Analyse discriminante et Régression logistique

L3 Mathématiques, parcours Statistique

Université Bretagne Sud

Salim LARDJANE

Maître de Conférences en Statistique

Table des matières

1 Introduction

L'Analyse Discriminante, également appelée *Reconnaissance de formes* est une approche mathématique d'une activité fondamentale de la cognition humaine.
L'être humain reconnaît sans peine les visages, comprend les paroles, lit les caractères écrits, retrouve ses clefs dans sa poche au toucher et décide si une pomme est bien mûre en se basant sur son odeur.
Les processus derrière cette capacité à reconnaître des formes déjà connues, c'est-à-dire à *discriminer* entre différentes possibilités au vu d'indices de types divers, sont complexes.
Les méthodes d'Analyse Discriminante qui seront présentées dans ce cours se basent sur une sorte particulière d'indices : les données statistiques.
Elles seront décrites dans le langage des mathématiques mais admettront souvent une interprétation intuitive.
L'Analyse Discriminante conçue ainsi se prête bien à une implémentation sur machine.
On cherche à concevoir et construire des machines capables de discriminer pour remplir différents types de tâches : reconnaissance automatique de la parole, identification d'empreintes digitales, reconnaissance optique de caractères, identification de séquences d'ADN, entre autres.
De plus, en mettant au point de tels systèmes, on acquiert des connaissance sur les systèmes de reconnaissance de forme existant dans la nature et plus particulièrement chez les êtres humains.
En retour, la connaissance que nous avons des systèmes de discrimination naturels peut nous inspirer des algorithmes d'Analyse Discriminante.

2 Théorie de la Décision

Considérons l'expérience aléatoire consistant à prélever au hasard un individu (objet) dans une population donnée.
On sait qu'on peut modéliser une telle expérience aléatoire par un *espace probabilisé*

$(\Omega, \mathcal{A}, \mathbb{P})$ où Ω est l'ensemble des résultats possibles de l'expérience, $\mathcal{A}$ l'ensemble des groupes d'individus observables et $\mathbb{P}$ une mesure de probabilité.

On suppose l'espace fondamental Ω partitionné en C classes $\omega_1, \ldots, \omega_C$.

On suppose les probabilités de ces classes $p(\omega_1) = \mathbb{P}(\omega_1), \ldots, p(\omega_C) = \mathbb{P}(\omega_C)$, connues *a priori*.

Ces probabilités a priori correspondent à la proportion des différentes classes dans la population générale considérée.

Si on souhaite affecter un objet à une classe *sans information supplémentaire*, on assigne l'objet à la classe ω_j si

$$p(\omega_j) = \max_k p(\omega_k)$$

Cette règle affecte toujours les objets à classer à la même classe. Pour des classes de même probabilité, les objets sont affectés arbitrairement à l'une ou l'autre.

A présent, supposons que nous disposons d'un *vecteur d'observations* ou *vecteur de mesures* x (effectuées sur l'objet à classer) et que nous souhaitons l'affecter à l'une des C classes.

On modélise x comme réalisation d'une variable aléatoire vectorielle X définie sur $(\Omega, \mathcal{A}, \mathbb{P})$.

On note Z la valeur aléatoire discrète, définie sur le même espace probabilisé, qui associe à chaque individu la classe à laquelle il appartient.

Ainsi, Z est à valeur dans $\{\omega_1, \omega_2, \ldots, \omega_C\}$.

On suppose que le nombre de valeurs possible de X dans chacune des classes est suffisamment élevé pour qu'on puisse l'assimiler à une variable *continue* et on note $p(x|\omega_i)$ la *densité conditionnelle* de X sachant que $Z = \omega_i$.

Alors, la probabilité que X prenne une valeur dans le pavé infinitésimal

$$\Delta x = [x_1, x_1 + dx_1] \times \cdots \times [x_p, x_p + dx_p]$$

sachant que Z prend la valeur ω_i est donnée par :

$$\mathbb{P}(X \in \Delta x | Z = \omega_i) = p(x|\omega_i) dx_1 \ldots dx_p.$$

On a noté $x_1, \ldots, x_p$ les ***composantes*** de x.
La densité de probabilité de X sur la population générale est alors donnée par :

$$p(x) = p(\omega_1)p(x|\omega_1) + \cdots + p(\omega_C)p(x|\omega_C)$$

Une façon de le voir est de considérer la probabilité que X prenne une valeur dans le pavé infinitésimal

$$\Delta x = [x_1, x_1 + dx_1] \times \cdots \times [x_p, x_p + dx_p]$$

On notera $dx = dx_1 \ldots dx_p$.
D'après le théorème des probabilités totales, nous pouvons écrire cette probabilité :

$$\begin{aligned} p(x)dx &= \mathbb{P}(X \in \Delta x) \\ &= \sum_{i=1}^{C} \mathbb{P}(X^{-1}\Delta x \cap \omega_i) \\ &= \sum_{i=1}^{C} \mathbb{P}(X \in \Delta x|\omega_i) \cdot \mathbb{P}(\omega_i) \\ &= \sum_{i=1}^{C} p(\omega_i)p(x|w_i)dx \end{aligned}$$

d'où le résultat.

Ce résultat montre en particulier que la population générale correspondant à la distribution des valeurs de X est un **mélange** *de C sous-populations correspondant à la distribution des valeurs de X sur chacune des classes ω_i, ce mélange ayant lieu dans les proportions $p(\omega_i)$.*

Une ***règle de décision*** probabiliste consiste à affecter x à la classe ω_j si la probabilité de ω_j *étant donné* x est maximale parmi les classes $\omega_1, \ldots, \omega_C$.
C'est-à-dire qu'on affecte x à la classe ω_j si

$$p(\omega_j|x) = \max_k p(\omega_k|x)$$

Cette règle de décision partitionne l'espace des mesures $\mathcal{X}$ en C régions $D_1, \ldots, D_C$ telles que si $x \in D_j$, alors x est affecté à la classe ω_j.

Les probabilités *a posteriori* $p(\omega_j|x)$ peuvent être exprimées à partir des probabilités *a priori* et des densités conditionnelles $p(x|\omega_i)$ à l'aide du théorème de Bayes, ce qui s'écrit :

$$p(\omega_i|x) = \frac{p(x|\omega_i)p(\omega_i)}{p(x)}$$

La *règle de décision* se met alors sous la forme : affecter x à la classe ω_j si

$$p(x|\omega_j)p(\omega_j) = \max_k p(x|\omega_k)p(\omega_k)$$

Cette règle est appelée ***règle de Bayes***.

Pour *deux classes*, la règle de Bayes s'écrit : affecter x à la classe w_1 si

$$\ell_r(x) = \frac{p(x|\omega_1)}{p(x|\omega_2)} \geq \frac{p(\omega_2)}{p(\omega_1)}$$

La fonction $\ell_r(x)$ est appelée *rapport de vraisemblance*.

On peut démontrer que la règle de Bayes est ***optimale*** en ce sens qu'elle minimise la probabilité de commettre une erreur de classification, c'est-à-dire d'affecter un objet à une classe alors qu'il appartient à une autre.

3 Option de rejet

Une façon de réduire davantage l'erreur de classification est de retenir une option de *rejet*.

En effet, le plus souvent, ce sont les classifications les plus incertaines qui contribuent principalement à l'erreur de classification.

Par conséquent, *rejeter* certains objets peut conduire à réduire l'erreur de classification.

Les objets rejetés peuvent être simplement oubliés, ou mis de côté jusqu'à ce qu'un complément d'information permette de prendre une décision.

Il y a un ***arbitrage*** à effectuer entre le taux d'erreur et le taux de rejet car certaines classifications correctes peuvent être converties en rejet si on retient cette option.

On partitionne l'espace des mesures $\mathcal{X}$ en deux régions complémentaires : R la région de *rejet* et A la région d'*acceptation* ou de *classification*. Elles sont définies par

$$R = \{x : \max_i p(\omega_i|x) < 1 - t\}$$

$$A = \{x : \max_i p(\omega_i|x) \geq 1 - t\}$$

où t est un *seuil*.

Plus la valeur du seuil est petite, plus grande est la région de rejet R.

Si t est choisi tel que

$$1 - t < \frac{1}{C}$$

ou, de façon équivalente,

$$t > \frac{C-1}{C}$$

alors la région de rejet est *vide*.

En effet, la valeur $\max_i p(\omega_i|x)$ ne peut être inférieure à $1/C$ (le démontrer à titre d'exercice).

Lorsque l'option de rejet est activée, la règle de décision est donc la suivante : si $x \in A$, on affecte x à une classe selon la règle de Bayes ; si $x \in R$, on rejette x.

4 Procédure de Neyman-Pearson

Une alternative à la règle de Bayes pour le problème à *deux classes* est fournie par la procédure de *test* de Neyman-Pearson.

Dans ce contexte, on distingue deux types d'erreur : on peut affecter un objet de classe ω_1 à la classe ω_2 ou affecter un objet de classe ω_2 à la classe ω_1.

Notons

$$\begin{aligned} \epsilon_1 &= \int_{D_2} p(x|\omega_1)dx \\ &= \text{probabilité d'erreur de type I} \end{aligned}$$

$$\begin{aligned} \epsilon_2 &= \int_{D_1} p(x|\omega_2)dx \\ &= \text{probabilité d'erreur de type II} \end{aligned}$$

La règle de décision de Neyman-Pearson est obtenue en minimisant l'erreur ϵ_1 sous la contrainte que ϵ_2 soit égale à une constante. Notons-la ϵ_0.
La classe ω_1 est appelée *classe positive* et la classe ω_2 *classe négative.*
ϵ_1 est appelée *taux de faux négatifs* et ϵ_2 *taux de faux positifs.*
En formulant le problème de recherche de la règle de Neyman-Pearson comme un problème d'optimisation sous contraintes, on démontre que la règle est donnée par : affecter x à ω_1 si

$$\frac{p(x|\omega_1)}{p(x|\omega_2)} \geq \mu$$

sinon, affecter x à ω_2.
μ est un *seuil* choisi de façon à ce que

$$\int_{D_1} p(x|\omega_2)dx = \epsilon_0$$

En général, μ ne peut être déterminé analytiquement et est obtenu *numériquement.*
Souvent, la performance de la règle de Neyman-Pearson est résumée à l'aide d'une courbe ROC ; celle-ci représente le taux de vrais positifs $(1 - \epsilon_1)$ en fonction du taux de faux positifs (ϵ_2) pour les différentes valeurs possibles de μ.

5 Fonctions discriminantes

L'utilisation de la règle de Bayes requiert de connaître (ou d'estimer) les densités conditionnelles $p(x|\omega_i)$.
On peut également, au lieu de faire des hypothèses sur les $p(x|\omega_i)$, les faire directement sur les *fonctions discriminantes* à utiliser.

Une **fonction discriminante** *est une fonction du vecteur de mesures x qui définit une règle de décision.*

Par exemple, dans le cas du problème à deux classes, on appelle fonction discriminante

toute fonction $h(x)$ telle que

$$D_1 = \{x : h(x) \geq k\}$$
$$D_2 = \{x : h(x) < k\}$$

où k est une constante.

Dans le cas de deux classes, une fonction discriminante *optimale* est

$$h(x) = \frac{p(x|\omega_1)}{p(x|\omega_2)}$$

avec $k = p(\omega_2)/p(\omega_1)$.

Les fonctions discriminantes ne sont pas uniques. Soit f une fonction monotone strictement croissante et $g(x) \equiv f(h(x))$. Alors,

$$D_1 = \{x : g(x) \geq k'\}$$
$$D_2 = \{x : g(x) < k'\}$$

où $k' = f(k)$.

