

Scoring

Marie Chavent

<http://www.math.u-bordeaux.fr/~machaven/>

2014-2015

1 Introduction

L'idée générale est d'affecter une note (un score) global à un individu à partir de notes partielles. A partir de cette note, on affectera cet individu à un groupe.

Un score peut être défini comme un outil statistique ou probabiliste de detection de risque. On peut citer entre autre :

→ Les scores de risque de type :

- risque de credit ou credit scoring : prédire le retard de remboursement de crédit.
- risque financier : prédire la bonne ou mauvaise santé d'une entreprise.
- risque médical.

→ Les scores en marketing :

- score d'attrition (churn en anglais) : prédire le risque qu'un client passe à la concurrence ou résilie son abonnement.
- score d'appétence : prédire l'appétence d'un client à acheter tel ou tel type de produit.

Les principales étapes de la construction d'un score sont :

1. Choix de la variable à expliquer Y . On parle aussi de critère à modéliser ou encore de variable cible. La variable Y sera qualitative généralement binaire à deux modalités. Chaque modalité définit un groupe.
2. Choix d'un vecteur de variables explicatives $X = (X^1, \dots, X^p)$. Les p variables explicatives $X^j, j = 1 \dots p$, peuvent être quantitatives ou qualitatives.
3. Choix des données et de l'échantillon. L'échantillon doit être représentatif de la population. On dispose alors d'un échantillon de taille n (nombre d'individus) sur lequel

sont mesurées simultanément les p variables explicatives X^1, \dots, X^p et la variable à expliquer Y .

4. Choix de la méthode de construction du score. Une fonction score $S(X)$ donne une note à un individu en fonction de ses valeurs sur $X = (X^1, \dots, X^p)$. Parmi les méthodes classiques de construction de scores, on peut citer :

→ L'analyse factorielle discriminante (AFD) :

- variables explicatives quantitatives.
- outil descriptif de visualisation des données.

→ L'analyse discriminante géométrique :

- variables explicatives quantitatives.
- affectation d'un nouvel individu au groupe "le plus proche".

→ L'analyse discriminante probabiliste :

- variables explicatives quantitatives
- Y et $X = (X^1, \dots, X^p)$ aléatoires.
- affectation d'un nouvel individu au groupe "le plus probable".
- analyse discriminante linéaire (ADL) ou discriminante linear analysis (LDA) : cas homoscédastique gaussien.
- analyse quadratique discriminante (AQD) ou quadratic discriminant analysis (QDA) : cas hétéroscédastique gaussien.

→ La régression logistique :

- variables explicatives quantitatives ou qualitatives.
- Y binaire et aléatoire et $X = (X^1, \dots, X^p)$ non aléatoire.

5. Construction d'une règle de décision. Dans le cas où Y est binaire, ses deux modalités forment deux groupes d'individus et on peut fixer un seuil c pour obtenir la règle suivante :

$$S(X) \leq c \Rightarrow \text{l'individu est affecté au groupe 1}$$

$$S(X) > c \Rightarrow \text{l'individu est affecté au groupe 2}$$

6. Evaluation du score (courbe ROC, AUC,...) et évaluation de la règle de décision (taux d'erreur, sensibilité, spécificité...) en utilisant les méthodes de l'échantillon test, de validation croisée....

On distingue donc deux problèmes :

1. construire un score,
2. construire une règle de décision.

↔ Exemples des fichiers infarctus.xls, farms.xls, credit.xls

Quelques liens.

<http://cedric.cnam.fr/~saporta/scoring.pdf>

<http://www.modulad.fr/archives/numero-38/Bardos-38/Bardos-38.pdf>

"Analyse discriminante, Application au risque et scoring financier", M. Bardos, Dunod.

<http://www.bentley.edu/centers/sites/www.bentley.edu.centers/files/csbiggs/Desbois.pdf>

<http://www.modulad.fr/archives/numero-30/desbois-30/desbois-30.pdf>

2 Notations et définitions

Les outils statistique de scoring rentrent dans le cadre de la modélisation d'une variable Y qualitative à K modalités à partir de p variables explicatives $X = (X^1, \dots, X^p)$ quantitatives ou qualitatives. On parle généralement dans ce cadre de classification (d'apprentissage) supervisée, chaque modalité de Y représentant une classe (un groupe) d'individus.

