=1}^{p} w_i f_i(\boldsymbol{x})$$

où les fonctions $f_i(\boldsymbol{x})$ sont des fonctions non paramétrées des variables (composantes du vecteur $\boldsymbol{x}$), dites *variables primaires*. Ces fonctions peuvent être considérées comme des *variables secondaires* z_i, de sorte que l'on écrira de manière générale un modèle linéaire en ses paramètres sous la forme

$$g(\boldsymbol{z},\boldsymbol{w}) = \sum_{i=1}^{p} w_i z_i$$

où les variables z_i peuvent être soit les variables primaires elles-mêmes, soit des variables secondaires déduites des variables primaires par une transformation non paramétrée (ou à paramètres fixés). On écrira aussi un tel modèle sous la forme

$$g(\boldsymbol{z},\boldsymbol{w}) = \boldsymbol{w} \cdot \boldsymbol{z}$$

où $\boldsymbol{w}$ et $\boldsymbol{z}$ sont des vecteurs de dimension p.

Sélection de variables pour les modèles linéaires en leurs paramètres

Ce problème a été abordé plus haut, dans la section consacrée à la sélection de modèles. Les méthodes décrites dans cette section sont directement applicables à la conception de modèles linéaires en leurs paramètres.

Apprentissage de modèles linéaires en leurs paramètres : la méthode des moindres carrés

Pour l'apprentissage des modèles linéaires en leurs paramètres, on choisit généralement comme fonction de perte le carré de l'erreur de modélisation

$$\pi\left[y^p, g(\boldsymbol{z},\boldsymbol{w})\right] = \left[y^p - g(\boldsymbol{z},\boldsymbol{w})\right]^2$$

de sorte que l'on cherche les paramètres pour lesquels la fonction de coût des moindres carrés $J(\boldsymbol{w})$ est minimum :

$$J(\boldsymbol{w}) = \sum_{k=1}^{N_A}\left(y_k^p - g(\boldsymbol{z}_k,\boldsymbol{w})\right)^2$$

où N_A est le nombre d'exemples de l'ensemble d'apprentissage, $\boldsymbol{z}_k$ est le vecteur des variables pour l'exemple k, et y_k^p est la valeur de la grandeur à modéliser pour l'exemple k.

Dans la section intitulée « Variable sonde et test de Fisher », on a défini la matrice des observations $\boldsymbol{Z}$, qui est une matrice à N lignes et p colonnes, dont l'élément z_{ij} est la valeur prise par la variable numéro j du modèle pour l'exemple i de l'ensemble d'apprentissage :

$$\boldsymbol{Z} = \begin{pmatrix} z_{11} & \cdots & z_{1,p} \\ z_{21} & \ddots & z_{2,p} \\ \vdots & \ddots & \vdots \\ z_{N,1} & \cdots & z_{N,p} \end{pmatrix}.$$

La fonction de coût peut alors se mettre sous la forme :

$$J(\boldsymbol{w}) = \left\|\boldsymbol{y}^p - \boldsymbol{g}(\boldsymbol{z},\boldsymbol{w})\right\|^2$$

où $\boldsymbol{y}^p$ est le vecteur dont les N composantes sont les valeurs de la grandeur à mesurer pour chacun des N exemples, et $\boldsymbol{g}(\boldsymbol{z}, \boldsymbol{w})$ est le vecteur dont les N composantes sont les prédictions du modèle pour chacun des exemples. Le vecteur $\boldsymbol{w}_{mc}$ est le vecteur pour lequel la fonction de coût est minimum :

$$\nabla_{\boldsymbol{w}} J = \left(\frac{dJ(\boldsymbol{w})}{d\boldsymbol{w}}\right)_{\boldsymbol{w}=\boldsymbol{w}_{mc}} = 0,$$

qui représente un ensemble de p équations, dont les p inconnues sont les paramètres w_i, $i = 1$ à p. Comme la fonction $J(\boldsymbol{w})$ est quadratique en fonction des w_i, sa dérivée par rapport à w_i est linéaire : il s'agit donc d'un système linéaire de p équations à p inconnues, appelées *équations canoniques*.

On montre facilement que cette équation s'écrit

$$\nabla_{\boldsymbol{w}} J = -2\boldsymbol{Z}^{\boldsymbol{T}}\left(\boldsymbol{y}^p - \boldsymbol{Z}\boldsymbol{w}_{mc}\right) = 0$$

où $\boldsymbol{Z}^{\boldsymbol{T}}$ désigne la transposée de la matrice $\boldsymbol{Z}$, soit encore

$$\boldsymbol{w}_{mc} = \left(\boldsymbol{Z}^{\boldsymbol{T}}\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}^{\boldsymbol{T}}\boldsymbol{y}^p.$$

Exemple

Considérons un modèle affine à une variable (p = 2) comme représenté sur la figure 1-27 :

$$g(\boldsymbol{x},\boldsymbol{w}) = w_1 + w_2 x.$$

Dans cet exemple, les points « expérimentaux » ont été obtenus en ajoutant à la fonction de régression $f(x) = 2 + 5\,x$ des réalisations d'une variable aléatoire gaussienne de moyenne nulle et d'écart-type égal à 3. Rappelons que, dans un problème réaliste, la fonction de régression est inconnue : l'objectif de l'apprentissage est de trouver un modèle qui soit aussi proche que possible de cette fonction inconnue.

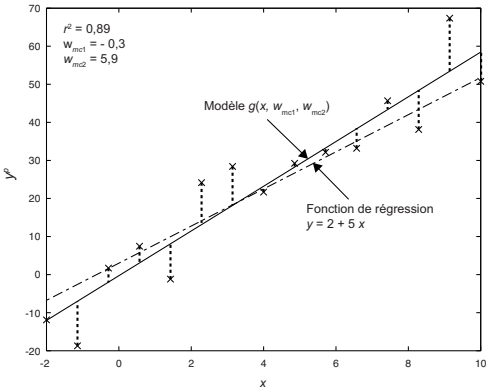


Figure 1-27. Points expérimentaux et modèle obtenu par la méthode des moindres carrés ; la somme des carrés des longueurs des segments en pointillés est minimale ; le coefficient de corrélation r^2 est défini ci-dessous, dans la section « Estimation de la qualité de l'apprentissage ».

La matrice des observations vaut $\boldsymbol{X} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_N \end{pmatrix}$, où x_i désigne la valeur prise par pour l'observation i de la variable x. On a alors :

$$\boldsymbol{X}^T\boldsymbol{X} = \begin{pmatrix} N & \sum_{k=1}^{N} x_k \\ \sum_{k=1}^{N} x_k & \sum_{k=1}^{N} (x_k)^2 \end{pmatrix}.$$

Par application de la relation $\boldsymbol{w}_{mc} = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{y}^p$, on trouve les paramètres du modèle affine :

$$w_{mc2} = \frac{N\sum_{k=1}^{N} x_k y_k^p - \sum_{k=1}^{N} x_k \sum_{k=1}^{N} y_k^p}{N\sum_{k=1}^{N} (x_k)^2 - \left(\sum_{k=1}^{N} x_k\right)^2} = \frac{\langle xy^p\rangle - \langle x\rangle\langle y^p\rangle}{\langle x^2\rangle - \langle x\rangle^2}$$

$$w_{mc1} = \frac{1}{N}\sum_{k=1}^{N} y_k^p - w_{mc2}\frac{1}{N}\sum_{k=1}^{N} x_k = \langle y^p\rangle - w_{mc2}\langle x\rangle$$

où $\langle u\rangle$ désigne la valeur moyenne de la grandeur u.

Remarque 1

La droite des moindres carrés passe par le centre de gravité des mesures.

En effet : $g(\langle x\rangle,\boldsymbol{w}) = w_{mc1} + w_{mc2}\langle x\rangle = \langle y^p\rangle - w_{mc2}\langle x\rangle + w_{mc2}\langle x\rangle = \langle y^p\rangle$.

Remarque 2

Si les données sont centrées ($\langle x\rangle = \langle y^p\rangle = 0$), la droite des moindres carrés passe par l'origine car $w_{mc1} = 0$. De plus : $w_{mc2} = \dfrac{\langle xy^p\rangle}{\langle x^2\rangle}$

Si, de plus, les données sont normalisées, on a en outre $\frac{1}{N}\sum_{k=1}^{N}(x - \langle x\rangle)^2 = 1 = \langle x^2\rangle$, par conséquent $w_{mc2} = \langle xy^p\rangle$.

Propriétés de la solution des moindres carrés

Un modèle obtenu par la méthode des moindres carrés possède des propriétés statistiques intéressantes qui justifient l'utilisation de la fonction de perte d'erreur quadratique, de préférence à d'autres fonctions de pertes envisageables telles que la valeur absolue de l'erreur.

Cas où le modèle est vrai

Supposons que le modèle linéaire postulé soit « vrai », c'est-à-dire que la fonction de régression inconnue appartienne effectivement à la famille des fonctions linéaires. Ce cas a déjà été rencontré plus haut (classification de deux ensembles d'observations issues de deux distributions gaussiennes de mêmes variances) ; le cas inverse a également été rencontré (modélisation de la fonction 10 sin$x$ / x par des polynômes). Les observations sont donc des réalisations de la variable aléatoire $Y^p = \boldsymbol{w}^p \cdot \boldsymbol{z} + \varepsilon$, avec $E_\varepsilon = 0$. En conséquence, $E_{Y^p} = \boldsymbol{w}^p \cdot \boldsymbol{z}$. Désignant par $\boldsymbol{Y}^p$ le vecteur des N observations, on a donc $E_{\boldsymbol{Y}^p} = \boldsymbol{Z}\boldsymbol{W}^p$.