$g(x)$ conduit donc aux mêmes décisions que $h(x)$.

Pour le problème à C classes, on définit C fonctions discriminantes $g_i(x)$ telles que

$$D_i = \{x : g_i(x) = \max_j g_j(x)\}$$

On peut définir des fonctions de discrimination *optimales* $g_i(x)$, pour le problème à C classes, en posant

$$g_i(x) = p(x|\omega_i)p(\omega_i)$$

obtenant de ce fait la règle de Bayes.

Mais, comme dans le cas de deux classes,*il existe d'autres fonctions discriminantes* qui conduisent aux mêmes décisions.

La différence essentielle entre l'approche par densités conditionnelles et l'approche par fonctions discriminantes est que la forme de la fonction discriminante est *spécifiée* directement par l'analyste et non plus *déduite* de la distribution sous-jacente.

Le choix de la fonction discriminante peut dépendre d'informations a priori sur les objets à classifier ou correspondre à une forme paramétrique particulière possédant de bonnes propriétés d'approximation.
Diverses formes de fonction discriminantes sont utilisées en pratique ; celles-ci sont de complexité différentes : les moins complexes sont les fonctions linéaires ; parmi les plus complexes il y a celles implémentées par des Réseaux de Neurones Multi-Couches (Deep Learning).

5.1 Fonctions discriminantes linéaires

Les ***fonctions discriminantes linéaires*** sont obtenues comme combinaison linéaire des composantes de $x = (x_1, \ldots, x_d)'$,

$$g(x) = w'x + w_0 = \sum_{i=1}^{d} w_i x_i + w_0$$

Une telle fonction est entièrement spécifiée par la donnée du vecteur de *poids* w et du *biais* ou *seuil* w_0.
L'équation $g(x) = 0$ définit un *hyperplan* orthogonal au vecteur w et de distance à l'origine égale à $|w_0|/||w||$.
La valeur prise par la fonction discriminante sur un vecteur x est une mesure de la distance entre x et cet hyperplan.
Une fonction discriminante linéaire est typiquement obtenue sous l'hypothèse de densités conditionnelles *gaussiennes* et *de même matrice de variance-covariance*.
On peut également procéder de façon directe et décider d'utiliser une fonction discriminante linéaire puis déterminer ses paramètres.
Une règle de décision basée sur une fonction discriminante linéaire est appelée ***machine linéaire***.
Un cas particulier de machine linéaire est fournit par la ***règle du plus proche voisin***.

5.1.1 Règle du plus proche prototype

Supposons qu'on dispose pour chacune des classes $\omega_1, \ldots, \omega_C$ d'un ***prototype*** p_i ($i = 1, \ldots, C$).

La règle du plus proche prototype consiste à affecter x à la classe ω_i associée au prototype p_i dont il est le plus proche.

La distance entre x et le prototype p_i sécrit

$$||x - p_i||^2 = x'x - 2x'p_i + p_i'p_i$$

et la règle du plus proche prototype consiste donc à comparer les valeurs $x'p_i - \frac{1}{2}p_i'p_i$ et à sélectionner la valeur maximale.

La fonction discriminante s'écrit donc

$$g_i(x) = w_i'x + w_{i0}$$

avec $w_i = p_i$ et $w_{i0} = -\frac{1}{2}||p_i||^2$.

Ainsi, la règle du plus proche prototype est une machine linéaire.

Lorsque les prototypes p_i sont les *moyennes des classes*, on obtient la règle de la moyenne la plus proche.

Notons que les régions de décision obtenues sont *convexes*, c'est-à-dire que deux points quelconques à l'intérieur d'une région peuvent être reliés par un segment *à l'intérieur* de celle-ci.

C'est une caractéristique générale des machines linéaires.

Une conséquence est que certaines classes, bien que séparables, ne peuvent être séparées par une machine linéaire.

Deux généralisations des machines linéaires permettant de résoudre ce problème sont les fonctions discriminantes linéaires par morceaux et les fonction discriminantes généralisées.

5.2 Fonctions discriminantes linéaires par morceaux

Elles sont obtenues pour la règle du plus proche voisin au lorsqu'il y a plusieurs prototypes par classe.

Supposons qu'il y a n_i prototypes dans la classe ω_i. Notons les prototypes $p_i^1, \ldots, p_i^{n_i}$ $(i = 1, \ldots, C)$.

On définit la fonction discriminante associée à la classe ω_i par

$$g_i(x) = \max_{j=1,\ldots,n_i} g_i^j(x)$$

où g_i^j est une fonction discriminante linéaire donnée par

$$g_i^j(x) = x' p_i^j - \frac{1}{2}(p_i^j)' p_i^j$$

avec $j = 1, \ldots, n_i$, $i = 1, \ldots, C$.

Un vecteur x est affecté à la classe pour laquelle $g_i(x)$ est le plus grand, c'est-à-dire à la classe du prototype le plus proche.

Lorsque chaque élément d'une base d'apprentissage (échantillon) est pris comme prototype, on obtient la *règle du 1-plus proche voisin classique.*

5.3 Fonctions discriminantes généralisées

Une fonction discriminante généralisée, également appelée ϕ-machine, est de la forme

$$g(x) = w' \phi(x) + w_0$$

où $\phi(x) = (\phi_1(x), \ldots, \phi_r(x))'$ est une fonction vectorielle de x.

En pratique, le problème est de choisir ϕ de façon appropriée.

Lorsque $r = d$, le nombre de variables, et $\phi_i(x) = x_i$, on obtient une fonction discriminante linéaire.

6 Estimation paramétrique de densité

Dans le début du cours, on a supposé connues les fonctions de densité conditionnelles $p(x|\omega_i)$.

Dans la pratique, c'est rarement le cas. On doit donc ***estimer*** ces densités à partir des données.

Si on postule un modèle *paramétrique* pour ces densités, alors le problème se ramène à celui de l'estimation d'un nombre fini de paramètres.
Afin d'estimer $p(x|\omega_i)$, on part des données

$$\mathcal{D}_i = \{x_1^i, \ldots, x_{n_i}^i\}$$

issues de la classe ω_i.
On suppose que la densité $p(x|\omega_i)$ est de *forme* connue à un paramètre θ_i près (θ_i peut être vectoriel).
Dans le cadre de l'approche estimative ou fréquentiste, on a recours à une estimation $\widehat{\theta}_i$ du paramètre θ_i basée sur $\mathcal{D}_i$ et on pose

$$\widehat{p}(x|\omega_i) = p(x|\widehat{\theta}_i)$$

Dans le cadre de l'approche prédictive ou bayésienne (non développée dans la suite, voir le cours de Statistique Bayésienne de MASTER), θ_i est considéré comme aléatoire et on écrit

$$p(x|\omega_i) = \int p(x|\theta_i)p(\theta_i|\mathcal{D}_i)d\theta_i$$

où $p(\theta_i|\mathcal{D}_i)$ peut être vue comme une fonction de pondération basée sur les données $\mathcal{D}_i$ ou comme une densité a posteriori de θ_i sachant $\mathcal{D}_i$.

6.1 Modèles gaussiens

La règle de décision la plus utilisée en pratique est basée sur le modèle gaussien

$$\begin{aligned} p(x|\omega_i) &= \frac{1}{(2\pi)^{d/2}|\Sigma_i|^{1/2}} \cdot \\ &\quad \cdot \exp\left\{-\frac{1}{2}(x-\mu_i)'\Sigma_i^{-1}(x-\mu_i)\right\} \end{aligned}$$

Elle consiste à affecter x à la classe pour laquelle $p(\omega_i|x)$ ou, de façon équivalente, $\log p(\omega_i|x)$ est maximal.

Or, on peut écrire

$$\begin{aligned}\log p(\omega_i|x) &= \log p(x|\omega_i) + \log p(\omega_i) - \log p(x) \\ &= -\frac{1}{2}(x-\mu_i)'\Sigma_i^{-1}(x-\mu_i) \\ &\quad -\frac{1}{2}\log|\Sigma_i| - \frac{d}{2}\log(2\pi) \\ &\quad +\log p(\omega_i) - \log p(x)\end{aligned}$$

Comme $\log p(x)$ ne dépend pas de la classe, la règle de décision obtenue consiste à affecter x à ω_i si $g_i(x) > g_j(x)$ pour tout $j \neq i$, où

$$\begin{aligned}g_i(x) &= \log p(\omega_i) - \frac{1}{2}\log|\Sigma_i| \\ &\quad -\frac{1}{2}(x-\mu_i)'\Sigma_i^{-1}(x-\mu_i)\end{aligned}$$

On obtient ainsi la *fonction discriminante gaussienne quadratique*.

Dans le cadre de l'approche *estimative*, les quantités μ_i et Σ_i apparaîssant dans la formule ci-dessus sont remplacées par leurs estimations à partir des données.

A cet effet, on utilise le plus souvent les estimations par maximum de vraisemblance

$$\widehat{\mu}_i = \frac{1}{n_i}\sum_{k=1}^{n_i} x_k^i$$

et

$$\widehat{\Sigma}_i = \frac{1}{n_i}\sum_{k=1}^{n_i}(x_k^i - \widehat{\mu}_i)(x_k^i - \widehat{\mu}_i)'$$

En les injectant dans la fonction discriminante, on obtient

$$\begin{aligned}\widehat{g}_i(x) &= \log p(\omega_i) - \frac{1}{2}\log|\widehat{\Sigma}_i| \\ &\quad -\frac{1}{2}(x-\widehat{\mu}_i)'\widehat{\Sigma}_i^{-1}(x-\widehat{\mu}_i)\end{aligned}$$

Lorsqu'elle est inconnue, $p(\omega_i)$ est en général estimée par

$$\widehat{p}(\omega_i) = \frac{n_i}{\sum_j n_j}$$

Le principal problème rencontré en utilisant des fonctions discriminantes gaussiennes est qu'une ou plusieurs des matrices $\widehat{\Sigma}_i$ peuvent être singulières (non inversibles).

On considère en général trois solutions à ce problème :

1 – Utiliser des matrices de variances-covariances diagonales

2 – Projeter les données dans un espace où $\widehat{\Sigma}_i$ est inversible, par exemple à l'aide d'une ACP

3 – Supposer que les matrices de covariances $\Sigma_1, \ldots, \Sigma_C$ sont identiques

Dans ce dernier cas, la fonction discriminante se simplifie et la règle de décision devient : affecter x à ω_i si $g_i(x) > g_j(x)$ pour tout $j \neq i$, où

$$g_i(x) = \log p(\omega_i) - \frac{1}{2}\widehat{\mu}_i' S_W^{-1} \widehat{\mu}_i + x' S_W^{-1} \widehat{\mu}_i$$

où S_W est la matrice de covariance commune des différentes classes.

La fonction discriminante ainsi définie est la *fonction discriminante gaussienne linéaire.*

L'estimation par maximum de vraisemblance de S_W est

$$\widehat{S}_W = \sum_{i=1}^{C} \frac{n_i}{n} \widehat{\Sigma}_i$$

où $n = \sum_j n_j$.

Une estimation non biaisée de S_W est fournie par

$$\widetilde{S}_W = \frac{n}{n - C} \widehat{S}_W$$

Dans les problèmes où les données sont issues de classes gaussiennes de matrices de covariances différentes, on peut ne pas avoir suffisamment de données pour estimer correctement ces différentes matrices de covariance.

Là également, on a recours de préférence à la fonction discriminante gaussienne linéaire, qui reste souvent intéressante en pratique.

Toutefois, les bonnes performance de cette dernière reposent essentiellement sur l'hypothèse de normalité. Il peut donc être nécessaire de transformer les données avant de l'utiliser.