Nous supposons disposer d'un échantillon de n observations de Y et de X : sur les n individus de l'échantillon, on a mesuré une variable qualitative à K modalités et p variables quantitatives. En effet, dans ce cours, on ne se place dans le cas où **toutes les variables explicatives sont numériques**. En notant y_i la valeur de la variable à expliquer mesurée sur les i ème individu, on obtient le vecteur $\mathbf{y} = (y_1, \dots, y_n) \in \{1, \dots, K\}^n$. En notant x_i^j la valeur de la j ème variable explicative mesurée sur le i ème individu, on obtient ainsi la matrice de données de dimension $n \times p$ suivante :

$$\mathbf{X} = \begin{array}{c|cccc} & 1 & \dots & j & \dots & p \\ \hline 1 & & & & & \\ \vdots & & & \vdots & & \\ i & \dots & & x_i^j & \dots & \\ \vdots & & & \vdots & & \\ n & & & & & \\ \hline \end{array}$$

Notons :

- $\mathbf{x}_i = (x_i^1, \dots, x_i^p)' \in \mathbb{R}^p$ une ligne de \mathbf{X} décrivant le i ème individu.
- $\mathbf{x}^j = (x_1^j, \dots, x_n^j)' \in \mathbb{R}^n$ une colonne de \mathbf{X} décrivant la j ème variable.
- E_k est le groupe des individus des l'échantillon qui possèdent la modalité k .
- $n_k = \text{card}(E_k)$ est le nombre d'individus qui possèdent la modalité k .

Si les n individus sont affectés des poids p_1, \dots, p_n , (tels que $\forall i = 1, \dots, n, p_i \leq 0$ et $\sum_{i=1}^n p_i = 1$) alors le poids de chaque groupe E_k est :

$$P_k = \sum_{i \in E_k} p_i$$

En général, on prend $p_i = \frac{1}{n}$ et donc $P_k = \frac{n_k}{n}$. On a alors les définitions suivantes :

- Le centre de gravité global est le vecteur de \mathbb{R}^p défini par :

$$\mathbf{g} = \sum_{i=1}^n p_i \mathbf{x}_i = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i.$$

- Le centre de gravité du groupe E_k est le vecteur de \mathbb{R}^p défini par :

$$\mathbf{g}_k = \frac{1}{P_k} \sum_{i \in E_k} p_i \mathbf{x}_i = \frac{1}{n_k} \sum_{i \in E_k} \mathbf{x}_i.$$

- La matrice $p \times p$ de variance-covariance globale est définie par :

$$\mathbf{V} = \sum_{i=1}^n p_i (\mathbf{x}_i - \mathbf{g})(\mathbf{x}_i - \mathbf{g})' = \frac{1}{n} \sum_{i=1}^n (\mathbf{x}_i - \mathbf{g})(\mathbf{x}_i - \mathbf{g})'.$$

- La matrice $p \times p$ de variance-covariance du groupe E_k est définie par :

$$\mathbf{V}_k = \frac{1}{P_k} \sum_{i \in E_k} p_i (\mathbf{x}_i - \mathbf{g}_k)(\mathbf{x}_i - \mathbf{g}_k)' = \frac{1}{n_k} \sum_{i \in E_k} (\mathbf{x}_i - \mathbf{g}_k)(\mathbf{x}_i - \mathbf{g}_k)'.$$

- La matrice $p \times p$ de variance-covariance intra-groupe est définie par :

$$\mathbf{W} = \sum_{k=1}^K P_k \mathbf{V}_k = \sum_{k=1}^K \frac{n_k}{n} \mathbf{V}_k.$$

- La matrice $p \times p$ de variance-covariance inter-groupe est définie par :

$$\mathbf{B} = \sum_{k=1}^K P_k (\mathbf{g}_k - \mathbf{g})(\mathbf{g}_k - \mathbf{g})' = \sum_{k=1}^K \frac{n_k}{n} (\mathbf{g}_k - \mathbf{g})(\mathbf{g}_k - \mathbf{g})'.$$

Cette matrice est la matrice de variance-covariance des p centres de gravités \mathbf{g}_k pondérés par P_k (en général $\frac{n_k}{n}$).

Remarque : en règle générale, \mathbf{W} est inversible tandis que \mathbf{B} n'est pas inversible. En effet, les K centres de gravité g_k sont dans un sous-espace de dimension $K - 1$ de \mathbb{R}^p , et si $K - 1 < p$ (ce qui est généralement le cas), alors la matrice \mathbf{B} (qui est de dimension $p \times p$) n'est pas inversible.