Propriété

Le vecteur des paramètres w_{mc} trouvés par la méthode des moindres carrés est un estimateur non biaisé des paramètres w^p de la fonction de régression.

Démonstration
On a vu plus haut que $\boldsymbol{w}_{mc} = \left(\boldsymbol{Z}^T\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}^T\boldsymbol{y}^p$. Par conséquent : $E_{\boldsymbol{w}_{mc}} = \left(\boldsymbol{Z}^T\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}^T E_{\boldsymbol{y}^p} = \left(\boldsymbol{Z}^T\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}^T\boldsymbol{Z}\boldsymbol{W}^p = \boldsymbol{W}^p$, ce qui prouve la propriété.

Théorème de Gauss-Markov

Théorème

Les paramètres des modèles obtenus par minimisation de la fonction de coût des moindres carrés sont les paramètres de variance minimum.

Ainsi, dans la mesure où c'est l'augmentation de la variance qui produit le surajustement, la minimisation de la fonction de coût des moindres carrés permet de limiter le phénomène (sans toutefois le supprimer, bien entendu). L'expression de la variance des paramètres est établie plus loin, dans la section « Variance des paramètres d'un modèle linéaire ».

Cas où le bruit est gaussien

Si le bruit ε est gaussien, de variance σ^2, les estimations des paramètres obéissent à une loi gaussienne.

De plus, on démontrera, dans la section « Variance des paramètres d'un modèle linéaire », que la variance des paramètres vaut $\left(\boldsymbol{Z}^T\boldsymbol{Z}\right)^{-1}\sigma^2$ (quelle que soit la distribution de ε).

La figure 1-28 présente les histogrammes des paramètres w_{mc1} et w_{mc2} pour l'exemple considéré sur la figure 1-27. Ces histogrammes ont été obtenus en engendrant 100 ensembles d'apprentissage correspondant à 100 réalisations différentes du bruit, et en effectuant l'apprentissage de 100 modèles par la méthode des moindres carrés. On observe bien des distributions gaussiennes, centrées sur les valeurs des paramètres de la fonction de régression ($w_1^p = 2$, $w_2^p = 5$).

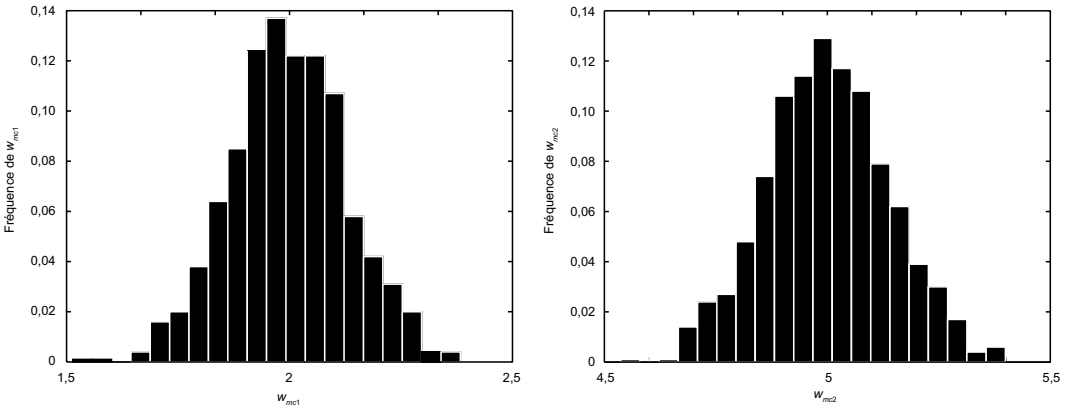


Figure 1-28. Distributions des paramètres d'un modèle linéaire avec bruit gaussien

Estimation de la qualité de l'apprentissage

La qualité d'un modèle linéaire est estimée par le coefficient de corrélation multiple r^2 entre les données et les prédictions.

Si U et V sont deux variables aléatoires, leur coefficient de corrélation $R_{U,V}$ est défini par

$$R_{U,V} = \frac{\mathrm{cov}_{U,V}}{\sigma_U \sigma_V} = \frac{E_{UV} - E_U E_V}{\sqrt{E_{U^2} - E_U^2}\sqrt{E_{V^2} - E_V^2}}$$

où $\mathrm{cov}_{U,V}$ désigne la covariance de U et V (voir la définition de la covariance de deux variables dans l'annexe « Éléments de statistiques » à la fin de ce chapitre).

Si U et V sont identiques, le coefficient de corrélation est une variable certaine qui vaut 1 ; si, au contraire, ces deux variables aléatoires sont indépendantes, le coefficient de corrélation vaut 0.

Comme cela a été fait à plusieurs reprises dans ce chapitre, considérons les données y^p et les prédictions du modèle comme des réalisations de variables aléatoires. On peut alors calculer une réalisation r de la variable R :

$$r = \frac{\sum_{k=1}^{N}\left(g(\boldsymbol{x},\boldsymbol{w}_{mc}) - \langle g(\boldsymbol{x},\boldsymbol{w}_{mc})\rangle\right)\left(y^p - \langle y^p\rangle\right)}{\sqrt{\sum_{k=1}^{N}\left(g(\boldsymbol{x},\boldsymbol{w}_{mc}) - \langle g(\boldsymbol{x},\boldsymbol{w}_{mc})\rangle\right)^2}\sqrt{\sum_{k=1}^{N}\left(y^p - \langle y^p\rangle\right)^2}} \quad (N >> 1).$$

Pour juger de la qualité du modèle, on utilise le *coefficient de détermination,* dont on démontre qu'il est une réalisation du *carré du coefficient de corrélation* entre les prédictions du modèle et les observations :

$$r^2 = \frac{\sum_{k=1}^{N}\left(g(x_k, w_{mc}) - \langle y^p\rangle\right)^2}{\sum_{k=1}^{N}\left(y_k^p - \langle y^p\rangle\right)^2}.$$

Si les variables sont centrées, cette expression se réduit à :

$$r^2 = \frac{\left(\left\langle xy^p \right\rangle\right)^2}{\left\langle x^2 \right\rangle \left\langle \left(y^p\right)^2 \right\rangle}.$$

Remarque
On retrouve ici la formule du carré du coefficient de corrélation introduit comme critère de pertinence dans la section « Sélection de variables » ; on trouve également dans cette section l'interprétation géométrique de ce coefficient.

Pour juger « visuellement » de la qualité d'un modèle, il est très commode d'utiliser son diagramme de dispersion, qui présente les valeurs prédites par le modèle en fonction des valeurs expérimentales correspondantes : les points de ce diagramme sont d'autant plus proches de la première bissectrice que la qualité de l'apprentissage est meilleure.

Remarque très importante
Rappelons qu'un apprentissage de très bonne qualité ne signifie pas que le modèle obtenu soit capable de généraliser correctement : un modèle qui a parfaitement appris les données d'apprentissage peut être surajusté, donc généraliser très mal. Il faut ainsi considérer le diagramme de dispersion sur les données d'apprentissage pour juger de la qualité de l'apprentissage, mais également le diagramme de dispersion sur des données non utilisées pour l'apprentissage, afin d'estimer la capacité de généralisation du modèle.

La figure 1-29 montre le diagramme de dispersion pour le modèle linéaire réalisé à partir des données d'apprentissage représentées sur la figure 1-27.

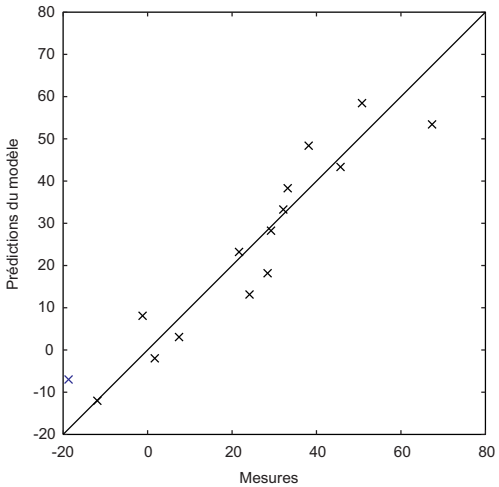


Figure 1-29. Diagramme de dispersion pour les données représentées sur la Figure 1-27.

Interprétation géométrique

La régression linéaire par la méthode des moindres carrés a une interprétation géométrique simple. Rappelons que le vecteur $\boldsymbol{w}_{mc}$ des paramètres du modèle

$$g(\boldsymbol{z},\boldsymbol{w}) = \sum_{i=1}^{p} w_i z_i = \boldsymbol{w} \cdot \boldsymbol{z}$$

est obtenu par la relation

$$\boldsymbol{w}_{mc} = \left(\boldsymbol{Z}^T\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}^T\boldsymbol{y}^p$$

où $\boldsymbol{Z}$ est la matrice des observations. Par conséquent, le vecteur $\boldsymbol{g}(\boldsymbol{z}, \boldsymbol{w}_{mc})$ des prédictions du modèle sur l'ensemble d'apprentissage est donné par

$$\boldsymbol{g}\left(\boldsymbol{z},\boldsymbol{w}_{mc}\right) = \boldsymbol{Z}\boldsymbol{w}_{mc} = \boldsymbol{Z}\left(\boldsymbol{Z}^T\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}^T\boldsymbol{y}^p$$

Or la matrice $\boldsymbol{Z}\left(\boldsymbol{Z}^T\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}^T$ (de dimensions N, N) n'est autre que la matrice de projection orthogonale sur les vecteurs colonnes de la matrice $\boldsymbol{Z}$. Le vecteur des prédictions du modèle sur l'ensemble d'apprentissage est donc la projection orthogonale du vecteur $\boldsymbol{y}^p$ sur le sous-espace de l'espace des observations défini par les vecteurs colonnes de la matrice des observations $\boldsymbol{Z}$. Ce dernier sous-espace est appelé « espace des estimations ».