7 Estimation non paramétrique de densité

Dans de nombreux cas, on ne peut supposer les densités connues à des paramètres près et on doit avoir recours à des méthodes *non paramétriques* d'estimation de densité.
Nous en considérerons quatre dans la suite : l'approche par histogramme, l'approche par k plus proches voisins, l'approche par développement en séries de fonctions et l'approche par noyaux.
C'est cette dernière qui est implémentée par la fonction R **density** par exemple.
Commençons par considérer quelques propriétés fondamentales des estimateurs de la densité.
Soit $X_1, \ldots, X_n$ des variables i.i.d. de dimension d, de densité $p(x)$.
Cette dernière vérifie nécessairement

$$p(x) \geq 0, \quad \int p(x)dx = 1$$

On souhaite estimer $p(x)$ à partir de réalisation des variables précédentes.
Un premier résultat est que si l'estimateur utilisé $\widehat{p}(x)$ est également une densité, alors *il est nécessairement biaisé*, c'est-à-dire qu'il existe une densité p continue et x tels que

$$\mathbb{E}[\widehat{p}(x)] \neq p(x)$$

On peut toutefois construire des estimateurs *asymptotiquement* sans biais, c'est-à-dire tels que pour toute densité p suffisamment régulière et en tout x,

$$\mathbb{E}[\widehat{p}(x)] \longrightarrow p(x), \quad n \to \infty$$

mais en pratique, on est limité par la taille de l'échantillon.
On peut également s'intéresser à l'*erreur quadratique moyenne* (EQM/MSE) entre $\widehat{p}(x)$ et $p(x)$.
Celle-ci est définie par

$$\mathrm{EQM}_x(\widehat{p}) = \mathbb{E}[(\widehat{p}(x) - p(x))^2]$$

ce qu'on peut également écrire

$$\mathrm{EQM}_x(\widehat{p}) = \mathrm{var}(\widehat{p}(x)) + \{\mathrm{biais}(\widehat{p}(x))\}^2$$

Lorsque $\mathrm{EQM}_x(\widehat{p}) \to 0$ pour tout $x \in \mathbb{R}^d$, on dit que $\widehat{p}$ est un estimateur *ponctuellement convergent de p en moyenne quadratique.*

On définit également l'erreur quadratique intégrée (EQI/ISE)

$$\mathrm{EQI}(\widehat{p}) \equiv \int [\widehat{p}(x) - p(x)]^2 dx$$

et l'erreur quadratique intégrée moyenne (EQIM/MISE)

$$\mathrm{EQIM}(\widehat{p}) \equiv \mathbb{E}\left[\int [\widehat{p}(x) - p(x)]^2 dx\right]$$

Comme l'espérance et l'intégrale peuvent être permutées, l'EQIM n'est autre que l'intégrale de l'EQM, c'est-à-dire la somme de l'intégrale de la variance et de l'intégrale du carré du biais.

En pratique, on n'utilise pas nécessairement des estimateurs de la densité qui sont des densités. On requiert surtout qu'ils soient ponctuellement convergent. On peut donc parfois envisager des estimateurs qui peuvent prendre des valeurs négatives. De même, l'estimateur par k plus proches voisins, très utilisé en pratique, a une intégrale infinie.

7.1 Histogrammes

L'histogramme est la méthode la plus ancienne d'estimation de la densité. C'est la méthode *classique* de construction d'une densité de probabilité à partir d'un échantillon.

En une dimension, sous sa forme usuelle, l'axe réel est partitionné en un nombre de cellules d'amplitudes égales et l'estimation de la densité au point x est donnée par

$$\widehat{p}(x) = \frac{n_j}{\sum_j n_j \cdot \Delta x}$$

où n_j est le nombre d'observations dans la cellule d'amplitude Δx qui contient x.

Cette formule se généralise en

$$\widehat{p}(x) = \frac{n_j}{\sum_j n_j \cdot V}$$

pour des observations multidimensionnelles, où V est le volume d'une cellule.

Bien que simple dans son principe et facile à implémenter, l'histogramme pose de nombreux problèmes.

Tout d'abord, l'approche est de moins en moins pratique quand la dimension de l'espace des observations devient plus grande.

Supposons que le domaine de variation de chaque variable est divisé en N cellules ; alors, en une dimension, l'histogramme est construit sur N cellules, en deux dimensions sur N^2, et en d dimensions sur N^d cellules.

Ainsi, en grande dimension, une quantité gigantesque de données est requise pour estimer la densité par histogramme. En six dimensions, si on suppose que l'intervalle de variation de chaque variable est divisé en dix cellules, l'histogramme sera construit sur un million de cellules et, à moins d'avoir suffisamment de données, l'estimation de la densité sera nulle en la plupart des points.

Un deuxième inconvénient est que l'estimation par histogramme est discontinue et passe brutalement à zéro aux bornes de son intervalle de définition.

On va considérer diverses approches permettant de résoudre ces problèmes.

7.2 Méthode naïve

Une première approche permettant de réduire le nombre de cellules requises en dimension élevée est de faire des hypothèses simplificatrices concernant la forme de la densité à estimer.

Si on suppose que les variables sont indépendantes, $p(x)$ se met sous la forme

$$p(x) = \prod_{i=1}^{d} p(x_i)$$

où les $p(x_i)$ sont les densités marginales des composantes de x.

Ce modèle est souvent appelé *modèle bayésien naïf*.

On utilise alors un histogramme pour chacune des densités marginales. En supposant que chacun de ces histogrammes est construit sur N cellules, l'estimation est basée sur dN cellules au lieu de N^d.

Une implémentation particulière de l'approche naïve donne

$$\widehat{p}(x) = \left\{ \prod_{i=1}^{d} \frac{n(x_r) + 1/C_r}{N(r) + 1} \right\}^B$$

où x_r désigne la r-ième composante de x, $n(x_r)$ est le nombre d'observations dans la même cellule que x_r sur la variable r, $N(r)$ est le nombre d'observations pour la variable r, C_r est le nombre de cellules pour la variable r et B est un *facteur d'association*. En pratique, on essaye $B = 1, 0.8, 0.5$.

Il est généralement admis que la méthode naïve peut donner de bons résultats en pratique.

7.3 Méthode de Lancaster

Les modèles de Lancaster fournissent une représentation de la densité conjointe à partir des densités marginales *sous l'hypothèse que les interactions (en un certain sens) au delà d'un certain ordre sont nulles.*

Si on suppose que les interactions d'ordre supérieur à $s = 1$ sont nulle, on obtient la méthode naïve.

Si on pose $s = 2$, alors on peut montrer que la densité conjointe s'exprime en fonction des densités marginales et des densités marginales d'ordre deux sous la forme

$$\begin{aligned} p(x) &= \left\{ \sum_{i<j} \frac{p(x_i, x_j)}{p(x_i)p(x_j)} - \left[\frac{p(p-1)}{2} - 1 \right] \right\} \cdot \\ &\quad \cdot p_{\text{ indep}}(x) \end{aligned}$$

où $p_{\text{ indep}}(x)$ désigne la fonction de densité sous l'hypothèse d'indépendance

$$p_{\text{ indep}}(x) = \prod_{i=1}^{d} p(x_i)$$

Une estimation de la densité conjointe est obtenue en injectant dans la formule de Lancaster des estimations par histogramme des $p(x_i, x_j)$ et des $p(x_i)$.

On peut par exemple retenir

$$\widehat{p}(x_i, x_j) = \frac{n(x_i, x_j) + 1/(C_i C_j)}{N(i, j) + 1}$$

où les $n(x_i, x_j)$ et $N(i, j)$ sont définis de la même façon que pour l'estimation naïve, et où

$$\widehat{p}(x_i) = \left\{ \frac{n(x_i) + 1/C_i}{N(i) + 1} \right\}^B$$

Un inconvénient de cette procédure d'estimation est qu'elle peut fournir des valeurs négatives.

La solution généralement retenue dans ce cas consiste à utiliser l'estimation naïve.

7.4 Méthode des k-plus proches voisins

La méthode des k plus proches voisins est basée sur les considérations suivantes.

La probabilité qu'une variable aléatoire X de densité $p(x)$ prenne sa valeur dans une région $V(x)$ de volume V centrée sur un point x est donnée par

$$\theta = \int_{V(x)} p(u)du$$

Lorsque le volume est petit, on peut écrire

$$\theta \approx p(x) \cdot V$$

Lorsqu'on dispose d'un échantillon i.i.d. de densité $p(x)$, on peut approximer cette probabilité par la proportion d'observations appartenant à $V(x)$.

Si on note k le nombre d'observations (parmi n) appartenant à $V(x)$, alors on a

$$\theta \approx \frac{k}{n} \approx p(x) \cdot V$$

d'où on déduit une estimation de la densité

$$\widehat{p}(x) = \frac{k}{nV}$$

La méthode des k plus proches voisins consiste à fixer k/n (donc k, à n donné) et à déterminer le volume V de la plus petite hypersphère $V(x)$ centrée sur x contenant k observations.

Par exemple, si on note $x_{(k)}$ le k-ième plus proche voisin de x dans $\mathbb{R}^d$ pour la norme euclidienne $||\cdot||$, alors on prend pour $V(x)$ l'hypersphère centrée sur x, de rayon $||x - x_{(k)}||$.

Le volume de cette hypersphère est donné par

$$V = 2||x - x_{(k)}||^d \pi^{d/2} \Gamma(d/2)$$

Le ratio k/nV fournit l'estimation voulue.

Cette approche est différente de celle de l'histogramme où on fixe *le volume* puis on détermine le nombre d'observations qu'il contient.

Le paramètre crucial est k. S'il est trop grand, l'estimation obtenue sera trop lisse et les détails seront perdus. S'il est trop faible, on obtient une estimation avec de nombreux pics.

Une caractéristique de l'estimation obtenue est qu'elle *n'est pas une densité*, étant d'intégrale infinie.

Toutefois, on peut démontrer que l'estimateur des k plus proches voisins est asymptotiquement sans biais et ponctuellement convergent en moyenne quadratique dès lors que

$$k(n) \to \infty, \quad n \to \infty$$

et

$$\frac{k(n)}{n} \to 0, \quad n \to \infty$$

7.4.1 Application à l'analyse discriminante

Disposant d'une estimation de la densité, on peut l'utiliser pour construire une règle de décision.

Supposons que parmi les k plus proches voisins de x, k_i appartiennent à la classe ω_i. On a donc

$$\sum_{i=1}^{C} k_i = k$$

Notons n_i le nombre d'observations issues de la classe ω_i. On a donc

$$\sum_{i=1}^{C} n_i = n$$

Notons V le volume de la plus petite hypersphère centrée sur x contenant k observations.

Alors, en combinant l'approche par ppv et par histogramme, on peut estimer la densité conditionnelle $p(x|\omega_i)$ par

$$\widehat{p}(x|\omega_i) = \frac{k_i}{n_i V}$$

et la probabilité a priori $p(\omega_i)$ par

$$\widehat{p}(\omega_i) = \frac{n_i}{n}$$

La règle de décision consiste à affecter x à ω_i si

$$\widehat{p}(x|\omega_i)\widehat{p}(\omega_i) \geq \widehat{p}(x|\omega_j)\widehat{p}(\omega_j) \text{ pour tout } j$$

ce qui s'écrit, en appliquant le théorème de Bayes,

$$\frac{k_i}{n_i V} \cdot \frac{n_i}{n} \geq \frac{k_j}{n_j V} \cdot \frac{n_j}{n}$$

pour tout j.