On a les relations suivantes classiques suivantes :

$$\mathbf{g} = \sum_{k=1}^K P_k \mathbf{g}_k = \sum_{k=1}^K \frac{n_k}{n} \mathbf{g}_k, \quad (1)$$

et

$$\mathbf{V} = \mathbf{W} + \mathbf{B}. \quad (2)$$

Les expressions données en (1) et (2) sont de “nature” assez usuelles en statistique :

- moyenne totale (globale) = moyenne (pondérée) des moyennes marginales,
- variance totale = moyenne (pondérée) des variances marginales + variance des moyennes marginales.

Exercice : démontrez (1) et (2).

3 L'analyse factorielle discriminante

L'analyse factorielle discriminante (AFD) est essentiellement descriptive. L'objectif est ici de chercher quelles sont les combinaisons linéaires des variables quantitatives qui permettent de séparer le mieux possible les k catégories et de donner une représentation graphique (comme pour les méthodes factorielles telles que l'analyse en composantes principales) qui rende au mieux compte de cette opération. Cette visualisation sur le ou les plans factoriels appropriés permet aussi de décrire les liaisons entre la variable à expliquer et les p variables explicatives.

3.1 Centrage des données

En AFD comme en analyse en composantes principales (ACP), on suppose que $\mathbf{g} = \mathbf{0}_p$, c'est à dire que les données sont centrées. Cela revient à translater le nuage de points de \mathbf{g} (ancien centre de gravité) vers l'origine. La “structure” du nuage de points n'est pas modifiée,

mais les calculs ultérieurs seront simplifiés. L'écriture des matrices de variance-covariance globale et inter-groupe sont simplifiées :

$$\mathbf{V} = \sum_{i=1}^n \frac{1}{n} \mathbf{x}_i \mathbf{x}_i' \quad \text{et} \quad \mathbf{B} = \sum_{k=1}^K \frac{n_k}{n} \mathbf{g}_k \mathbf{g}_k',$$

et on obtient l'écriture matricielle $\mathbf{V} = \frac{1}{n} \mathbf{X}' \mathbf{X}$.

3.2 Axes, facteurs et variables discriminantes

L'AFD consiste à rechercher de nouvelles variables, qui seront appelées variables discriminantes, correspondant à des directions de \mathbb{R}^p qui séparent le mieux possible en projection les K groupes d'individus.

L'objectif. Il s'agit de trouver une nouvelle variable, combinaison linéaire des variables explicatives, qui “discrimine” au mieux les groupes définis par les modalités de la variable à expliquer. Cette variable notée \mathbf{s} est définie ici comme un vecteur de \mathbb{R}^n , combinaison linéaire des vecteurs $\mathbf{x}^1, \dots, \mathbf{x}^p$ (colonnes de \mathbf{X}) :

$$\mathbf{s} = \mathbf{X}\mathbf{u} = u_1 \mathbf{x}^1 + \dots + u_p \mathbf{x}^p,$$

où $\mathbf{u} = (u_1, \dots, u_p)' \in \mathbb{R}^p$ est le vecteur des coefficients de cette combinaison linéaire. Il va donc falloir définir :

- comment mesurer que \mathbf{s} “discrimine” bien,
- comment trouver \mathbf{u} pour que $\mathbf{s} = \mathbf{X}\mathbf{u}$ “discrimine” au mieux.

Définitions. En AFD, on muni \mathbb{R}^p d'une métrique \mathbf{M} et on projette les n points de \mathbb{R}^p $\mathbf{x}_1, \dots, \mathbf{x}_n$ (les n lignes de \mathbf{X}) sur un axe Δ de vecteur directeur \mathbf{a} . On effectue des projections \mathbf{M} -orthogonales et \mathbf{a} est \mathbf{M} -normé à 1. La liste des coordonnées $s_i = \mathbf{x}_i' \mathbf{M} \mathbf{a}$ des individus sur Δ forme la nouvelle variable \mathbf{s} et on a donc :

$$\mathbf{s} = (s_1, \dots, s_n)' = \mathbf{X} \mathbf{M} \mathbf{a} = \mathbf{X} \mathbf{u},$$

où

- $\mathbf{a} \in \mathbb{R}^p$ est appelé l'axe discriminant ($\mathbf{a}' \mathbf{M} \mathbf{a} = 1$),
- $\mathbf{u} \in \mathbb{R}^p$ est appelé le facteur discriminant ($\mathbf{u}' \mathbf{M}^{-1} \mathbf{u} = 1$),
- \mathbf{s} est appelé la variable discriminante.

Exemple.