Remarque
La matrice $\boldsymbol{Z}\left(\boldsymbol{Z}^T\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}^T$ est souvent appelée « matrice chapeau » et notée *H*. En effet, le vecteur des estimations effectuées par le modèle à partir des observations y est souvent noté $\hat{\boldsymbol{y}}$, donc $\hat{\boldsymbol{y}} = \boldsymbol{H}\boldsymbol{y}$: la matrice *H* est la matrice qui « met un chapeau » sur *y*.

L'interprétation géométrique de la méthode des moindres carrés est illustrée sur la figure 1-30, pour un modèle affine, dans le cas où l'espace des observations est de dimension 3. Dans cet espace, la matrice des observations a pour expression :

$$\boldsymbol{Z} = \begin{pmatrix} 1 & z_1 \\ 1 & z_2 \\ 1 & z_3 \end{pmatrix}.$$

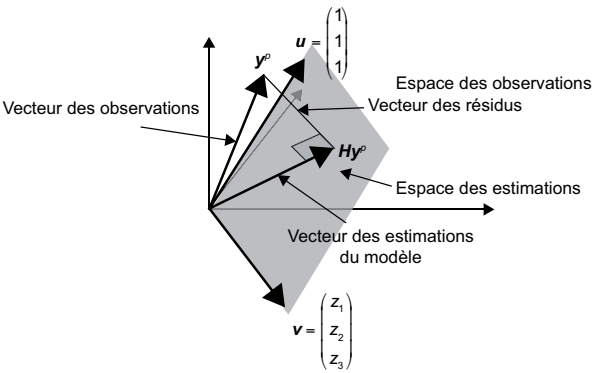


Figure 1-30. Méthode des moindres carrés : interprétation géométrique

L'espace des estimations est donc le sous-espace défini par les vecteurs colonnes de $\boldsymbol{Z}$, notés $\boldsymbol{u}$ et $\boldsymbol{v}$ respectivement. Le vecteur des prédictions du modèle pour l'ensemble d'apprentissage, ou vecteur des estimations, est la projection orthogonale du vecteur des observations $\boldsymbol{y}^p$ sur le sous-espace des estimations. Le vecteur des différences entre les mesures et les prédictions sur l'ensemble d'apprentissage est appelé vecteur des résidus. Le carré de son module est donc la somme des carrés des erreurs sur les éléments de l'ensemble d'apprentissage. De tous les vecteurs qui joignent l'extrémité de $\boldsymbol{y}^p$ à un point du sous-espace des estimations, c'est celui qui a le plus petit module.

Dilemme biais-variance pour les modèles linéaires

Dans les sections « Deux exemples académiques d'apprentissage supervisé » et « Dilemme biais-variance », on a constaté sur plusieurs exemples que, pour les modèles linéaires, ce dilemme est gouverné par le rapport du nombre de paramètres au nombre d'exemples. Ce résultat va maintenant être démontré de manière générale pour les modèles linéaires.

Variance des paramètres d'un modèle linéaire

Les paramètres d'un modèle linéaire obtenu par la méthode des moindres carrés sont donnés par la relation

$$\boldsymbol{w}_{mc} = \left(\boldsymbol{Z}^T\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}^T\boldsymbol{y}^p$$

où $\boldsymbol{Z}$ est la matrice des observations. Si l'on considère que les observations sont des réalisations de variables aléatoires, le vecteur des paramètres est lui-même une réalisation d'un vecteur aléatoire $\boldsymbol{W}_{mc} = \left(\boldsymbol{Z}^T\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}^T\boldsymbol{Y}^p$. Si les mesures de $\boldsymbol{y}^p$ sont indépendantes et de même variance σ^2, la variance du vecteur aléatoire $\boldsymbol{Y}^p$ est la matrice

$$\mathrm{var}_{\boldsymbol{Y}^p} = \boldsymbol{I}_{NN}\sigma^2.$$

où $\boldsymbol{I}_{NN}$ est la matrice identité de dimension *N*. La variance du vecteur des paramètres d'un modèle linéaire obtenu par la méthode des moindres carrés est donc :

$$\mathrm{var}_{\boldsymbol{W}_{mc}} = \left(\boldsymbol{Z}^T\boldsymbol{Z}\right)^{-1}\sigma^2.$$

Démonstration

D'après la propriété rappelée ci-dessous dans la section « variance d'un vecteur aléatoire », on a :

$$\text{var}_{\boldsymbol{W}_{mc}} = \left(\boldsymbol{Z}^T\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}^T\,\text{var}_{\boldsymbol{Y}^p}\left(\left(\boldsymbol{Z}^T\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}^T\right)^T = \left(\boldsymbol{Z}^T\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}^T\left(\left(\boldsymbol{Z}^T\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}^T\right)^T\sigma^2$$
$$= \left(\boldsymbol{Z}^T\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}^T\boldsymbol{Z}\left(\boldsymbol{Z}^T\boldsymbol{Z}\right)^{-1}\sigma^2 = \left(\boldsymbol{Z}^T\boldsymbol{Z}\right)^{-1}\sigma^2$$

Variance de la prédiction d'un modèle linéaire

On a vu, dans la section « Dilemme biais-variance », que l'erreur de prédiction théorique est donnée par la relation

$$P^2 = \sigma^2 + E_z\left[\text{var}\left[G(z,\boldsymbol{W})\right]\right] + E_z\left[E\left[f(z) - G(z,\boldsymbol{W})\right]\right]^2.$$

où $E_z(U)$ désigne l'espérance mathématique de la variable aléatoire U, considérée comme fonction du vecteur aléatoire z.

La prédiction du modèle au point z est ici $G\left(z,\boldsymbol{W}_{mc}\right) = z\cdot\boldsymbol{W}_{mc}$, qui peut s'écrire, sous forme matricielle : $G\left(z,\boldsymbol{W}_{mc}\right) = z^T\boldsymbol{W}_{mc}$. Par conséquent :

$$\text{var}\left(G\left(z,\boldsymbol{W}_{mc}\right)\right) = z^T\,\text{var}_{\boldsymbol{W}_{mc}}z = z^T\left(\boldsymbol{Z}^T\boldsymbol{Z}\right)^{-1}z\sigma^2.$$

Si les variables sont normalisées et centrées comme recommandé dans la section « Prétraitement des données », $\left(\boldsymbol{Z}^T\boldsymbol{Z}\right)^{-1} \approx \frac{1}{N}\boldsymbol{I}_{NN}$ si $p << N$, de sorte que $\text{var}\left(G\left(z,\boldsymbol{W}_{mc}\right)\right) \approx \frac{1}{N}z^Tz$.

D'autre part : $E_z\left(z^Tz\right) = E_z\left(\sum_{k=1}^{p} z_k^2\right) = \sum_{k=1}^{p}E_z\left(z_k^2\right) = \sum_{k=1}^{p}\left(E_z\left(z_k\right)\right)^2 + \sum_{k=1}^{p}\text{var}_{z_k}$. Les données étant supposées normalisées et centrées, le premier terme de la somme est nul, et le second est égal à p. Il reste donc :

$$E_z\left[\text{var}\left[G(z,\boldsymbol{W})\right]\right] = \frac{p}{N}.$$

Ainsi, on retrouve le fait que, lorsque l'on augmente le nombre de paramètres du modèle (par exemple en augmentant le degré du polynôme dans le cas d'un modèle polynomial) le terme de variance augmente. La figure 1-31 montre l'évolution de la variance en fonction du nombre de paramètres, pour l'exemple décrit dans la section « Un exemple de modélisation pour la prédiction », avec $N = 100$ exemples pour l'apprentissage, et des polynômes de degré 1 à 20. Comme pour les résultats présentés sur la figure 1-11, les espérances mathématiques portant sur Y^p sont estimées par les moyennes sur 100 ensembles d'apprentissage, et l'espérance mathématique portant sur z est estimée par une moyenne sur 1 000 points de test. On observe que la variance augmente linéairement avec le nombre de paramètres, la pente de la droite valant $1/N$, conformément à la relation démontrée ci-dessus.

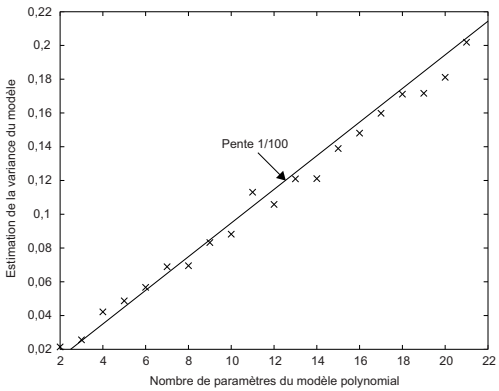


Figure 1-31. Variance d'un modèle polynomial en fonction du degré du polynôme (N = 100, p = 2 à 21)

Remarque

Dans l'exemple décrit par la figure 1-11, la variance (représentée par le symbole *x*) ne varie pas linéairement avec le degré du polynôme. Ceci est dû au fait que l'expression de la variance que l'on vient d'établir est vraie dans la limite des très grands ensembles d'apprentissage (*N* infini) ; pour *N* = 100 cette relation est raisonnablement bien vérifiée (figure 1-31) mais ce n'est pas le cas si *N* vaut seulement 15 (figure 1-11).

Sélection de modèles linéaires

La sélection de modèles linéaires peut être effectuée par les méthodes décrites dans la section intitulée « Sélection de modèles » : validation simple, validation croisée, leave-one-out. Cette dernière méthode est efficace mais gourmande en temps de calcul. On décrit ci-dessous une alternative intéressante au leave-one-out, qui est économe en temps de calcul : l'estimation du PRESS (Predicted REsidual Sum of Squares) pour les modèles linéaires, et le *leave-one-out virtuel* pour les modèles non linéaires.