Ainsi, on affecte x à la classe ω_i si

$$k_i \geq k_j \text{ pour tout } j$$

Autrement dit, *on affecte x à la classe la plus représentée parmi ses k plus proches voisins.*

En cas de présence d'ex-aequo, on procède en général de l'une des façons suivantes :

1 – on choisit une classe aribtrairement parmi les ex-aequo

2 – on affecte x à la classe ω_i, parmi les ex-aequo, dont la moyenne (calculée sur k_i observations) est la plus proche de x

3 – on affecte x à la classe la plus proche parmi les ex-aequo

On peut démontrer que l'erreur de classification asymptotique e de la règle du plus proche voisin ($k = 1$) vérifie

$$e_B \leq e \leq e_B(2 - \frac{Ce_B}{C-1})$$

où e_B désigne l'erreur de la règle de Bayes, donc de la règle optimale.

Ainsi, pour de grands échantillons, l'erreur de classification par plus proche voisin est bornée supérieurement par deux fois l'erreur de la règle de Bayes.
Elle fournit également une borne inférieure pour celle-ci puisqu'on déduit de l'inégalité précédente que

$$\frac{C-1}{C} - \sqrt{\frac{C-1}{C}}\sqrt{\frac{C-1}{C} - e} \leq e_B \leq e$$

7.5 Développement en séries de fonctions

On souhaite approximer une densité $p(x)$ par une somme pondérée de fonctions de base orthogonales.
Plus précisément, on suppose que la densité admet le développement

$$p(x) = \sum_{i=1}^{\infty} a_i \phi_i(x)$$

où les $\{\phi_i\}$ vérifient

$$\int k(x)\phi_i(x)\phi_j(x)dx = \lambda_i \delta_{ij}$$

pour le *noyau* ou *fonction de pondération* $k(x)$, où $\delta_{ij} = 1$ si $i = j$, 0 sinon.
Il découle des relations précédentes que

$$\lambda_i a_i = \int k(x)\phi_i(x)p(x)dx$$

Etant donné $X_1, \ldots, X_n$ échantillon i.i.d. de densité $p(x)$, on peut estimer les a_i sans biais à l'aide de la formule

$$\lambda_i \widehat{a}_i = \frac{1}{n}\sum_{j=1}^{n} k(X_j)\phi_i(X_j)$$

On en déduit l'estimateur par développement en série orthogonale de $p(x)$:

$$\widehat{p}_n(x) = \sum_{i=1}^{s} \frac{1}{n\lambda_i} \sum_{j=1}^{n} k(X_j)\phi_i(X_j)\phi_i(x)$$

où s est le nombre de termes retenus dans le développement.
Les coefficients $\widehat{a}_i$ peuvent être calculés récursivement à l'aide de la formule

$$\begin{aligned} \lambda_i \widehat{a}_i(r+1) &= \frac{r}{r+1}\lambda_i \widehat{a}_i(r) \\ &\quad + \frac{1}{r+1} k(X_{r+1})\phi_i(X_{r+1}) \end{aligned}$$

où $\widehat{a}_i(r+1)$ est l'estimateur de a_i fonction de $X_1, \ldots, X_{r+1}$.

La formule précédente est un avantage de la méthode : on peut mettre les estimations à jour au fur et à mesure qu'on reçoit de nouvelles observations, sans avoir besoin de conserver l'ensemble des données en mémoire.

Un autre avantage de la méthode est que l'estimation finale obtenue consiste en un vecteur de coefficients ; elle est donc facile à sauvegarder et à récupérer.

La méthode présente par contre l'inconvénient d'être limitée, en pratique, à des données de dimension faible. Le nombre de coefficients croît en effet exponentiellement avec la dimension.

De plus, l'estimation obtenue n'est pas nécessairement une densité et elle n'est pas nécessairement positive, ce qui peut poser problème dans certaines applications.

Diverses fonctions sont utilisées comme fonctions de base : celles-ci incluent les fonctions trigonométriques sur $[0, 1]$, les polynômes de Legendre sur $[-1, 1]$, ainsi que les polynôme de Laguerre sur $[0, \infty[$.

Les fonctions de bases les plus utilisées pour des densités à support non borné sont les fonctions d'Hermite

$$\phi_k(x) = \frac{H_k(x)}{(2^k k! \sqrt{\pi})^{1/2}} \, e^{-x^2/2}$$

où $H_k(x)$ est le polynôme d'Hermite d'ordre k

$$H_k(x) = (-1)^k \exp(x^2) \frac{d^k}{dx^k} \exp(-x^2)$$

La régularité et les performances de l'estimateur obtenu dépendent du nombre de termes s retenus dans le développement.

Retenir trop peu de termes conduit à une approximation trop grossière (lissage excessif).

Diverses procédures de choix de s ont donc été proposées voir l'article d'Izenmann, Journal of the American Statistical Association 86 (1991), pages 205-223.

7.6 Méthodes à noyau

Afin d'estimer la densité p en un point x, la méthode des k plus proches voisins part du plus petit voisinage sphérique de x contenant k voisins puis détermine son volume. Sous sa forme la plus simple, la méthode du noyau (également appelée estimateur de Parzen-Rosenblatt) procède en fixant le volume d'un voisinage, en général cubique, centré sur x, puis détermine le nombre d'observations contenues dans ce voisinage. De ce point de vue, elle est analogue à l'histogramme, si ce n'est qu'à chaque point x correspond un voisinage différent.

En dimension 1, on peut dire que l'histogramme est construit sur des "fenêtres" fixes, alors que l'estimateur à noyau est construit sur des "fenêtres" *mobiles*.

Plus précisément, soit $X_1, \ldots, X_n$ un échantillon i.i.d. de densité p. Notons P la fonction de répartition associée à p :

$$P(x) = \int_{-\infty}^{x} p(u)du$$

Un estimateur naturel de la fonction de répartition est donné par

$$\widehat{P}(x) = \frac{\text{nombre d'observations } \leq x}{n}$$

p étant la dérivée de P, une idée, pour obtenir un estimateur de p, serait de dériver $\widehat{P}$.

Malheureusement, $\widehat{P}$ étant une fonction en escaliers, sa dérivée, lorsqu'elle existe est toujours égale à zéro.

On peut toutefois utiliser comme estimateur de la densité

$$\widehat{p}(x) = \frac{\widehat{P}(x+h) - \widehat{P}(x-h)}{2h}$$

où h est un (petit) nombre positif.

$\widehat{p}(x)$ correspond à la proportion d'observations appartenant à l'intervalle $]x-h, x+h]$, divisé par $2h$.

On peut mettre cet estimateur sous la forme

$$\widehat{p}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right)$$

avec $K(z) = 0$ si $|z| > 1$ et $K(z) = 1/2$ sinon.

L'estimateur de la densité ainsi défini est *discontinu*. Pour corriger cet inconvénient, on généralise l'estimateur en prenant une fonction de pondération K (noyau) plus "lisse".

De façon analogue, en dimension d, étant donné un ensemble d'observations $x_1, \ldots, x_n$ et un noyau K, l'estimation de la fonction de densité au point x par la méthode à noyau est donnée par

$$\widehat{p}(x) = \frac{1}{nh^d} \sum_{i=1}^{n} K(\frac{x - x_i}{h})$$

h est appelé *fenêtre*, *largeur de fenêtre* ou *paramètre de lissage*.

Des exemples usuels de noyaux en dimension 1 sont les suivants.

Rectangulaire : $K(z) = 1/2$ pour $|z| < 1$, 0 sinon

Triangulaire : $K(z) = 1 - |z|$ pour $|z| < 1$, 0 sinon

Gaussien : $K(z) = e^{-z^2/2}/\sqrt{2\pi}$

Epanechnikov : $K(z) = \frac{3}{4}(1 - z^2/5)/\sqrt{5}$ pour $|z| < \sqrt{5}$, 0 sinon

Biweight : $\frac{15}{16}(1 - z^2)^2$ pour $|z| < 1$, 0 sinon

Les noyaux *multivariés* sont généralement à symétrie radiale (sphérique).

Les exemples classiques sont le noyau gaussien

$$K(z) = (2\pi)^{-d/2} \exp(-z'z/2)$$

et le noyau d'Epanechnikov multivarié

$$K(z) = (1 - z'z)(d + 2)/(2c_d), \quad |z| < 1$$

0 sinon, où

$$c_d = \pi^{d/2}/\Gamma((d/2) + 1)$$

est le volume de la sphère unité en dimension d.

Lorsqu'on souhaite avoir des paramètres de lissage différents pour les différentes variables, on utilise des *noyaux produits*, ce qui nous donne des estimations de la forme

$$\widehat{p}(x) = \frac{1}{n} \frac{1}{h_1 \cdots h_d} \sum_{i=1}^{n} \prod_{j=1}^{d} K_j \left(\frac{[x - x_i]_j}{h_j} \right)$$

Les K_j sont des noyaux univariés quelconques, typiquement choisis parmi ceux mentionnés précédemment.

On utilise parfois également des estimations définies par

$$\widehat{p}(x) = \frac{1}{n}\sum_{i=1}^{n} |H|^{-1/2} K(H^{-1/2}(x - x_i))$$

où K est un noyau multivarié à symétrie sphérique et H est une matrice symétrique définie positive.

Dans le cadre de l'analyse discriminante, on prend généralement

$$H = h_i^2 \widehat{\Sigma}_i$$

pour la classe ω_i, où h_i est un facteur d'échelle et $\widehat{\Sigma}_i$ est la matrice de variance-covariance empirique.

7.6.1 Méthodes à noyau : propriétés

Si $K(z) \geq 0$ pour tout z et $\int K(z)dz = 1$, alors l'estimation de la densité $\widehat{p}(x)$ vérifie également $\widehat{p}(x) \geq 0$ pour tout x et $\int \widehat{p}(x)dx = 1$; c'est donc une densité.

Toutefois, dans ce cas, l'estimateur à noyau est *biaisé* pour tout échantillon fini.

Il est toutefois *asymptotiquement* sans biais et convergent, moyennant des conditions supplémentaires sur le noyau et les conditions suivantes sur la fenêtre :

(*i*) $\lim h_n = 0$

(*ii*) $\lim n h_n^d = \infty$

7.6.2 Choix de la fenêtre

Une des facteurs critiques de performance de la méthode à noyau est le bon choix de la largeur de fenêtre.

Si h est trop petit, l'estimation obtenue est une suite de n pics centrés sur les observations. Si h est trop grand, le lissage est excessif et l'estimation obtenue ne rend pas suffisamment compte de la structure de la densité.

Le choix optimal de h dépend de plusieurs facteurs :

— du nombre d'observations

— de la densité inconnue

— du noyau retenu

— du critère d'optimalité utilisé

Les techniques les plus utilisées en pratique pour déterminer h sont les suivantes :

1. Prendre pour h la distance moyenne entre les observations et leur k-ième plus proche voisin, avec $k = 10$ par exemple
2. Déterminer la valeur de h qui minimise l'Erreur Quadratique Moyenne Intégrée (MISE) entre la densité et la densité estimée. Pour un noyau gaussien à symétrie sphérique, il est suggéré de prendre

$$h = \sigma \left(\frac{4}{d+2} \right)^{1/d+4} n^{-1/d+4}$$

avec

$$\sigma = \frac{1}{d} \sum_{i=1}^{d} s_{ii}$$

où les s_{ii} sont les termes diagonaux d'une estimation de la matrice de variance-covariance.

La spécification précédente est adaptée à des observations approximativement gaussiennes, mais elle peut conduire à une estimation trop lisse dans le cas d'une densité multimodale.

Une valeur un peu plus faible de h peut alors être adaptée.

En analyse discriminante, on essaye plusieurs valeurs et on compare les taux d'erreur de classification obtenus pour chacune.

3. Il existe des façons plus sophistiquées de choisir h, basées sur la validation croisée.

   Dans l'approche par vraisemblance, par exemple, h est choisi de façon à maximiser

$$\prod_{i=1}^{n} \widehat{p}_i(x_i)$$

où $\widehat{p}_i(x_i)$ est l'estimation de la densité basée sur toutes les observations moins la i-ème.