3.2.1 Le critère à optimiser

On définit maintenant le critère qui mesurant la capacité d'un axe à discriminer les groupes. Un axe sera discriminant si les groupes sont bien séparés en projection. Pour cela, on utilise la variance intra-groupe et la variance inter-groupe de la variable discriminante \mathbf{s} constituée des coordonnées des projections des individus sur l'axe. On a les définitions suivantes :

— La variance de \mathbf{s} est définie par :

$$\begin{aligned} Var(\mathbf{s}) &= \frac{1}{n} \sum_{i=1}^n (s_i - \bar{s})^2 \\ &= \mathbf{u}'\mathbf{V}\mathbf{u} \end{aligned} \tag{3}$$

où \bar{s} est la moyenne de s .

— La variance intra-groupe de \mathbf{s} est définie par :

$$\begin{aligned} Intra(\mathbf{s}) &= \sum_{k=1}^K \frac{n_k}{n} \sum_{i \in E_k} \frac{1}{n_k} (s_i - \bar{s}_k)^2 \\ &= \mathbf{u}'\mathbf{W}\mathbf{u} \end{aligned} \tag{4}$$

où \bar{s}_k est la moyenne de s dans le groupe k .

— La variance inter-groupe de \mathbf{s} est définie par :

$$\begin{aligned} Inter(\mathbf{s}) &= \sum_{k=1}^K \frac{n_k}{n} (\bar{s} - \bar{s}_k)^2 \\ &= \mathbf{u}'\mathbf{B}\mathbf{u} \end{aligned} \tag{5}$$

Exercice : Démontrez (3) (4) et (5).

On a donc la relation classique $Var(\mathbf{s}) = Intra(\mathbf{s}) + Inter(\mathbf{s})$ qui s'écrit :

$$\mathbf{u}'\mathbf{V}\mathbf{u} = \mathbf{u}'\mathbf{W}\mathbf{u} + \mathbf{u}'\mathbf{B}\mathbf{u} \tag{6}$$

La quantité $\mathbf{u}'\mathbf{V}\mathbf{u}$ étant indépendante des groupes, minimiser $\mathbf{u}'\mathbf{W}\mathbf{u}$ est équivalent à maximiser $\mathbf{u}'\mathbf{B}\mathbf{u}$.

On aura une bonne discrimination des groupes si :

- les centres de gravité projetés sont bien éloignés i.e. $Inter(\mathbf{s}) = \mathbf{u}'\mathbf{W}\mathbf{u}$ est maximum,
- les groupes projetés ne sont pas trop dispersés i.e. $Intra(\mathbf{s}) = \mathbf{u}'\mathbf{B}\mathbf{u}$ est minimum.

Le critère à maximiser,

$$\frac{\mathbf{u}'\mathbf{B}\mathbf{u}}{\mathbf{u}'\mathbf{V}\mathbf{u}} \in [0, 1], \quad (7)$$

est la proportion de la variance de \mathbf{s} expliquée par \mathbf{y} encore appelé le rapport de corrélation entre \mathbf{s} et \mathbf{y} . Ce critère désigne également le pouvoir discriminant de l'axe Δ . Dans le cadre de la discrimination parfaite, ce rapport serait égal à 1. Plus généralement, la discrimination est d'autant meilleure que le ratio est proche de 1.

3.2.2 Solution du problème d'optimisation

On veut maximiser le rapport entre la variance inter-groupe et la variance totale. La première variable discriminante $\mathbf{s} = \mathbf{X}\mathbf{u}$ est telle que le facteur discriminant \mathbf{u} vérifie :

$$\max_{\mathbf{u} \in \mathbb{R}^p} \frac{\mathbf{u}'\mathbf{B}\mathbf{u}}{\mathbf{u}'\mathbf{V}\mathbf{u}}. \quad (8)$$

Rappels sur la maximisation du quotient de deux formes quadratiques. Soient \mathbf{A} et \mathbf{B} deux matrices symétriques de même dimension, \mathbf{B} étant supposée inversible. Alors, le rapport $\frac{\mathbf{u}'\mathbf{A}\mathbf{u}}{\mathbf{u}'\mathbf{B}\mathbf{u}}$ est maximal pour \mathbf{u}_1 vecteur propre de $\mathbf{B}^{-1}\mathbf{A}$ associé à la plus grande valeur propre notée λ_1 , la valeur du maximum étant λ_1 .

Exercice : Démontrez le résultat ci-dessus.

On déduit du rappel ci-dessus que, pour notre problème d'optimisation (8), le vecteur \mathbf{u} solution est le vecteur propre de $\mathbf{V}^{-1}\mathbf{B}$ associé à la plus grande valeur propre λ_1 .