Rappelons que le leave-one-out consiste à retirer un exemple k de l'ensemble des données disponibles, à effectuer l'apprentissage du modèle $g(\boldsymbol{z}, \boldsymbol{w}^{-k})$ avec toutes les autres données, et à calculer l'erreur de modélisation (ou *résidu*) sur l'exemple retiré des données :

$$r_k^{-k} = y_k^p - g(\boldsymbol{x}, \boldsymbol{w}^{-k}).$$

La procédure est itérée pour chaque exemple disponible, et le score de leave-one-out est calculé :

$$E_t = \sqrt{\frac{1}{N}\sum_{k=1}^{N}\left(r_k^{-k}\right)^2}\ .$$

Dans le cas de modèles linéaires, il est possible de calculer ce score de manière exacte, *en effectuant un seul apprentissage avec toutes les données disponibles*.

PRESS (Predicted REsidual Sum of Squares)

Montrons cette propriété dans le cas simple d'un modèle linéaire à un seul paramètre *w.* Dans ce cas, la matrice $\boldsymbol{Z}$ se réduit à un vecteur dont les composantes sont les N mesures z_i de la variable z, et la matrice $\boldsymbol{Z}^T\boldsymbol{Z}$ se réduit à un scalaire :

$$\left(\boldsymbol{Z}^T\boldsymbol{Z}\right)^{-1} = \frac{1}{\sum_{k=1}^{N}\left(z_k\right)^2}.$$

Si l'on effectue l'apprentissage avec les N exemples disponibles, le paramètre w_{mc} vaut alors :

$$w_{mc} = \left(\boldsymbol{Z}^T\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}^T\boldsymbol{y}^p = \frac{\sum_{k=1}^{N} z_k y_k^p}{\sum_{k=1}^{N} z_k^2}.$$

Supposons que l'on retire l'exemple i de l'ensemble des données disponibles, et que l'on effectue l'apprentissage avec tous les autres exemples. Le paramètre du modèle devient :

$$w_{mc}^{-i} = \frac{\sum_{\substack{k=1\\k\neq i}}^{N} z_k y_k^p}{\sum_{\substack{k=1\\k\neq i}}^{N} z_k^2} = \frac{\sum_{k=1}^{N} z_k y_k^p - z_i y_i^p}{\sum_{\substack{k=1\\k\neq i}}^{N} z_k^2}.$$

L'influence du retrait de l'exemple i sur le modèle se traduit donc par la variation de son unique paramètre :

$$w_{mc}^{-i} - w_{mc} = \frac{\sum_{k=1}^{N} z_k y_k^p - z_i y_i^p}{\sum_{\substack{k=1 \\ k\neq i}}^{N} z_k^2} - \frac{\sum_{k=1}^{N} z_k y_k^p}{\sum_{k=1}^{N} z_k^2} = -z_i \frac{r_i}{\sum_{\substack{k=1 \\ k\neq i}}^{N} z_k^2}$$

où r_i est le résidu (erreur de modélisation) sur l'exemple i *lorsque celui-ci est dans l'ensemble d'apprentissage* :

$$r_i = y_i^p - w_{mc} z_i = y_i^p - \frac{\sum_{k=1}^{N} z_k y_k^p}{\sum_{k=1}^{N} z_k^2} z_i \,.$$

Montrons à présent que l'on peut calculer l'erreur r_i^{-i} commise *lorsque l'exemple i a été retiré de l'ensemble d'apprentissage* en fonction de r_i :

$$r_i^{-i} - r_i = -\left(w_{mc}^{-i} - w_{mc}\right) z_i = z_i^2 \frac{r_i}{\sum_{\substack{k=1 \\ k\neq i}}^{N} z_k^2} = z_i^2 \frac{r_i}{\sum_{k=1}^{N} z_k^2 - z_i^2} \,,$$

et par conséquent :

$$r_i^{-i} = \frac{r_i}{1 - h_{ii}} \quad \text{avec} \quad h_{ii} = \frac{z_i^2}{\sum_{k=1}^{N} z_k^2} .$$

Cette relation rend donc inutile la réalisation de N apprentissages successifs, puisque l'on peut calculer exactement l'erreur de modélisation qui aurait été commise sur l'exemple i si celui-ci avait été retiré de l'ensemble d'apprentissage.

La quantité h_{ii} est appelée *levier* de l'exemple i, compris entre 0 et 1. Elle est présentée de manière plus détaillée dans la section suivante.

À partir de cette relation, on peut définir le PRESS (Predicted REsidual Sum of Squares) E_p, par analogie avec le score de leave-one-out E_t :

$$E_p = \sqrt{\frac{1}{N} \sum_{k=1}^{N} \left(\frac{r_i}{1 - h_{ii}} \right)^2} \,.$$

Dans le chapitre 2, une extension de ces résultats aux modèles non linéaires sera présentée sous le nom de « leave-one-out virtuel ».

Les leviers

Ce résultat peut être étendu au cas où le modèle possède p paramètres. Le levier de l'exemple i est alors l'élément diagonal i de la matrice chapeau

$$\boldsymbol{H} = \boldsymbol{Z}\left(\boldsymbol{Z}^T\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}^T.$$

Cette matrice étant une matrice de projection orthogonale, les leviers possèdent les propriétés suivantes (aisément vérifiées sur l'expression des leviers dans le cas d'un modèle à un seul paramètre, présenté dans la section précédente) :

$$0 < h_{ii} < 1 \; ; \; \sum_{i=1}^{N} h_{ii} = p.$$

Cette dernière relation fournit une interprétation intéressante des leviers : *le levier de l'exemple i est la proportion des paramètres qui est utilisée pour modéliser l'exemple i*. Ainsi, un exemple qui possède un grand levier a une grande importance pour le modèle : en d'autres termes, le modèle est très sensible au bruit présent sur la mesure de y^p pour l'exemple i. Il y a un risque de surajustement à l'exemple i.

Cet effet est illustré sur la figure 1-32.

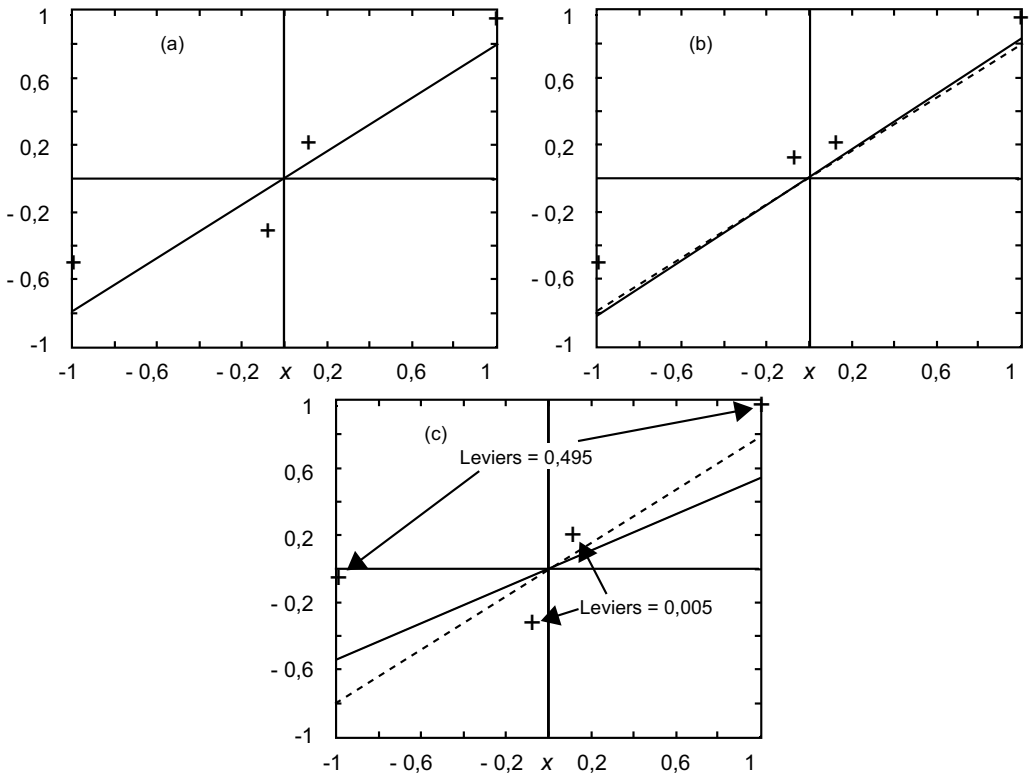


Figure 1-32. Interprétation des leviers

On dispose de 4 points expérimentaux, et l'on postule un modèle à un paramètre. La figure (a) montre le modèle linéaire ainsi obtenu. Supposons qu'une autre mesure effectuée en $x = -0{,}1$ donne un résultat différent, comme indiqué sur la figure (b) ; on obtient alors le modèle représenté en trait plein, très peu différent du modèle précédent, représenté en pointillé. Supposons en revanche que ce soit le point en $x = -1$ qui soit affecté (figure (c)). On obtient alors le modèle représenté en trait plein, très différent du modèle initial. On observe ainsi que le point situé en $x = -1$ a beaucoup plus d'influence sur le modèle que le point situé en $x = -0{,}1$. Ceci se traduit par des leviers de valeurs très différentes, dans un facteur à peu près égal à 100 : les points situés en $x = -1$ et $x = 1$ sont 100 fois plus importants pour le modèle que les points situés en $x = -0{,}1$ et $x = +0{,}1$. Les expériences qui ont été effectuées pour obtenir ces deux résultats étaient donc à peu près inutiles : il aurait été plus profitable de répéter les mesures en $x = -1$ et $x = +1$, afin de « moyenner » le bruit en ces points. On note que, conformément à ce qui a été indiqué plus haut, la somme des leviers est égale à 1, qui est le nombre de paramètres du modèle postulé.