Un inconvénient de cette approche est qu'elle s'avère plutôt mauvaise pour des distributions de type log-normal ou Student, par exemple.

4. Dans le cas univarié, diverses autres méthodes ont été proposées.

   Parmi celle-ci la méthode "plug-in" consiste à injecter une estimation de $S \equiv \int (p''(x))^2 dx$ dans l'expression de la fenêtre optimale pour le critère EQMI (MISE) asymptotique

   $$h = \left(\frac{c_0}{c_1^2 S n} \right)^{1/5}$$

   où $c_0 = \int K^2(z)dz$ et $c_1 = \int z^2 K(z)dz$.

   Cette estimation peut être obtenue à l'aide d'un estimateur à noyau convenablement défini : voir l'article de Sheather & Jones dans Journal of The Royal Statistical Society série B 53 (1991), pages 683-690, pour une implémentation satisfaisante en pratique.

7.6.3 Choix du noyau

Un autre choix à effectuer lors de l'utilisation de la méthode à noyau concerne le noyau lui-même.

En pratique, les noyaux les plus utilisés sont le noyau gaussien en dimension 1 et, en dimensions supérieures, des noyaux produits de noyaux gaussiens unidimensionnels.

Diverses études ont montré que le choix du noyau n'était pas un facteur critique de performance de la méthode à noyau.

7.6.4 Implémentations

Sous R, la méthode à noyau est implémentée, entre autres, par la fonction **density** pour des données en dimension 1 et par **bkde2D** (package KernSmooth) pour des données en dimension 2.

Sous SAS, l'option **kernel** de la **proc univariate** permet d'obtenir des estimations à noyau de la densité pour des données en dimension 1. La **proc kde** permet d'obtenir des estimations à noyau de la densité pour des données en dimension 1 ou 2.

8 Pouvoir discriminant et matrice de confusion

On étudiera dans la suite trois aspects de la performance des techniques de discrimination :

1. Pouvoir discriminant : qualité attendue de la classification pour des données *non observées*

2. Fiabilité : qualité de l'estimation des probabilités a posteriori

3. Courbe ROC : indicateur de performance pour le problème à deux classes

8.1 Pouvoir discriminant

Il existe plusieurs mesure du pouvoir discriminant, la plus utilisée étant le *taux d'erreur de classification*, parfois plus simplement appeleée *taux d'erreur*.
Il est en général très compliqué d'obtenir une expression *analytique* du taux d'erreur ; c'est pourquoi celui-ci est en général *estimé* à partir des données disponibles.
Il existe de nombreux travaux sur l'estimation du taux d'erreur ; toutefois, celui-ci a l'inconvénient d'être une mesure de performance *unidimensionnelle*.
De plus, il traite de la même façon toutes les classifications correctes et toutes les classifications incorrectes.

8.2 Matrice de confusion

En plus du calcul du taux d'erreur, il est recommandé de calculer une *matrice de confusion.*

L'élement (i, j) de cette matrice est le nombre d'observations dans la classe ω_j classées à tort dans la classe ω_i par la règle d'affectation.

La matrice de confusion permet d'obtenir une *décomposition* de l'erreur de classification.

Dans le cas de l'analyse discriminante à deux classes, on note souvent les deux classes 0 et 1 respectivement.

La classe 1 est appelée *classe positive.*

La classe 0 est appelé *classe négative.*

La matrice de confusion prend alors la forme suivante :

		Classe réelle	
		0	1
Classe	0	vrais négatifs (TN)	faux négatifs (FN)
prédite	1	faux positifs (FP)	vrais positifs (TP)

On appelle :

vrais positifs (en anglais *true positives*) les exemples positifs correctement classifiés ;

faux positifs (en anglais *false positives*) les exemples négatifs étiquetés positifs par le modèle ;

vrais négatifs (en anglais *true negatives*) les exemples négatifs correctement classifiés ;

faux négatifs (en anglais *false negatives*) les exemples positifs étiquetés négatifs par le modèle.

Les faux positifs sont également appelés *fausses alarmes* ou *erreurs de type I*, par opposition aux *erreurs de type II* qui sont les faux négatifs.

On peut obtenir de nombreux *critères d'évaluation* d'une analyse discriminante à partir de la matrice de confusion.

En voici quelques uns.

On appelle *rappel* (en anglais *recall*) ou *sensibilité* (en anglais *sensitivity*) le *taux* de vrias positifs, c'est-à-dire la proportion d'exemples positifs correctement identifiés comme tels :

$$\text{Rappel} = \frac{TP}{TP + FN}$$

Il est très facile d'avoir un bon rappel en prédisant que *tous* les exemples sont positifs. Ainsi, *le critère du rappel ne peut être utilisé seul.*

On lui adjoint souvent *la précision.*

On appelle *précision*, ou *valeur positive prédictive* (en anglais *positive predictive value,PPV*), la proportion de prédictions correctes parmi les prédictions positives :

$$\text{Précision} = \frac{TP}{TP + FP}$$

De même que l'on peut facilement avoir un très bon rappel au détriment de la précision, il est facile d'obtenir une bonne précision au détriment du rappel : il suffit de faire très peu de prédictions positives, ce qui réduit le risque qu'elles soient erronées.

Remarque : l'anglais distingue *precision* (la précision ci-dessus) et *accuracy*, qui est la proportion d'observations correctement classées, soit le complémentaire à 1 du taux d'erreur, aussi traduit par *précision* en français. On devra donc faire attention en utilisant ce terme.

Pour résumer rappel et précision en un seul nombre, on calcule la *F-mesure.*

On appelle *F-mesure* (en anglais *F-score* ou *F1-score*) la moyenne harmonique de la précision et du rappel :

$$\begin{aligned} F &= 2\,\frac{\text{Précision} \cdot \text{Rappel}}{\text{Précision} + \text{Rappel}} \\ &= \frac{2TP}{2TP + FP + FN} \end{aligned}$$

On appelle *spécificité* le taux de vrais négatifs, autrement dit la proportion d'exemples négatifs correctement identifiés comme tels :

$$\text{Spécificité} = \frac{TN}{FP + TN}$$

Exemple : On considère un test clinique pour illustrer les critères précédents. Il ne s'agit pas d'un modèle statistique d'analyse discriminante mais d'un frottis de dépistage du col de l'utérus : il s'agit d'un examen simple et moins invasif qu'un examen histologique, qui doit être interprété par un expert et qui sert de vérité terrain.

Les résultats d'une expérience menée sur 4000 femmes âgées de 40 ans et plus sont présentés dans le tableau suivant :

	Vérité terrain		
	Cancer	Pas de cancer	Total
Frottis +	190	210	400
Frottis -	10	3590	3600
Total	200	3800	4000

Le rappel est de 95%, la spécificité de 94.5% mais la précision ne vaut que 47.5%.

Ainsi, ce test est un bon outil de dépistage : la probabilité de n'avoir pas de cancer quand le frottis est négatif est élevée ($3590/3600 \approx 99.7\%$).

Cependant, c'est un mauvais outil diagnostique, au sens où la probabilité de fausse alarme est élevée.

9 Les taux d'erreur des méthodes de discrimination

Notons les données d'apprentissage

$$Y = \{y_i : i = 1, \ldots, n\}$$

chaque donnée y_i étant composée de deux parties

$$y_i' = (x_i', z_i')$$

où $\{x_i : i = 1, \ldots, n\}$ sont les valeurs prises par le vecteur de variables explicatives (observations, mesures) et où $\{z_i : i = 1, \ldots, n\}$ permettent d'identifier les classes associées.

Plus précisément, on supposera que z_i est un vecteur de dimension C, tel que

$$(z_i)_j = 1 \quad \text{si} \quad x_i \in \omega_j$$

et 0 sinon.

On notera $\omega(z_i)$ la classe correspondant à z_i.

Notons la règle d'apprentissage obtenue à l'aide de l'échantillon d'apprentissage $\eta(x; Y)$.

$\eta(x; Y)$ correspond à la classe à laquelle est affecté x au vu des données Y, par la technique de discrimination utilisée.

Notons $Q(\omega(z), \eta(x; Y))$ la fonction de perte

$$Q(\omega(z), \eta(x; Y)) = 0 \quad \text{si} \quad \omega(z) = \eta(x; Y)$$

(classification correcte) et 1 sinon (classification incorrecte).

9.1 Taux d'erreur réel

Le *taux d'erreur réel*, également appelé *taux d'erreur conditionnel*, noté e_T ou e_r, est la probabilité de mal classer une observation $y' = (x', z')$ tirée au hasard dans la population mère en utilisant la règle d'affectation obtenue sur l'échantillon d'apprentissage.

$$\begin{aligned} e_T &= \mathbb{E}[\mathbb{I}(\omega(z) \neq \eta(x; Y))|Y] \\ &= \mathbb{P}(\omega(z) \neq \eta(x; Y)|Y) \\ &= \mathbb{E}[Q(\omega(z), \eta(x; Y))|Y] \end{aligned}$$

e_T est aléatoire car *fonction* de Y. A Y fixé, il est constant.

9.2 Taux d'erreur apparent

Le taux d'erreur apparent e_A, ou *taux d'erreur par resubstitution* est obtenu en utilisant l'échantillon d'apprentissage pour estimer le taux d'erreur

$$e_A = \frac{1}{n} \sum_{i=1}^{n} Q(\omega(z_i), \eta(x_i; Y))$$

Celui-ci fournit une approximation du taux d'erreur réel. Il peut être *biaisé* de façon très optimiste, particulièrement pour des règles d'affectation complexes et des ensembles d'apprentissage de taille très limitée, *i.e.* lorsqu'il y a risque de *surajuster* les données.

Dans ce contexte, on appelle *biais* l'écart $e_T - e_A$. Il s'agit d'une *variable aléatoire fonction de* Y, tout comme e_T et e_A. A Y fixé, le biais est constant.

9.3 Taux d'erreur espéré

Le *taux d'erreur espéré* ou *taux d'erreur moyen* e_E est l'espérance du taux d'erreur réel relativement à des échantillons d'apprentissage de taille donnée

$$e_E = \mathbb{E}(e_T) = \mathbb{E}[\mathbb{E}(Q(\omega(z), \eta(x; Y))|Y)]$$

9.4 Taux d'erreur bayésien

Le *taux d'erreur bayésien*, également appelé *taux d'erreur de Bayes* ou *taux d'erreur optimal*, e_B, est le minimum théorique du taux d'erreur réel.

Il correspond au taux d'erreur réel lorsque la règle d'affectation obtenue sur l'échantillon d'apprentissage est identique à la règle de Bayes basée sur les probabilités conditionnelles $p(\omega_i|x)$, $i = 1, \ldots, C$.

10 La validation croisée

10.1 Validation croisée (hold-out)

La méthode de validation croisée hold-out consiste à scinder les données en deux ensembles mutuellement exclusifs : l'ensemble d'apprentissage et l'ensemble de test ou de validation.

La règle d'affectation est obtenue à partir de l'échantillon d'apprentissage et sa performance est évaluée sur l'échantillon test.

Cette méthode n'utilise pas les données de façon optimale puisqu'elle n'en utilise qu'une partie pour obtenir la règle d'affectation et fournit en général une estimation *pessimiste* du taux d'erreur.

Toutefois, elle permet d'obtenir des *intervalles de crédibilité (bayésiens)* pour le taux d'erreur réel à partir d'un échantillon test de taille n, tiré selon la même loi que l'échantillon d'apprentissage.