En pratique, si on utilise la métrique $\mathbf{M} = \mathbf{V}^{-1}$ par exemple, on peut choisir parmi tous les vecteurs propres colinéaires, celui qui est \mathbf{V} -normé à 1 ($\mathbf{u}'\mathbf{V}\mathbf{u} = Var(\mathbf{s}) = 1$). Si on utilise une autre métrique \mathbf{M} , l'axe discriminant $\mathbf{a} = \mathbf{M}^{-1}\mathbf{u}$ est modifié mais le facteur discriminant \mathbf{u} est le même (à une constante près) et la variable discriminante $\mathbf{s} = \mathbf{X}\mathbf{M}\mathbf{a} = \mathbf{X}\mathbf{u}$ (coordonnées des projections des individus sur la droite engendrée par l'axe discriminant) est donc également la même (à une constante près). La valeur du maximum, c'est à dire le pouvoir discriminant,

quand à lui est inchangé. Le choix de la métrique \mathbf{M} est donc indifférente pour la construction de variables et des facteurs discriminants, seule la norme change. Avec la métrique $\mathbf{M} = \mathbf{W}^{-1}$ par exemple, si \mathbf{u} est \mathbf{W} -normé à 1, alors $\mathbf{u}'\mathbf{W}\mathbf{u} = \text{Intra}(\mathbf{s}) = 1$.

On a alors les définitions suivantes :

- Le premier facteur discriminant est \mathbf{u}_1 , le premier vecteur propre de $\mathbf{V}^{-1}\mathbf{B}$.
- La première variable discriminante est $\mathbf{s}_1 = \mathbf{X}\mathbf{u}_1$.
- Le pouvoir discriminant est $\lambda_1 = \frac{\mathbf{u}_1'\mathbf{B}\mathbf{u}_1}{\mathbf{u}_1'\mathbf{V}\mathbf{u}_1}$.

Remarques :

- Cas où $\lambda_1 = 1$: en projection sur la droite discriminante Δ , les dispersions intra-groupe sont nulles ; cela correspond au fait que les K sous-nuages sont donc chacun dans des hyperplans orthogonaux à Δ . Il y a alors évidemment discrimination parfaite si les centres de gravités se projettent en des points différents sur la droite Δ .
- Cas où $\lambda_1 = 0$: ici le meilleur axe discriminant ne permet pas de séparer les K centres de gravité g_k . C'est par exemple le cas si ces derniers sont confondus ; cela peut correspondre au cas de figure où les K sous-nuages sont concentriques et donc aucune séparation linéaire n'est possible. Remarquons qu'il peut cependant exister une possibilité de discrimination non linéaire (du type quadratique par exemple, si l'on pense à la distance au centre de gravité qui permet de séparer les groupes dans le cas de sous-nuages concentriques).
- Notons enfin que la valeur propre λ est une mesure pessimiste du pouvoir discriminant. En effet, même s'il est possible de discriminer parfaitement les groupes, on peut avoir $\lambda < 1$.

La décomposition spectrale de $\mathbf{V}^{-1}\mathbf{B}$ donne aussi les autres axes discriminants de l'AFD. En effet, on cherche ensuite $\mathbf{s}_2 = \mathbf{X}\mathbf{u}_2$, non corrélée à \mathbf{s}_1 , telle que $\frac{\mathbf{u}_2'\mathbf{B}\mathbf{u}_2}{\mathbf{u}_2'\mathbf{V}\mathbf{u}_2}$ soit maximum et ainsi de suite. On peut montrer que les vecteurs propres de $\mathbf{V}^{-1}\mathbf{B}$ sont les facteurs discriminants et les valeurs propres $\lambda_1, \dots, \lambda_{K-1}$ sont les pouvoirs discriminants.

On pourra construire $K - 1$ axes discriminants car le rang de $\mathbf{V}^{-1}\mathbf{B}$ est au plus $\min(p, K - 1)$ et on suppose que $K - 1 < p$.

Remarques :

- Il faut que la matrice de variance-covariance \mathbf{V} soit inversible et que donc la matrice de données \mathbf{X} soit de plein rang. Il faudra donc faire attention en AFD, comme en

régression linéaire multiple, au problème de la multicolinéarité des colonnes de \mathbf{X} .