Cette illustration numérique met en lumière l'intérêt des *plans d'expériences*, qui permettent de choisir les mesures les plus judicieuses pour établir un modèle prédictif précis.

Moindres carrés par orthogonalisation de Gram-Schmidt

Dans la section « Apprentissage de modèles linéaires en leurs paramètres », on a présenté une détermination algébrique du vecteur des paramètres pour lesquels la fonction de coût des moindres carrés est minimale, ainsi qu'une interprétation géométrique de ce résultat. La solution algébrique nécessite le calcul de l'inverse d'une matrice. La méthode d'orthogonalisation de Gram-Schmidt permet d'obtenir le même résultat de manière itérative, paramètre par paramètre ; elle est simple à comprendre dans le cadre de l'interprétation géométrique de la méthode des moindres carrés. Elle a déjà été rencontrée dans le cadre de la sélection de modèle, dans la section « Méthode de la variable sonde ».

On considère l'espace des observations, de dimension N, dans lequel la grandeur à modéliser est représentée par un vecteur $\boldsymbol{y}^p$, et chacune des variables est représentée par un vecteur $\boldsymbol{z}_i$, $i = 1$ à p ; rappelons que p est le nombre de paramètres du modèle et que N est le nombre d'observations de l'ensemble d'apprentissage. L'algorithme est une application simple du théorème des trois perpendiculaires :

- choisir une variable i représentée par le vecteur $\boldsymbol{z}_i$;
- projeter $\boldsymbol{y}^p$ sur la direction de $\boldsymbol{z}_i$, ce qui fournit le paramètre w_{mci} de la variable i : $w_{mci} = \dfrac{\boldsymbol{y}^p \cdot \boldsymbol{z}_i}{\|\boldsymbol{z}_i\|}$;
- projeter le vecteur des résidus $\boldsymbol{r}_i = \boldsymbol{y}^p - w_{mci}\boldsymbol{z}_i$, le vecteur $\boldsymbol{y}^p$, et tous les vecteurs $\boldsymbol{z}_{ji}$ sur le sous-espace orthogonal à $\boldsymbol{z}_i$;
- projeter la projection de $\boldsymbol{y}^p$ sur la projection d'un deuxième vecteur $\boldsymbol{z}_j$, ce qui fournit un deuxième paramètre du modèle ;
- itérer jusqu'à épuisement des variables du modèle.

La figure 1-33 présente l'algorithme dans le cas $N = 3$, $p = 2$. Les prédictions du modèle pour l'ensemble d'apprentissage sont représentées par $\boldsymbol{g}(\boldsymbol{z}, \boldsymbol{w})$, projection orthogonale de $\boldsymbol{y}^p$ sur l'espace des estimations, qui est donc une combinaison linéaire de $\boldsymbol{z}_1$ et $\boldsymbol{z}_2$. On peut obtenir ce vecteur en projetant d'abord sur un des vecteurs des variables (ici $\boldsymbol{z}_1$), puis en projetant orthogonalement $\boldsymbol{r}_1$ et $\boldsymbol{z}_2$ sur le sous-espace orthogonal à $\boldsymbol{z}_1$. Ce résultat s'obtient par application répétée du théorème des trois perpendiculaires.

Cet algorithme est celui qui est utilisé pour établir le classement des variables candidates en vue de la sélection de variables. La seule différence réside dans le fait que les projections ne se font pas dans n'importe quel ordre, mais en tenant compte des corrélations entre les vecteurs, comme indiqué dans la section « méthode de la variable sonde ».

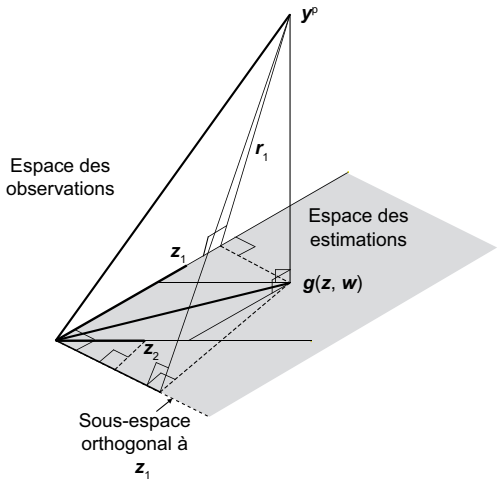


Figure 1-33. Moindres carrés par Gram-Schmidt

Éléments de statistiques

Cette introduction aux statistiques, à l'usage du lecteur peu familier avec celles-ci, termine ce chapitre introductif. Il existe de très nombreux ouvrages classiques (par exemple, [MOOD 1974], [WONNACOTT 1990]) auxquels le lecteur peut se référer pour plus de détails, notamment pour la démonstration de certains résultats.

Qu'est-ce qu'une variable aléatoire ?

Une variable aléatoire est une abstraction commode pour représenter une grandeur (par exemple, le résultat d'une mesure) lorsque sa valeur n'est pas certaine. On considère alors que la valeur de cette variable est la *réalisation* d'une variable aléatoire ; cette dernière est entièrement déterminée par sa « densité de probabilité » (ou simplement « densité », ou encore « distribution » ou « loi »).

Définition
Soit $p_Y(y)$ la densité de probabilité d'une variable aléatoire Y : la probabilité pour que la valeur d'une réalisation de Y soit comprise entre y et $y+dy$ vaut $p_Y(y)dy$.

Ainsi, si l'on traite une grandeur mesurable comme une variable aléatoire, *on fait comme si* le résultat de la mesure de cette grandeur était le résultat d'un tirage au sort dans un ensemble de valeurs possibles de y, avec la distribution (généralement inconnue) $p_Y(y)$. Utiliser une variable aléatoire pour modéliser le résultat d'une mesure ne signifie pas du tout que l'on considère la grandeur mesurée comme régie par des lois non déterministes : la variable aléatoire est un outil mathématique, dont l'utilisation est très commode lorsque les facteurs qui déterminent le résultat de la mesure ne sont pas connus, ou sont connus mais non maîtrisés ni mesurés.

Ainsi, le lancer d'un dé est un phénomène parfaitement déterministe, qui obéit à toutes les lois de la physique : si l'on connaissait la position initiale de la main du joueur, si l'on pouvait mesurer la vitesse initiale du dé, et si l'on connaissait les caractéristiques mécaniques de la matière dont sont constitués le dé et la table sur laquelle on le lance, on pourrait prédire exactement le résultat du lancer. Dans la pratique, comme toutes ces grandeurs ne sont pas connues et pas mesurées, il est commode de *modéliser* ce résultat comme la réalisation d'une variable aléatoire. Dans ce cas particulier, cette variable Y est une variable *discrète*, qui ne peut prendre que 6 valeurs, et, pour un dé non pipé, la probabilité de réalisation de chacune de ces valeurs est égale à 1/6.

De même, les méthodes statistiques sont susceptibles de prévoir les résultats d'une élection, alors que chaque citoyen ne vote pas au hasard, mais en fonction de ses convictions.

Propriété

La densité de probabilité $p_Y(y)$ est la dérivée première de la fonction de répartition ou probabilité cumulée : $p_Y(y) = \dfrac{dF_Y(y)}{dy}$ avec $F_Y(y) = \text{Probabilité}(Y \leq y)$.

Remarque

Toute réalisation y de la variable aléatoire Y étant comprise entre $-\infty$ et $+\infty$, on a évidemment $F_Y(-\infty) = 0$, $F_Y(+\infty) = 1$ et $\int_{-\infty}^{+\infty} p_Y(y)dy = 1$.

Variable certaine

Une variable certaine de valeur y_0 est une variable aléatoire dont la densité de probabilité est une distribution de Dirac $\delta(y - y_0)$.

Exemples de densités de probabilités (ou lois)

Densité de probabilité uniforme

Une variable aléatoire Y a une distribution uniforme si sa densité de probabilité vaut $p_Y(y) = 1/(b-a)$ sur un intervalle $[a, b]$, et est nulle partout ailleurs.

Densité de probabilité gaussienne

La distribution gaussienne $p_Y(y) = \dfrac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\dfrac{(y-\mu)^2}{2\sigma^2}\right)$ est très fréquemment utilisée. μ est la *moyenne* de la gaussienne et σ (>0) est son *écart-type*. La figure 1-34 représente une *distribution normale centrée réduite* (ou simplement *loi normale*), qui est une distribution gaussienne avec $\mu = 0$ et $\sigma = 1$. Les aires hachurées indiquent que la probabilité pour qu'une réalisation d'une variable suivant une loi normale soit comprise entre –1 et +1 vaut environ 0,68, et que la probabilité pour qu'elle soit entre –2 et +2 vaut environ 0,96.

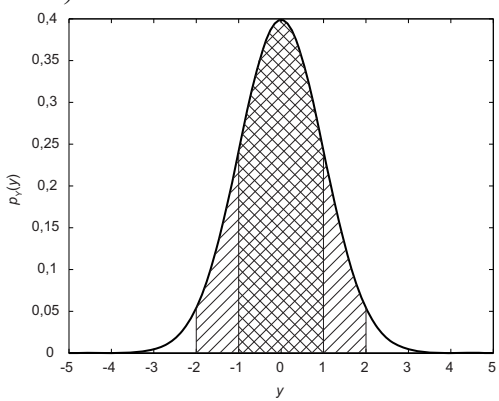


Figure 1-34. Loi normale

Autres densités de probabilité

Les distributions de Pearson (ou du χ^2), de Student et de Fisher sont présentées plus loin.