10.2 Intervalles de crédibilité bayésiens

Notons le taux d'erreur réel e_T et k le nombre d'observations mal classifiées (sur l'échantillon test). k est une variable aléatoire fonction de Y et de n.

Sachant e_T, k est de loi binômiale

$$p(k|e_T) = C_n^k e_T^k (1 - e_T)^{n-k}$$

A l'aide du théorème de Bayes, on obtient

$$p(e_T|k) = \frac{p(k|e_T)p(e_T)}{\int p(k|e_T)p(e_T)de_T}$$

En supposant que $p(e_T)$ ne varie pas en fonction de e_T (loi a priori uniforme) et étant donné que $p(k|e_T)$ correspond à la loi binômiale, on obtient que e_T est de loi a posteriori Beta de paramètres $k+1$, $n-k+1$:

$$p(e_T|k) = \frac{e_T^k (1 - e_T)^{n-k}}{\int e_T^k (1 - e_T)^{n-k} de_T}$$

La densité a posteriori précédente fournit un résumé complet de l'information pouvant être obtenue sur le taux d'erreur à partir de l'échantillon test.

Elle peut être résumée de diverses façons ; une approche standard consiste à en déduire une valeur supérieur et une valeur inférieur pour le taux d'erreur réel.

Plus précisément, pour une valeur donnée de α (typiquement, 0.05), il existe de nombreux intervalles dans lesquels e_T prend sa valeur avec une probabilité $1 - \alpha$ (sachant k). Ceux-ci sont appelés *intervalle de crédibilité* ou *intervalle de confiance bayésiens.*

Parmi ceux-ci, on retient généralement l'intervalle de crédibilité de *densité a posteriori maximale.*

Celui-ci possède la propriété additionnelle qu'en chaque point qu'il contient, la densité a posteriori prend une valeur supérieure aux valeurs qu'elle prend en dehors de l'intervalle.

On peut démontrer qu'il s'agit également de l'intervalle de crédibilité de niveau $1-\alpha$ *d'amplitude la plus faible.*

Formellement, il s'agit de l'intervalle E_α défini par

$$E_\alpha = \{e_T : p(e_T|k) \geq c\}$$

où c est choisie de façon à avoir

$$\int_{E_\alpha} p(e_T|k)de_T = 1 - \alpha$$

Pour des densité *multimodales* E_α peut être composé du *plusieurs intervalles.*

Toutefois, pour la loi Beta, E_α est bien un intervalle unique, de bornes inférieure et supérieure $\epsilon_1(\alpha)$ et $\epsilon_2(\alpha)$, toutes deux fonctions de k et n, vérifiant

$$0 \leq \epsilon_1(\alpha) < \epsilon_2(\alpha) \leq 1$$

Par exemple, pour 4 exemples mal classés sur un échantillon test de taille 20, la région de crédibilité de niveau 95% est $[0.069, 0.399]$.

A partir des considérations précédentes, on peut également déterminer la taille d'échantillon minimale pour obtenir un intervalle de crédibilité de niveau donné, d'amplitude inférieure à un seuil fixé.

10.3 Validation croisée (leave-one-out)

La méthode de validation croisée par leave-one-out consiste à déterminer une règle d'affectation en utilisant $n-1$ exemples parmi n, puis à tester celle-ci sur l'observation restante.

Cette démarche est répétée pour les n sous-ensembles de taille $n-1$ de l'échantillon d'apprentissage.

Pour n grand, la méthode est très gourmande en temps de calcul puisqu'elle requiert l'obtention de n règles d'affectations.

Toutefois, elle est approximativement sans biais, bien que cela soit obtenu au prix d'un accroissement de la variance de l'estimateur du taux d'erreur réel.

Formellement, notons Y_j l'échantillon d'apprentissage obtenu en supprimant l'observation x_j.

L'erreur de validation croisée par leave-one-out est donnée par

$$e_{VC} = \frac{1}{n} \sum_{j=1}^{n} Q(\omega(z_j), \eta(x_j; Y_j))$$

Il s'agit de la proportion d'observations mises de côté à tour de rôle qui sont mal classées par la règle d'affectation obtenue sur les observations restantes.

Une généralisation de la méthode appelée parfois *validation croisée par rotation* est obtenue en partitionnant l'échantillon d'apprentissage en ν échantillons, à effectuer l'apprentissage sur $\nu - 1$ échantillons, puis à tester la règle obtenue sur l'échantillon restant.

La procédure est répétée en prenant pour échantillon test chacun des ν échantillons à tour de rôle.

La validation croisée par rotation constitue un compromis entre la validation croisée par holdout et par leave-one-out.

Le biais est réduit par rapport à la VC holdout et le temps de calcul est réduit par rapport à la VC leave-one-out.

Pour $\nu = n$, on obtient la validation croisée par leave-one-out et pour $\nu = 2$ on obtient une variante de la validation croisée par holdout où échantillon d'apprentissage et

échantillon test sont interchangés à tour de rôle.

11 Le Jackknife en Analyse discriminante

Le jackknife est une procédure permettant d'obtenir un estimateur du taux d'erreur *réel* moins biaisé que le taux d'erreur apparent.

Comme estimateur du taux d'erreur réel, le taux d'erreur apparent a un biais d'ordre $1/n$ pour un échantillon d'apprentissage de taille n.

La méthode du Jackknife permet de réduire le biais à un terme en $1/n^2$.

Soit t_n une statistique quelconque basée sur n observations $x_1, \ldots, x_n$.

On fait l'hypothèse que pour m suffisamment grand, on a

$$\mathbb{E}(t_m) = \theta + \frac{a_1(\theta)}{m} + \frac{a_2(\theta)}{m^2} + O(m^{-3})$$

où θ est une constante et où a_1 et a_2 ne dépendent pas de m.

Soit $t_n^{(j)}$ la même statistique calculée à partir des mêmes observations à l'exclusion de x_j et soit

$$t_n^{(.)} = \frac{1}{n} \sum_{j=1}^{n} t_n^{(j)}$$

On obtient alors

$$\begin{aligned} \mathbb{E}[t_n^{(.)}] &= \frac{1}{n} \sum_{j=1}^{n} \left(\theta + \frac{a_1(\theta)}{n-1} + O(n^{-2}) \right) \\ &= \theta + \frac{a_1(\theta)}{n-1} + O(n^{-2}) \end{aligned}$$

Par conséquent, la statistique

$$t_J = n t_n - (n-1) t_n^{(.)}$$

a un biais d'ordre n^{-2}.

t_J est l'*estimateur par jackknife* déduit de t_n.

On va appliquer à présent ces considérations à l'estimation du taux d'erreur.

L'estimateur par jackknife déduit du taux d'erreur apparent est donné par

$$\begin{aligned} e_J^0 &= n e_A - (n-1) e_A^{(.)} \\ &= e_A + (n-1)(e_A - e_A^{(.)}) \end{aligned}$$

Ici, $e_A^{(j)}$ est le taux d'erreur apparent obtenu lorsque la j-ème donnée est exclue de l'échantillon d'apprentissage :

$$e_A^{(j)} = \frac{1}{n-1} \sum_{k=1, k \neq j}^{n} Q(\omega(z_k), \eta(x_k; Y_j))$$

Comme estimateur du taux d'erreur *espéré* (qui est constant), le biais de e_J^0 est de l'ordre de n^{-2}, mais comme estimateur du taux d'erreur *réel* (qui est constant à Y fixé), il est toujours d'ordre n^{-1}.

Afin de réduire le biais de e_J^0, comme estimateur du taux d'erreur *réel*, à un terme du second ordre, on utilise

$$e_J = e_A + (n-1)(\widetilde{e}_A - e_A^{(.)})$$

où

$$\widetilde{e}_A = \frac{1}{n^2} \sum_{j=1}^{n} \sum_{k=1}^{n} Q(\omega(z_k), \eta(x_k; Y_j))$$

12 Le Bootstrap en Analyse Discriminante

Le bootstrap est une technique permettant d'estimer le biais du taux d'erreur apparent, comme estimateur du taux d'erreur réel, à échantillon d'apprentissage fixé, et donc de le corriger.

Il est implémenté de la façon suivante.

Notons nos données $Y = \{(x_i', z_i)' : i = 1, \ldots, n\}$.

Soit $\widehat{F}$ la *fonction de répartition empirique* des x_i.

Dans le cas d'un échantillonnage global, c'est la distribution affectant la masse $1/n$ à chaque observation x_i.

Dans le cas d'un échantillonnage par classe, notons $\widehat{F}_i$ la distribution affectant la masse $1/n_i$ à chaque point x_k de la classe ω_i, où n_i désigne l'effectif de la classe ω_i.

1. Générer un nouvel échantillon de données (appelé *échantillon bootstrap*) $Y^b = \{(\widetilde{x}_i', \widetilde{z}_i')' : i = 1, \ldots, n\}$ selon la loi empirique correspondant au plan d'échantillonnage.

2. Obtenir une règle d'affectation sur la base de Y^b.

3. Calculer le taux d'erreur apparent sur Y^b ; on le notera $\widetilde{e}_A$.

4. Calculer le taux d'erreur réel pour la règle d'affectation obtenue (en considérant Y comme population mère) ; on le notera $\widetilde{e}_c$.
5. Calculer $w_b = \widetilde{e}_A - \widetilde{e}_c$

6. Répéter les étapes 1 à 5 B fois

7. L'estimation du biais du taux d'erreur apparent par bootsrap est donnée par

$$W_{\text{boot}} = \mathbb{E}[\widetilde{e}_A - \widetilde{e}_c]$$

   où l'espérance est relative au plan d'échantillonnage des Y^b, ainsi

$$W_{\text{boot}} = \frac{1}{B} \sum_{b=1}^{B} w_b$$

8. La version corrigée du taux d'erreur apparent est donnée par

$$e_A^{(B)} = e_A - W_{\text{boot}}$$

A l'étape 1, dans le cas d'un échantillonnage global, n échantillons indépendants sont tirés selon la loi $\widehat{F}$; certains peuvent apparaître plusieurs fois dans Y^b et il peut arriver qu'une ou plusieurs classes ne soient pas représentées.
Dans le cas d'un échantillonnage par classes, n_i échantillons sont générés selon chaque loi $\widehat{F}_i$, $i = 1, \ldots, C$; ainsi, toutes les classes sont représentées dans l'échantillon boostrap, dans les mêmes proportions que dans l'échantillon initial.

Le nombre d'échantillons bootstrap B utilisés pour estimer W_{boot} peut être limité en pratique pour des raisons calculatoires.

Dans le cas de l'estimation du taux d'erreur, B est en général compris entre 25 et 100.

Il existe diverses variantes de l'approche précédente, notamment, le bootstrap double, le bootstrap randomisé et l'estimateur 0.632 (faire des recherches sur ceux-ci à titre d'exercice).

13 Fiabilité des méthodes de discrimination

La *fiabilité* également appelée *imprécision* d'une règle de décision quantifie la qualité de l'approximation des probabilités a posteriori d'appartenance aux différentes classes obtenue à l'aide de cette règle.

Ainsi, on ne s'intéresse pas simplement à la classe ω_i pour laquelle $p(\omega_j|x)$ est maximale, mais aux valeurs $p(\omega_j|x)$ elle-mêmes.

Il n'est pas facile d'estimer la fiabilité d'une règle de décision, et ce pour deux raisons :

1. Les vraies probabilités a posteriori sont inconnues
2. Certaines règles de décision ne founrissent pas d'estimations *explicites* des probabilités a posteriori.

Une règle de décision peut avoir un bon pouvoir discriminant tout en ayant une fiabilité faible.

Dans le problème à deux classes, il suffit pour celà par exemple qu'elle utilise le seuil de décision optimal, tout en estimant mal les probabilités a posteriori.