Densités de probabilités conjointes

Soit $p_{X,Y}(x,y)$ la densité de probabilité conjointe de deux variables aléatoires X et Y : la probabilité pour qu'une réalisation de X soit comprise entre x et x+dx et qu'une réalisation de Y soit comprise entre y et y+dy vaut $p_{X,Y}(x,y)\,dx\,dy$.

Variables aléatoires indépendantes

Deux variables aléatoires X et Y sont indépendantes si la probabilité de réalisation d'une des variables est indépendante de la probabilité de réalisation de l'autre. On a donc $p_{X,Y}(x,y) = p_X(x)\,p_Y(y)$.

Densités de probabilités conditionnelles

Soient deux variables aléatoires X et Y. La probabilité pour qu'une réalisation de la variable Y soit comprise entre y et $y+dy$ lorsque la variable X prend la valeur x est notée $p_Y(y|x)dy$, où $p_Y(y|x)$ est la densité de probabilité de y sachant x ou densité de probabilité conditionnelle de y. On a donc

$$p_{X,Y}(x,y) = p_Y(y|x)\,p_X(x) = p_X(x|y)\,p_Y(y)$$

Remarque :
Si les variables sont indépendantes : $p_Y(y|x) = p_Y(y)$ et $p_X(x|y) = p_X(x)$.

Vecteur aléatoire

Un vecteur aléatoire est un vecteur dont les composantes sont des variables aléatoires.

Espérance mathématique d'une variable aléatoire

Définition

L'espérance mathématique d'une variable aléatoire Y est $E_Y = \int_{-\infty}^{+\infty} y p_Y(y)\,dy$.

L'espérance mathématique d'une variable aléatoire est donc le premier moment de sa densité de probabilité.

Propriétés

Il est facile de démontrer les propriétés suivantes :

- l'espérance mathématique d'une somme de variables aléatoires est la somme des espérances mathématiques des variables aléatoires ;
- l'espérance mathématique du produit de deux variables *indépendantes* est égale au produit de leurs espérances mathématiques ;
- l'espérance mathématique d'une variable certaine de valeur y_0 est égale à y_0 ;
- si une variable Y obéit à une distribution uniforme sur un intervalle [a, b], son espérance mathématique vaut (a+b)/2 ;
- si une variable Y suit une loi gaussienne de moyenne μ, son espérance mathématique vaut μ.

Comme nous l'avons vu dans la section « Éléments de la théorie de l'apprentissage », l'objectif de tout apprentissage est d'obtenir une estimation fiable de l'espérance mathématique de la grandeur à modéliser. À cet effet, il est utile d'introduire le concept d'estimateur.

Estimateur non biaisé

Un estimateur est une variable aléatoire, fonction d'une ou plusieurs variables aléatoires *observables* ; une variable aléatoire est observable si ses réalisations sont mesurables.

Définition

Un estimateur H d'un paramètre de la distribution d'une variable aléatoire observable Y est dit « non biaisé » si son espérance mathématique E_H est égale à ce paramètre. Alors une réalisation de H constitue une estimation non biaisée du paramètre de la distribution.

Estimateur non biaisé d'une variable certaine

D'après la définition précédente, un estimateur d'une variable certaine est non biaisé si son espérance mathématique est égale la valeur de la variable certaine.

Ainsi, chercher à estimer les paramètres $\boldsymbol{w}$ d'un modèle, c'est-à-dire faire l'apprentissage d'un modèle, revient à chercher des estimateurs non biaisés des paramètres, ces derniers étant considérés comme des variables certaines. C'est cette approche, dite *fréquentiste*, qui est décrite dans le présent ouvrage. L'approche *bayesienne* qui considère les paramètres du modèle comme des variables aléatoires, permet également d'obtenir d'excellents résultats, comme décrit par exemple dans [NEAL 1996] ; la description de cette approche sort du cadre de cet ouvrage.

La moyenne est un estimateur non biaisé de l'espérance mathématique

Supposons que l'on ait effectué N mesures d'une grandeur Y, dans des conditions supposées identiques. On modélise cette grandeur par une variable aléatoire dont l'espérance mathématique E_Y est inconnue. Le résultat y_i de la mesure i peut être considéré comme une réalisation d'une variable aléatoire Y_i. Supposons que le résultat d'une mesure n'affecte pas les résultats des autres mesures, ce qui est raisonnable pour une expérience bien conçue : toutes ces variables aléatoires sont donc mutuellement indépendantes, et, puisque les mesures ont été effectuées dans des conditions identiques, elles ont des distributions de probabilité identiques ; elles ont donc notamment la même espérance mathématique E_Y.

Considérons la variable aléatoire $M = (Y_1 + Y_2 + \ldots + Y_N) / N$. Puisque l'espérance mathématique d'une somme de variables aléatoires est la somme des espérances mathématiques de ces variables, on a évidemment $E_M = E_Y$: l'espérance mathématique de la variable aléatoire M (appelée « moyenne ») est bien égale à l'espérance mathématique de la variable aléatoire Y. La grandeur $m = (y_1 + y_2 + \ldots + y_N) / N$, réalisation de l'estimateur de l'espérance mathématique de Y, constitue une estimation non biaisée de cette dernière.

Il reste à évaluer la qualité de cette estimation : le fait qu'elle soit non biaisée ne garantit pas qu'elle soit précise : sa précision dépend du nombre et de la « qualité » des mesures effectuées, c'est-à-dire de la dispersion des mesures autour de l'espérance mathématique. Pour caractériser numériquement cette dispersion, on utilise la notion de *variance.*

Variance d'une variable aléatoire

Définition

La variance d'une variable aléatoire Y de distribution $p_Y(y)$ est la quantité $\text{var}_Y = \sigma^2 = \int_{-\infty}^{+\infty} \left[y - E_Y\right]^2 p_Y(y)dy.$

La variance est donc le deuxième moment centré de la distribution de probabilité.

Remarque

La variance est également l'espérance mathématique de $\left[Y - E_Y\right]^2$: $\text{var}_Y = E_{(Y-E_Y)^2}$.

Propriétés

- Une variable certaine a une variance nulle.
- $\text{var}_Y = E_{Y^2} - \left(E_Y\right)^2$.
- $\text{var}_{aY} = a^2 \text{var}_Y$.
- Si une variable aléatoire obéit à une distribution uniforme sur un intervalle [a, b], sa variance vaut $(b–a)^2/12$.
- Si une variable aléatoire obéit à une loi gaussienne d'écart-type σ, sa variance vaut σ^2.

Estimateur non biaisé de la variance d'une variable aléatoire

Rappelons que, pour introduire l'estimateur moyenne M (estimateur non biaisé de l'espérance mathématique), on a considéré que N mesures, mutuellement indépendantes, d'une grandeur Y ont été effectuées, et elles ont été modélisées comme des réalisations de variables aléatoires Y_i de distributions identiques.

Estimateur non biaisé de la variance

La variable aléatoire $S^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(Y_i - M\right)^2$ est un estimateur non biaisé de la variance de Y.

Si l'on dispose de N résultats de mesures y_i, il faut donc, pour estimer la variance, calculer d'abord la valeur de la moyenne $m = \frac{1}{N}\sum_{i=1}^{N} y_i$, puis calculer l'estimation de la variance par la relation : $s^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(y_i - m\right)^2$.

L'estimation de la variance permet donc d'évaluer, de manière quantitative, la dispersion des résultats des mesures autour de leur moyenne. La moyenne étant elle-même une variable aléatoire, elle possède une variance : on pourrait effectuer plusieurs séries de mesures, calculer la moyenne de chacune de ces séries, puis estimer la variance de la moyenne, laquelle caractériserait la dispersion de l'estimation de la grandeur à modéliser. Néanmoins, cette procédure est lourde puisqu'elle requiert que l'on effectue plusieurs séries de mesures, dans des conditions supposées identiques.

Covariance de deux variables aléatoires

La covariance de deux variables aléatoires U et V est définie par :

$$\mathrm{cov}_{U,V} = E_{(U-E_U)(V-E_V)} = E_{UV} - E_U E_V.$$

Remarque

On a vu plus haut que

$\mathrm{var}_Y = E_{(Y-E_Y)^2}$.

La variance d'une variable aléatoire est donc la covariance de cette variable et d'elle-même.

Variance d'un vecteur aléatoire

Étant donné un vecteur aléatoire $\boldsymbol{U} = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_p \end{pmatrix}$, de dimension p, sa variance est la matrice (p, p) donnée par :

$$\mathrm{var}_U = \begin{pmatrix} \mathrm{var}_{u_1} & \mathrm{cov}_{u_1,u_2} & \cdots & \mathrm{cov}_{u_1,u_p} \\ \mathrm{cov}_{u_1,u_2} & \mathrm{var}_{u_2} & \cdots & \cdots \\ \vdots & \vdots & \ddots & \vdots \\ \cdots & \cdots & \cdots & \mathrm{var}_{u_p} \end{pmatrix}.$$

Propriété
Si $\boldsymbol{A}$ est une matrice certaine : $\mathrm{var}_{\boldsymbol{AU}} = \boldsymbol{A}\,\mathrm{var}_{\boldsymbol{U}}\,\boldsymbol{A}^T$.

Autres distributions utiles

Loi de χ^2 (ou de Pearson)

Si une variable aléatoire X est la somme des carrés de N variables gaussiennes indépendantes, elle obéit à une loi de χ^2 (ou de Pearson) à N degrés de liberté. Alors $E_X = N$ et $\mathrm{var}_X = 2N$.

Loi de Student

Si Y_1 est une variable de distribution normale, et si Y_2 est une variable aléatoire, indépendante de Y_1, obéissant à une loi de Pearson à N degrés de liberté, alors la variable aléatoire $Z = \dfrac{Y_1}{\sqrt{Y_2 / N}}$ obéit à une loi de Student à N degrés de liberté.

Loi de Fisher

Si Y_1 est une variable aléatoire de Pearson à N_1 degrés de liberté, et si Y_2 est une variable aléatoire de Pearson à N_2 degrés de liberté, alors la variable aléatoire $Z = \dfrac{Y_1 / N_1}{Y_2 / N_2}$ obéit à une loi de Fisher à N_1 et N_2 degrés de liberté.

Intervalles de confiance

Dans les sections précédentes, nous avons vu que l'estimation d'une grandeur dépend à la fois du nombre d'expériences et de la variabilité des observations. On peut combiner élégamment la taille de l'échantillon et sa variabilité pour évaluer la différence qui peut exister entre l'estimation d'une grandeur et sa « vraie » valeur.

Définition

Un intervalle de confiance, au seuil de confiance $1 - \alpha$, pour une variable aléatoire *Y*, est un intervalle qui, avec une probabilité $1 - \alpha$, contient la valeur de l'espérance mathématique de *Y*.

En conséquence, plus l'intervalle de confiance est petit, plus on peut avoir confiance en l'estimation de la grandeur à modéliser.

Ainsi, supposons que l'on ait réalisé 100 ensembles de mesures ; à partir de celles-ci, on peut calculer 100 moyennes, 100 estimations de la variance, et 100 intervalles de confiance à 95 % ($\alpha = 0{,}05$). Alors, pour 95 % de ces ensembles de données, l'intervalle de confiance contient la moyenne ; on ne peut évidemment pas garantir que, pour un ensemble particulier de mesures, la vraie valeur soit à l'intérieur de l'intervalle de confiance calculé à partir de cet ensemble de mesures.

Conception d'un intervalle de confiance

Pour concevoir un intervalle de confiance pour une variable aléatoire *Y*, il faut trouver une variable aléatoire *Z*, fonction de *Y*, dont la distribution $p_Z(z)$ soit *connue* et *indépendante de Y*. Puisque la distribution $p_Z(z)$ est connue, il est facile de résoudre l'équation $\Pr(z_1 < z < z_2) = \int_{z_1}^{z_2} p_Z(z)dz = 1 - \alpha$: il suffit d'inverser la fonction de répartition de *Z*, c'est-à-dire trouver la valeur z_1 de z telle que $\Pr(z < z_1) = \alpha / 2$, et la valeur z_2 de z telle que $\Pr(z > z_2) = \alpha / 2$. Une fois déterminées les valeurs de z_1 et de z_2, on inverse la fonction $Z(Y)$ afin de trouver les valeurs a et b de y telles que $\Pr(a < y < b) = 1 - \alpha$.

Exemple : conception d'un intervalle de confiance pour la moyenne

Le tout premier exemple d'apprentissage qui a été considéré dans ce chapitre consistait en l'estimation de l'unique paramètre w d'un modèle constant ; on a vu que ce paramètre n'était autre que l'espérance mathématique de la grandeur à modéliser. On a également vu que la moyenne est un estimateur non biaisé de l'espérance mathématique. On se pose donc la question suivante : étant donné un ensemble de mesures d'une grandeur, dont on a calculé la moyenne pour estimer son espérance mathématique, quelle confiance peut-on accorder à cette estimation ?

Supposons donc, comme précédemment, que *N* expériences ont été effectuées, et que l'on peut modéliser les résultats de ces expériences comme *N* réalisations de variables aléatoires Y_i indépendantes et de même distribution. De plus, supposons que la distribution commune à ces variables est une distribution gaussienne de moyenne μ et de variance σ^2.

Il est facile de démontrer que la somme de N variables gaussiennes indépendantes est une variable gaussienne dont la moyenne est la somme des moyennes, et dont la variance est la somme des variances. Ici les distributions des N variables sont identiques, dont la moyenne est une gaussienne de moyenne $N\mu$ et de variance $N\sigma^2$. Leur moyenne M obéit donc à une loi gaussienne de moyenne μ et de variance σ^2/N ; par conséquent la variable aléatoire $\dfrac{M-\mu}{\sigma/\sqrt{N}}$ obéit à une loi normale (gaussienne de moyenne nulle et de variance unité).

Rappelons que l'on cherche à établir deux bornes pour l'espérance mathématique μ, qui doivent être de la forme $m \pm a$, où m est la moyenne des mesures et a le demi-intervalle de confiance. On peut prévoir que l'intervalle de confiance croît avec la variance des mesures et décroît avec leur nombre. Comme indiqué plus haut, l'estimateur non biaisé de la variance est la variable aléatoire $S^2 = \dfrac{1}{N-1}\sum_{i=1}^{N}(Y_i - M)^2$. Il est commode de normaliser cette variable en la divisant par son espérance mathématique σ^2 ; les variables Y_i étant supposées gaussiennes, la variable aléatoire M est également gaussienne, donc $(N-1)\,S^2/\sigma^2$ est la somme de $N-1$ variables gaussiennes indépendantes (il n'y a que $N-1$ variables indépendantes puisque M dépend des Y_i) ; elle obéit donc à une loi de Pearson.

D'autre part, comme indiqué plus haut, la variable aléatoire $\dfrac{M-\mu}{\sigma/\sqrt{N}}$ obéit à une loi normale.

Par conséquent, la variable aléatoire $Z = \dfrac{\dfrac{M-\mu}{\sigma/\sqrt{N}}}{\sqrt{S^2/\sigma^2}} = \dfrac{M-\mu}{\sqrt{S^2/N}}$ obéit à une loi de Student à $N-1$ degrés de liberté. La distribution de Student étant symétrique, il suffit alors de chercher la valeur de z_0 telle qu'une variable de Student soit comprise entre $-z_0$ et $+z_0$ avec la probabilité $1-\alpha$, soit encore telle qu'une variable de Student soit comprise entre $-\infty$ et z_0 avec la probabilité $\alpha/2$. À partir des résultats expérimentaux, on peut calculer une réalisation m de M, une réalisation s de S, et une réalisation z de Z par les relations $m = \dfrac{1}{N}\sum_{i=1}^{N} y_i$, $s = \sqrt{\dfrac{1}{N-1}\sum_{i=1}^{N}(y_i - m)^2}$ et $z = \dfrac{m-\mu}{\sqrt{s^2/N}}$. Avec une probabilité $1-\alpha$, l'estimation m de μ se trouve à l'intérieur de l'intervalle de confiance si z est dans l'intervalle $[-z_0, +z_0]$:

$$-z_0 < \frac{m-\mu}{\sqrt{s^2/N}} < +z_0$$

soit

$$m - z_0\sqrt{s^2/N} < \mu < m + z_0\sqrt{s^2/N}\,.$$

L'intervalle de confiance recherché est donc l'intervalle centré sur l'estimation de la moyenne m, et de demi-largeur $z_0\sqrt{s^2/N}$.

La figure 1-35 représente l'inverse de la distribution de probabilité cumulée d'une variable de Student, pour différentes valeurs de N. On observe que, au-delà de $N = 10$, la distribution devient à peu près indépendante de N (elle est d'ailleurs très voisine d'une distribution normale) ; pour un niveau de confiance de 0,95, on voit que $z_0 \approx 2$ pour $N \geq 10$, de sorte que la largeur de l'intervalle de confiance pour est à peu près $2\sqrt{s^2 / N} = 2s / \sqrt{N}$. La largeur de l'intervalle de confiance est donc proportionnelle à s, donc au bruit de mesure, et inversement proportionnelle à la racine carrée du nombre d'exemples : une grande variabilité dans les mesures doit être compensée par une grande taille de l'échantillon.

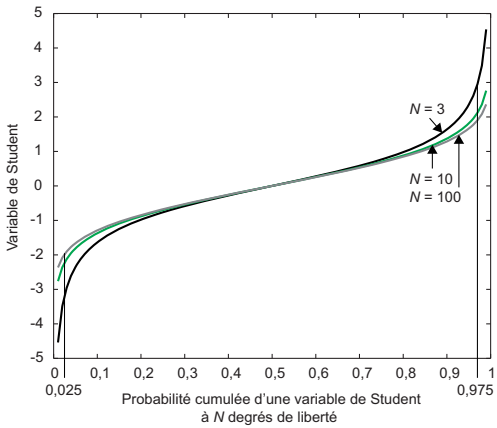


Figure 1-35. Inverse de la probabilité cumulée d'une variable de Student

À titre d'exemple, on a simulé 10 000 séries de 100 mesures en engendrant des réalisations d'une variable aléatoire selon une loi normale. Pour chaque série de mesures, la moyenne, l'estimateur de la variance, et l'intervalle de confiance déterminé ci-dessus, au niveau de confiance 0,95 ont été calculés : dans 95,7% des cas, l'espérance mathématique des « mesures » (égale à zéro) se trouve bien à l'intérieur de l'intervalle de confiance.

On a donc établi ici un intervalle de confiance pour l'estimation de l'espérance mathématique, ou, en d'autres termes, de l'unique paramètre d'un modèle constant. Il est très important de pouvoir fournir un intervalle de confiance sur les prédictions fournies par un modèle. On en rencontrera de nombreux exemples dans cet ouvrage.