Pourquoi est-il souhaitable d'avoir une bonne fiabilité ? Un bon pouvoir discriminant ne suffit-il pas ?

Dans certains cas, il est en effet suffisant d'obtenir une règle de décision avec un bon pouvoir discriminant, par exemple une règle de décision qui permette d'atteindre l'erreur optimale (de Bayes).

Toutefois, si l'on souhaite prendre une décision basée sur des considérations de *coût*

ou si l'on souhaite utiliser les résultats de l'analyse discriminante pour des analyses subséquentes, il est important d'avoir une règle *fiable*.

Une mesure classique de fiabilité est la suivante :

$$R = \sum_{i=1}^{C} \frac{1}{n} \sum_{k=1}^{n} \{\phi_i(x_k)[z_{ik} - \widehat{p}(\omega_i|x_k)]\}$$

où $z_{ik} = 1$ si $x_k \in \omega_i$ et 0 sinon, et où ϕ_i est une statistique de test ; par exemple, on peut prendre

$$\phi_i(x_k) = (1 - \widehat{p}(\omega_i|x_k))^2.$$

Une autre façon de quantifier la fiabilité d'une règle de décision et d'obtenir des intervalles de confiance pour les probabilités à posteriori d'appartenance aux différentes classes.

A ce propos, voir l'ouvrage de McLachlan : Discriminant Analysis and Statistical Pattern Recognition, éditions Wiley 1992.

14 Les courbes ROC et leur utilisation

La courbe ROC (Receiver Operating Characteristic) a été introduite lors du premier cours, dans le contexte de la règle de décision de Neyman-Pearson pour le problème à deux classes, comme moyen de visualiser la performance d'une règle de décision afin de sélectionner un *seuil* de décision.

La courbe ROC représente le taux de vrais positifs $1 - \epsilon_1$ (TPR) sur l'axe des ordonnées, en fonction du taux de faux positifs ϵ_2 (FPR) sur l'axe des abscisses.

En épidémiologie, on dit qu'une courbe ROC représente la *sensibilité* d'une règle de décision en fonction de son *anti-spécificité*.

En pratique, la courbe ROC *optimale* (celle obtenue à partir des *vraies* densités conditionnelles $p(x|\omega_i)$) est *inconnue*, tout comme le taux d'erreur réel.

Elle doit être *estimée* à partir d'une règle de décision empirique et d'un échantillon test indépendant.

Comme pour l'estimation du taux d'erreur réel, il est possible d'utiliser des méthodes de types validation croisée ou bootstrap.
Des règles de décision différentes produisent des courbes ROC différentes, caractérisant leur performance.
Toutefois, le plus souvent, on souhaite disposer d'un *nombre* caractérisant la performance d'une règle de décision, plutôt que d'une *courbe*, de façon à pouvoir comparer facilement la performance de diverses règles de décision.
Lors du premier cours, on a vu que la règle de décision optimale était définie sur la base du rapport de vraisemblance ; sous l'hypothèse que le coût d'une classification correcte est nul et que les coûts des erreurs de type I et de type II sont égaux, x est affecté à la classe ω_1 si

$$\frac{p(x|\omega_1)}{p(x|\omega_2)} \geq \frac{p(\omega_2)}{p(\omega_1)}$$

autrement dit si :

$$p(\omega_1|x) \geq 0.5$$

La règle de décision optimale correspond donc à un point de la courbe ROC déterminé par les coût relatifs des mauvaises classifications et par les probabilités a priori.
Le coût optimal (perte optimale) est donné par

$$L = p(\omega_1)\epsilon_1 + p(\omega_2)\epsilon_2$$

où ϵ_i désigne la probabilité de mal classer un objet appartenant à la classe ω_i.
La courbe ROC représente $1 - \epsilon_1$ en fonction de ϵ_2.
Dans le plan de la courbe ROC (*i.e.* le plan $(\epsilon_2, 1 - \epsilon_1)$), les droites de coût constant, appelées *droites d'iso-performance* sont les droites parallèles de pente $p(\omega_2)/p(\omega_1)$.
Le coût correspondant croît du coin supérieur gauche au coin inférieur droit du plan.
Les valeurs réalisables du coût correspondent aux droites intersectant la courbe ROC.
La solution de coût *minimum* correspond à la droite d'iso-performance *tangente* à la courbe ROC. Au point de tangence, la courbe ROC a une dérivée égale à $p(\omega_2)/p(\omega_1)$. On obtient alors la règle optimale de Bayes).

Pour différentes valeurs des coûts relatifs et des probabilités a priori, les courbes d'iso-performance sont en général de pentes différentes et le coût minimum est atteint en un point différent de la courbe ROC.

Une stratégie alternative consiste à comparer la distribution globale de $\widehat{p}(x) \equiv p(\omega_1|x)$ pour des échantillons issus respectivement des classes ω_1 et ω_2.

On s'attend à ce que les valeurs de $p(\omega_1|x)$ soient plus élevées pour les observations issues de ω_1 que pour les observations issues de ω_2.

En général, plus l'écart entre les deux est important, meilleure est la règle de décision.

Une mesure de la séparation de ces deux distributions est l'*aire* sous la courbe ROC (AUC). Celle-ci fournit une valeur numérique unique, basée sur la courbe ROC et qui ne prend pas en compte les coûts λ_{ij}.

Celle-ci ignore de ce fait l'information dont on peut disposer sur le coût relatif des erreurs de type I et de type II et elle est indépendante du seuil de décision retenu.

Cet avantage peut devenir un désavantage lors de la *comparaison* de différentes règles. En effet, si deux courbes ROC *se croisent*, chacune sera, en général, supérieure à l'autre pour *certaines* valeurs du seuil. L'AUC ne rend pas compte de ce fait.

L'aire sous la courbe ROC est généralement calculée en appliquant la règle de décision à un échantillon test.

Pour une règle de décision qui produit des estimations de $p(\omega_1|x)$ de façon directe, on peut obtenir des valeurs $\{f_1, \ldots, f_{n_1}; f_i = p(\omega_1|x_i), x_i \in \omega_1\}$ et $\{g_1, \ldots, g_{n_2}; g_i = p(\omega_1|x_i), x_i \in \omega_2\}$.

On les utilise ensuite pour obtenir une mesure de séparation des distributions de $\widehat{p}(x)$ pour les observations issues de ω_1 et ω_2 respectivement.

On commence par ordonner les estimations $\{f_1, \ldots, f_{n_1}, g_1, \ldots, g_{n_2}\}$ par ordre croissant. On note r_i le rang de la i-ème observation issue de ω_1.

L'AUC est alors estimée par

$$\widehat{A} = \sum_{i=1}^{n_1} r_i - \frac{1}{2} n_1(n_1 + 1)$$

Une approche alternative consiste à obtenir une estimation de la courbe ROC à partir

de la règle de décision retenue en faisant varier un seuil, puis à utiliser une technique d'intégration numérique (par exemple, la méthode des trapèzes).
Plus l'AUC est élevée, plus la règle de décision est meilleure. Une règle de décision qui n'apporte rien par rapport à une affectation au hasard produit une courbe ROC approximativement confondue avec la première bissectrice (AUC=1/2).

15 Régression logistique

15.1 Modèle dichotomique

Plaçons-nous tout d'abord dans le cas de deux classes ω_1, ω_2.
L'hypothèse de base de la discrimination logistique est que la différence entre les logarithmes des densités de classes est une fonction affine des variables explicatives x.
Formellement, on écrit

$$\log\left(\frac{p(x|\omega_1)}{p(x|\omega_2)}\right) = \beta_0 + \beta' x$$

Ce modèle est exact dans diverses situations, notamment lorsque :

1. les densités conditionnelles sont gaussiennes, de matrices de variances-covariances identiques

2. les densités conditionnelles sont discrètes multivariées et conformes à un modèle log-linéaire avec des termes d'interactions entres groupes égaux

3. les cas 1 et 2 sont combinés : certaines variables explicatives sont continues et d'autres discrètes

L'hypothèse est donc vérifiées par un grand nombre de densités ; de plus, il a été vérifiée *empiriquement* qu'elle était souvent pertinente pour des données réelles non gaussiennes.

On démontre facilement que l'hypothèse faite équivaut à

$$p(\omega_2|x) = \frac{1}{1 + \exp(\beta_0^* + \beta' x)}$$

et

$$p(\omega_1|x) = \frac{\exp(\beta_0^* + \beta' x)}{1 + \exp(\beta_0^* + \beta' x)}$$

où

$$\beta_0^* = \beta_0 + \log \frac{p(\omega_1)}{p(\omega_2)}$$

Or la règle de Bayes consiste à affecter x à ω_1 si

$$\frac{p(\omega_1|x)}{p(\omega_2|x)} \geq 1$$

et à ω_2 sinon.

Sous notre hypothèse, on voit que cette règle se ramène à affecter x à ω_1 si

$$\beta_0^* + \beta' x \geq 0$$

et à ω_2 sinon.

On obtient donc ainsi une fonction discriminante *linéaire*.

Celle-ci étant obtenue à partir d'expressions *paramétriques* des densités conditionnelles, on peut utiliser la méthode du *maximum de vraisemblance* pour estimer les paramètres y apparaissant.

15.2 Maximum de vraisemblance

Afin d'estimer les paramètres de la fonction discriminante logistique par maximum de vraisemblance, on peut utiliser un algorithme d'optimisation non linéaire à partir d'expressions de la fonction de vraisemblance et de ses dérivées.

L'analyse de la procédure d'estimation doit faire intervenir le *plan d'échantillonnage* utilisé pour obtenir les données observées.

Trois types de plans d'échantillonnage sont particulièrement utilisés :

1. Echantillonnage dans la population globale ; cela revient à prélever un échantillon issu de la distribution de mélange ;

2. Echantillonnage conditionnellement à x ; x est fixé et un ou plusieurs échantillons sont obtenus pour chaque valeur de x ; ceux-ci peuvent appartenir à ω_1 ou ω_2 ;

3. Echantillonnage dans chaque classe séparément ; cela revient à prélever des échantillons issus de chacune des densités conditionnelles $p(x|\omega_i)$ $(i = 1, 2)$.

Il est possible de démontrer que les estimations par maximum de vraisemblance sont en fait indépendantes du plan d'échantillonnage retenu, bien que le plan 3 ne permette d'obtenir qu'une estimation de β_0 et pas de β_0^*, qui est le coefficient requis pour la discrimination.

On suppose dans la suite que le plan d'échantillonnage est de type 1, c'est-à-dire qu'on prélève un échantillon dans la population globale.

On admettra que vraisemblance conditionnelle des observations est donnée par

$$L = \prod_{r=1}^{n_1} p(\omega_1|x_{1r}) \prod_{r=1}^{n_2} p(\omega_2|x_{2r})$$

où x_{ir} désigne la r-ème observation de la classe ω_i.

Ainsi, maximiser L, revient à maximiser

$$\ell = \sum_{r=1}^{n_1} \log(p(\omega_1|x_{1r})) + \sum_{r=1}^{n_2} \log(p(\omega_2|x_{2r}))$$

c'est-à-dire

$$\ell = \sum_{r=1}^{n_1} (\beta_0^* + \beta' x_{1r}) - \sum_x \log\{1 + \exp(\beta_0^* + \beta' x)\}$$

où $\sum_x$ désigne la somme sur l'ensemble des observations.

Le gradient de ℓ relativement aux paramètres est donné par

$$\begin{aligned} \frac{\partial \ell}{\partial \beta_0^*} &= n_1 - \sum_x p(\omega_1|x) \\ \frac{\partial \ell}{\partial \beta_j} &= \sum_{r=1}^{n_1} (x_{1r})_j - \sum_x p(\omega_1|x) x_j \end{aligned}$$

pour $j = 1, \ldots, d$.

Disposant d'une expression formelle de la log-vraisemblance et de ses dérivées, on peut utiliser un algorithme d'optimisation non-linéaire pour déterminer un maximum local de ℓ.

On peut prend généralement comme condition initiale de l'algorithme un vecteur de paramètres nul.

A deux exceptions près, la vraisemblance admet un maximum unique atteint en un β fini.

Ces deux exceptions sont les suivantes :

1. Les deux classes sont *linéairement séparables*. L'estimation par maximum de vraisemblance est alors mal définie.

   Il est facile de vérifier à chaque itération de l'algorithme d'optimisation si $\beta_0^* + \beta' x$ sépare les deux classes ou non. Lorsque c'est le cas, on peut arrêter l'algorithme.

2. Les données sont discrètes et l'effectif d'une modalité est nul dans l'une des deux classes. Dans ce cas également, l'estimation par maximum de vraisemblance est mal définie.

   Certaines hypothèses supplémentaires peuvent permettre de traiter ce dernier cas : voir l'article d'Anderson dans Applied Statistics 23 (1974), pages 397-404.

15.3 Modèle polytomique

Dans le cas du problème à C classes, on fait l'hypothèse fondamentale que

$$\log\left(\frac{p(x|\omega_i)}{p(x|\omega_C)}\right) = \beta_{i0} + \beta_i' x$$

pour $i = 1, \ldots, C-1$.

On peut démontrer, comme dans le cas dichotomique, que les probabilités a posteriori des classes sont de la forme

$$p(\omega_i|x) = \frac{\exp(\beta^*_{i0} + \beta'_i x)}{1 + \sum_{k=1}^{C-1} \exp(\beta^*_{k0} + \beta'_k x)}$$

pour $i = 1, \ldots, C-1$ et

$$p(\omega_C|x) = \frac{1}{1 + \sum_{k=1}^{C-1} \exp(\beta^*_{k0} + \beta'_k x)}$$

avec

$$\beta^*_{i0} = \beta_{i0} + \log(p(\omega_i)/p(\omega_C))$$

Les fonctions discriminantes obtenues sont linéaires et la règle de Bayes devient : affecter x à la classe ω_i si

$$\beta^*_{i0} + \beta'_i x = \max\{\beta^*_{k0} + \beta'_k x\} > 0$$

sinon affecter x à ω_C.

On admettra que la vraisemblance conditionnelle des observations est donnée par

$$L = \prod_{i=1}^{C} \prod_{r=1}^{n_i} p(\omega_i|x_{ir})$$

avec les notations du cas dichotomique.

Maximiser L équivaut à maximiser

$$\ell = \sum_{i=1}^{C} \sum_{r=1}^{n_i} \log(p(\omega_i|x_{ir}))$$

Les dérivées de ℓ sont données par

$$\begin{aligned} \frac{\partial \ell}{\partial \beta^*_{i0}} &= n_i - \sum_x p(\omega_i|x) \\ \frac{\partial \ell}{\partial (\beta_i)_l} &= \sum_{r=1}^{n_i} (x_{ir})_l - \sum_x p(\omega_i|x) x_l \end{aligned}$$

où $\sum_x$ désigne la somme sur l'ensemble des observations.

Comme dans le cas dichotomique, si les classes sont séparables, l'estimation par maximum de vraisemblance est mal définie ; l'algorithme peut être arrêté dès lors qu'il y a séparation complète.

Dans le cas discret, s'il existe des modalités de fréquence nulle, l'estimation par maximum de vraisemblance est mal définie ; on peut alors utiliser la procédure d'Anderson (1974).

16 La régression logistique sous SAS

16.1 Tests d'hypothèses

On peut réécrire l'hypothèse de base de la régression logistique pour le problème à deux classes sous la forme

$$\begin{aligned} \log \frac{p(\omega_1|x)}{p(\omega_2|x)} &= \beta_0^* + \beta' x \\ &= \beta_0^* + \beta_1 x_1 + \cdots + \beta_d x_d \end{aligned}$$

On a vu comment estimer les coefficients de ce modèle par maximum de vraisemblance. De même pour le modèle polytomique.
Soit l'hypothèse nulle ($H_0 : \gamma = 0$) où γ est un vecteur de coefficients de dimension $q \leq d+1$. On notera α les coefficients restants.
Les tests disponibles sous SAS pour tester cette hypothèse sont les suivants.

16.1.1 Test de Wald

Sous H_0,

$$W = \widehat{\gamma}' \Sigma_{(1)}^{-1}(\widehat{\beta}) \widehat{\gamma}$$

suit asymptotiquement une loi du χ^2 à q degrés de liberté, où $\Sigma_{(1)}^{-1}(\widehat{\beta})$ est la sous-matrice issue de $\Sigma^{-1}(\widehat{\beta})$ associée aux q coefficients de γ.
$\Sigma(\widehat{\beta})$ est calculée à l'aide de la formule

$$\Sigma(\widehat{\beta}) = -H^{-1}(\widehat{\beta})$$

où $H(\beta)$ désigne la matrice *hessienne* de la log-vraisemblance.

16.1.2 Test du rapport de vraisemblance

On considère le rapport

$$R = \frac{L(\widehat{\alpha}, 0)}{L(\widehat{\beta})}$$

où $\beta = (\alpha, \gamma)$ (dans le désordre) et où $(\widehat{\alpha}, 0)$ est l'estimateur du maximum de vraisemblance de β sous l'hypothèse H_0.

Sous l'hypothèse H_0, $G \equiv -2 \log R$ suit asymptotiquement une loi du χ^2 à q degrés de liberté. La statistique G est appelée *déviance*.

16.1.3 Test du score

Il est basé sur la distribution du gradient de $\log L(\beta)$ par rapport à β.

Sous H_0, la statistique

$$\mathrm{SC} = \left\{ \frac{\partial \log L(\beta)}{\partial \beta} \right\}'_{\beta=(\widehat{\alpha},0)} [\Sigma(\beta)]_{\beta=(\widehat{\alpha},0)} \left\{ \frac{\partial \log L(\beta)}{\partial \beta} \right\}_{\beta=(\widehat{\alpha},0)}$$

suit asymptotiquement une loi du χ^2 à q degrés de liberté.

Remarques. Le test du rapport de vraisemblance est le plus fiable en pratique.

Le test de Wald ne nécessite pas l'estimation des paramètres sous H_0 et le test du score ne nécessite l'estimation des paramètres que sous H_0.

Ces deux derniers tests sont les plus avantageux en termes de temps de calcul.

16.2 Critères de validité générale

Des critères de validité du modèle ont été proposé par divers auteurs et intégrés dans SAS.

Le critère d'Akaiké :

$$\mathrm{AIC} = -2 \log L(\widehat{\beta}) + 2(d+1)$$

Le critère de Schwartz :

$$\mathrm{SC} = -2 \log L(\widehat{\beta}) + (d+1) \log n$$

où n est le nombre total d'observations.

Ces deux critères sont utilisés pour comparer deux modèles différents portant sur le même type de données.

Le modèle pour lequel ces statistiques sont les plus faibles sera préféré.

16.3 Odds et Odds ratio

On se place dans le cadre de la régression logistique pour le problème à deux classes *avec une seule variable explicative continue*.

Le modèle s'écrit donc

$$\log \frac{p(\omega_1|x)}{p(\omega_2|x)} = \beta_0^* + \beta_1 x$$

On appelle *odds, côte* ou *chance* de la classe ω_1 (codée 1/event) le rapport

$$o(\omega_1|x) = \frac{p(\omega_1|x)}{1 - p(\omega_1|x)} = \frac{p(\omega_1|x)}{p(\omega_2|x)}$$

On appelle *rapport de chance* ou *odds ratio* le rapport

$$\psi = \frac{o(\omega_1|x+1)}{o(\omega_1|x)}$$

Il mesure l'influence d'un accroissement d'une unité de x sur la chance d'appartenir à ω_1 (event) plutôt qu'à ω_2 (non-event).

Un calcul simple montre que

$$\log \frac{p(\omega_1|x+1)}{p(\omega_2|x+1)} - \log \frac{p(\omega_1|x)}{p(\omega_2|x)} = \beta_1$$

d'où

$$\psi = \exp \beta_1$$

et

$$\widehat{\psi} = \exp \widehat{\beta}_1$$

Lorsque x n'est pas continue mais dichotomique, de valeurs $a < b$, on démontre par un calcul analogue que

$$\psi = \frac{o(\omega_1|x=b)}{o(\omega_1|x=a)} = \exp(\beta_1(b-a))$$

d'où

$$\widehat{\psi} = \exp(\widehat{\beta}_1(b-a))$$

Dans le cas de plusieurs variables explicatives, un odds ratio peut être calculé pour chaque variable *en fixant la valeur des autres variables explicatives.*

SAS détermine également un *intervalle de confiance* pour chaque odds ratio calculé. Si la borne inférieure de cet intervalle de confiance est *supérieure à 1*, on dit que *la variable est significative.* Une valeur plus élevée de la variable accroît la probabilité d'appartenir à ω_1 (event).

16.4 Table d'association

SAS fournit également une table d'association entre probabilité prédites et réponses observées.

Celle-ci est obtenue de la façon suivante.

On considère toutes les paires possibles d'observations de l'échantillon pour lesquelles la classe d'appartenance est différente. On a donc $n_1 \times n_2$ paires possibles.

On estime la probabilité pour que chaque observation d'une paire donnée appartienne à ω_1.

On affecte l'observation pour laquelle la probabilité est la plus grande à ω_1 et l'autre à ω_2. Si l'affectation obtenue et l'appartenance effective sont identiques, on dit qu'il y a concordance ; sinon, il y a discordance. Si les deux probabilités sont identiques, on a des ex-aequo ("tied").

Pour que le modèle soit bon, il faut que le pourcentage de concordants soit très élevé et le pourcentages de discordants très petit.

Le pourcentage de concordants correspond à l'aire sous la courbe ROC (AUC). Celle-ci est également fournie par la statistique c de la table d'association.

16.5 Table de classification

La table de classification fournit les matrices de confusion pour différentes valeurs du seuil de décision (sur $p(\omega_1|x)$) ainsi que le taux de faux positifs et le taux de vrais positifs correspondants.
Pour l'obtenir on peut utiliser l'option **ctable** de l'instruction **model** de la **proc logistic**.
Les données fournies permettent de choisir un seuil de décision et de construire une courbe ROC.

17 Régression logistique sous R

Sous R, la régression logistique est implémentée par la fonction **glm** (cf. cours de Modèles Linéaires Généralisés en Master 1 DSMS).
La régression logistique polytomique est implémentée par les fonctions **multinom** (package **nnet**) et **mlogit** (package **mlogit**).

18 Bibliographie

1. C.-A. Azencott, *Introduction au Machine Learning*, Dunod 2018.
2. M. Bardos. *Analyse discriminante*, Dunod 2001.
3. G. Celeux, J.-P. Nakache. *Analyse discriminante sur variables qualitatives*, Polytechnica 1994.
4. B. Dubuisson. *Diagnostic et reconnaissance des formes*, Hermes 1990.
5. J.-P. Nakache, J. Confais. *Statistique explicative appliquée*, Technip 2003.
6. B. D. Ripley. *Pattern recognition and Neural networks*, Cambridge UP 2002.
7. G. Saporta. *Probabilités, Analyse des Données et Statistique*, Technip 2011.
8. M. Tenenhaus. *Statistique, Méthodes pour décrire, expliquer et prévoir*, Dunod 2007.
9. A. Webb. *Statistical Pattern Recognition*, 2nd ed., Wiley 2002.