Tests d'hypothèse

On a vu plus haut que des étapes importantes dans la conception d'un modèle par apprentissage artificiel, telles que la sélection de variables ou la sélection de modèles, nécessitent de prendre des décisions (sélectionner ou rejeter un modèle ou une variable) à partir des informations disponibles, qui sont généralement en nombre limité. Il faut donc prendre ces décisions de manière raisonnée. Les tests d'hypothèse sont les outils appropriés pour ce genre de situation. Ils permettent de faire une hypothèse et d'établir une des deux conclusions suivantes, avec un risque d'erreur fixé :

- les données confirment cette hypothèse,
- le fait que les données semblent confirmer cette hypothèse est simplement le résultat d'un concours de circonstances improbable, lié à la petite taille de l'échantillon et à la variabilité des mesures.

De nombreux tests d'hypothèses, adaptés à une grande variété de situations, ont été proposés (voir par exemple [LEHMANN 1993]).

Le principe d'un test d'hypothèse est le suivant : pour tester la validité d'une hypothèse (appelée « hypothèse nulle » et traditionnellement notée H_0), on cherche à établir l'expression d'une variable aléatoire qui suit une loi connue si l'hypothèse nulle est vraie, et dont on peut calculer une réalisation à partir des données disponibles. Si la probabilité pour que cette réalisation se trouve dans un intervalle donné est « trop faible », on considère que la probabilité pour que l'hypothèse nulle soit vraie est trop faible : on la rejette donc.

À titre de première illustration, supposons qu'un modèle prédise que la grandeur à modéliser, par exemple l'unique paramètre d'un modèle constant, a une certaine valeur w_0. On dispose d'un ensemble de N observations de cette grandeur, et l'on veut savoir si elles confirment l'hypothèse selon laquelle la grandeur a pour « vraie » valeur w_0. Ces mesures sont modélisées comme des réalisations de N variables aléatoires Y_i supposées gaussiennes, d'espérance mathématique μ et de variance σ^2. L'hypothèse nulle est donc $H_0 : w_0 = \mu$, et l'hypothèse alternative est $w_0 \neq \mu$.

Nous avons vu dans la section précédente que, si l'hypothèse nulle est vraie, c'est-à-dire si $w_0 = \mu$, la variable aléatoire $Z = \dfrac{M - w_0}{\sqrt{S^2 / N}}$, obéit à une loi de Student à $N - 1$ degrés de liberté (M est l'estimateur de l'espérance mathématique, S^2 est l'estimateur de la variance). À partir des N données disponibles, on peut calculer une réalisation z de cette variable aléatoire. D'autre part on peut calculer la valeur z_0 telle que la probabilité pour qu'une réalisation de la variable aléatoire soit à l'extérieur de l'intervalle $[- z_0, +z_0]$ est égale au risque choisi $1 - \alpha$. Si la réalisation observée z est à l'extérieur de cet intervalle, on peut considérer que les données ne confirment pas de manière significative l'hypothèse H_0 ; on rejette donc celle-ci, avec un risque $1 - \alpha$ de se tromper. En outre, il faut définir le niveau de risque d'erreur, noté $1 - \alpha$, que l'on est disposé à admettre, l'erreur consistant à rejeter l'hypothèse nulle alors quelle est vraie (erreur de type 1).

Supposons par exemple qu'une théorie prévoie qu'une grandeur vaut $w_0 = 1$. Supposons que l'on dispose de 100 mesures de cette grandeur, dont la moyenne m vaut 2 et l'écart-type vaut $s = 10$: ces mesures sont donc très dispersées autour de la moyenne. On se pose la question : ces données confirment-elles l'hypothèse selon laquelle w_0 vaut 1 ? La réalisation de la variable aléatoire z vaut

$$z = \frac{m - w_0}{\sqrt{s^2 / N}} = 1.$$

En se reportant à la figure 1-35, on voit que $z_0 \approx 2$ (pour $\alpha = 0{,}95$), de sorte que z est dans l'intervalle $[- z_0, +z_0]$. On accepte donc l'hypothèse nulle au vu des données disponibles. À l'inverse, si les données disponibles ont toujours pour moyenne $m = 2$, mais avec une dispersion beaucoup plus petite, par exemple $s = 3$, alors $z = 3{,}3$; dans ces conditions, on est amené à rejeter l'hypothèse nulle.

La « certitude » avec laquelle on accepte l'hypothèse nulle est exprimée par la « p-valeur » de la réalisation z de la variable aléatoire Z. C'est la probabilité pour qu'une réalisation de Z soit à l'extérieur de l'intervalle $\left[-|z|, +|z|\right]$ si l'hypothèse nulle est vraie : la p-valeur de z_0 est donc $1 - \alpha$. Ainsi, dans l'exemple précédent, la p-valeur de $z = 1$ vaut 0,32, ce qui signifie que l'on est raisonnablement sûr de ne pas se tromper en acceptant l'hypothèse nulle (figure 1-36). En revanche, la p-valeur de $z = 3{,}3$ vaut $8\ 10^{-3}$: accepter l'hypothèse nulle serait donc extrêmement risqué.

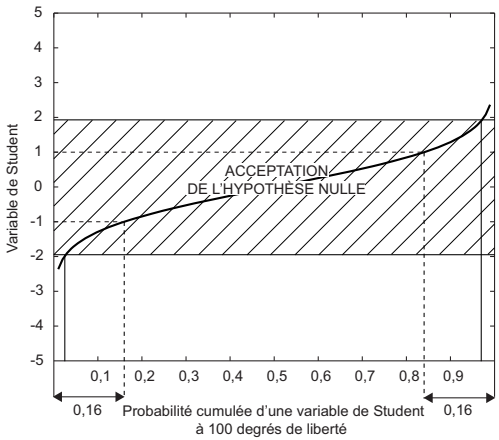


Figure 1-36. p-valeur de z = 1

Notons que la p-valeur de $z = 0$ vaut 1, ce qui veut dire que l'on accepte l'hypothèse nulle avec la plus grande certitude possible ; c'est naturel, puisque $z = 0$ correspond au cas où la moyenne est égale à la valeur postulée de l'espérance mathématique.

Remarque

Dans ce cas particulier, le test d'hypothèse consiste à regarder si la valeur de la moyenne dont on fait l'hypothèse se trouve dans l'intervalle de confiance calculé au paragraphe précédent, et à rejeter l'hypothèse nulle si cette valeur est à l'extérieur de cet intervalle.

Un autre exemple de test d'hypothèses (test de Fisher) est décrit dans la section « Sélection de variables ».

Conclusion

Dans ce chapitre, les fondements de l'apprentissage statistique et de sa mise en œuvre ont été décrits de manière succincte ; on en trouvera une présentation beaucoup plus détaillée dans [HASTIE 2001] par exemple. Pendant longtemps, les efforts de recherche en apprentissage artificiel ont porté essentiellement sur les familles de modèles et les algorithmes d'apprentissage. Le nombre et la variété des applications, leur difficulté et leur exigence croissantes, ont rendu nécessaires la mise en place d'un corps de doctrine et d'une méthodologie qui englobent tous les aspects de la conception de modèle par apprentissage statistique : sélection de variables, sélection de modèle, planification d'expériences, estimation d'intervalles de confiance sur les prédictions, sont au moins aussi importantes que l'apprentissage lui-même. Les méthodes qui ont été décrites ou esquissées dans ce chapitre peuvent être mises en œuvre pour la plupart des grandes familles de modèles. Les chapitres suivants de cet ouvrage sont consacrés à différents types de modèles – réseaux de neurones, cartes auto-organisatrices, machines à vecteurs supports – dont on montrera les spécificités, la mise en œuvre, et les applications.

Bibliographie

BJÖRCK A. [1967], Solving linear least squares problems by Gram-Schmidt orthogonalization. *BIT*, 7, p. 1-27.

CHEN S., BILLINGS S. A., LUO W. [1989], Orthogonal least squares methods and their application to non-linear system identification, *International Journal of Control*, 50, p. 1873-1896.

DRAPER N. R., SMITH H. [1998], *Applied regression analysis,* John Wiley & Sons.

DREYFUS G., GUYON I. [2006], Assessment Methods, in *Feature Extraction, Foundations and Applications*, I. Guyon, S. Gunn, M. Nikraveh, L. Zadeh, eds. (Springer), p. 65-88.

GUYON I., GUNN S., NIKRAVESH M., ZADEH L. [2006], *Feature Extraction, Foundations and Applications*, Springer.

HASTIE T, TIBSHIRANI R., FRIEDMAN J. [2001], *The elements of statistical learning, data mining, inference and predictions*, Springer.

KULLBACK S. [1959], *Information Theory and Statistics,* Dover Publications.

LAGARDE DE J. [1983], *Initiation à l'analyse des données*, Dunod, Paris.

LEHMANN E. L. [1993], *Testing statistical hypotheses*, Chapman & Hall.

MOOD A. M., GRAYBILL F. A., BOES D. C. [1974], *Introduction to the Theory of Statistics*, McGraw-Hill.

NEAL R. M. [1996] *Bayesian Learning for Neural Networks*, Springer.

SEBER G. A. F. [1977], *Linear Regression Analysis*, Wiley

STOPPIGLIA H. [1997], *Méthodes statistiques de sélection de modèles neuronaux ; applications financières et bancaires*, Thèse de Doctorat de l'Université Pierre et Marie Curie, Paris. *Disponible sur le site http://www.neurones.espci.fr.*

STOPPIGLIA H., DREYFUS G., DUBOIS R., OUSSAR Y. [2003], Ranking a Random Feature for Variable and Feature Selection, *Journal of Machine Learning Research,* p. 1399-1414.

VAPNIK V. [1998], *The nature of statistical learning theory,* Springer.

WONNACOTT T. H., WONNACOTT R. J. [1990], *Statistique économie-gestion-sciences-médecine*, Economica, 4e édition, 1990.