- L'AFD peut être vue comme une ACP (Analyse en composante principale) des centres de gravités g_k avec la métrique \mathbf{V}^{-1} .
- L'AFD est aussi un cas particulier de l'analyse canonique. Elle correspond en effet à l'analyse canonique des tableaux \mathbf{A} et \mathbf{X} , où \mathbf{A} (de dimension $n \times K$) représente l'ensemble des variables indicatrices associées à la variable qualitative à expliquer.

3.2.3 Les anglo-saxons

En AFD, d'autres critères ont été utilisés pour définir les facteurs discriminants et donc les variables discriminantes. Les anglo-saxons utilisent souvent comme critère pour mesurer la qualité de la discrimination d'une variable discriminante $\mathbf{s} = \mathbf{XMa} = \mathbf{Xu}$:

$$\frac{\mathbf{u}'\mathbf{Bu}}{\mathbf{u}'\mathbf{Wu}} = \frac{Inter(\mathbf{s})}{Intra(\mathbf{s})}.$$

Le problème d'optimisation s'écrit :

$$\max_{\mathbf{u} \in \mathbb{R}^p} \frac{\mathbf{u}'\mathbf{Bu}}{\mathbf{u}'\mathbf{Wu}}. \quad (9)$$

La solution est le vecteur propre \mathbf{u} de $\mathbf{W}^{-1}\mathbf{B}$ associé à la plus grande valeur propre μ .

Ce problème d'optimisation est équivalent en terme de facteurs discriminants au problème (8) mais les maximums sont différents. En d'autres termes, les matrices $\mathbf{W}^{-1}\mathbf{B}$ et $\mathbf{V}^{-1}\mathbf{B}$ ont les mêmes vecteurs propres et $\mu = \frac{\lambda}{1-\lambda}$.

Preuve : On suppose \mathbf{W} inversible.

$$\mathbf{u} \text{ vecteur propre de } \mathbf{V}^{-1}\mathbf{B} \Rightarrow \mathbf{V}^{-1}\mathbf{Bu} = \lambda\mathbf{u}$$

$$\Rightarrow \mathbf{Bu} = \lambda\mathbf{Vu} = \lambda(\mathbf{B} + \mathbf{W})\mathbf{u}$$

$$\Rightarrow (1 - \lambda)\mathbf{Bu} = \lambda\mathbf{Wu}$$

$$\Rightarrow \mathbf{Bu} = \frac{\lambda}{1 - \lambda}\mathbf{Wu}$$

$$\Rightarrow \mathbf{W}^{-1}\mathbf{Bu} = \frac{\lambda}{1 - \lambda}\mathbf{u}$$

$$\Rightarrow \mathbf{u} \text{ vecteur propre de } \mathbf{W}^{-1}\mathbf{B} \text{ et } \frac{\lambda}{1 - \lambda} \text{ est la valeur propre } \mu$$

■

On retrouve donc les mêmes variables discriminantes qu'avec l'autre critère (au signe et à une constante près). En revanche le critère λ varie entre 0 et 1 et s'interprète facilement tandis que le critère μ varie entre zéro et l'infini et s'interprète donc "moins bien".

3.2.4 Le cas particulier de deux groupes

En pratique, on étudie souvent le cas particulier où $K = 2$ (bon client/mauvais client, sain/malade...). On ne recherche ici qu'une seule variable discriminante puisque $K - 1 = 1$. Nous allons voir qu'il est possible de définir analytiquement l'expression de l'unique valeur propre de $\mathbf{V}^{-1}\mathbf{B}$ et du vecteur propre associé.

Dans un premier temps, on montre que :

- a) $\mathbf{B} = \frac{n_1 n_2}{n^2}(\mathbf{g}_1 - \mathbf{g}_2)(\mathbf{g}_1 - \mathbf{g}_2)'$.
- b) $\mathbf{V}^{-1}(\mathbf{g}_1 - \mathbf{g}_2)$ est vecteur propre de $\mathbf{V}^{-1}\mathbf{B}$,
 $\lambda = \frac{n_1 n_2}{n^2}(\mathbf{g}_1 - \mathbf{g}_2)' \mathbf{V}^{-1}(\mathbf{g}_1 - \mathbf{g}_2)$ est la valeur propre de $\mathbf{W}^{-1}\mathbf{B}$.

Exercice : Montrez les résultats ci-dessus.

On en déduit qu'avec la métrique $\mathbf{M} = \mathbf{V}^{-1}$ et le critère $\frac{\mathbf{u}'\mathbf{B}\mathbf{u}}{\mathbf{u}'\mathbf{V}\mathbf{u}} = \frac{Intra(\mathbf{s})}{Var(\mathbf{s})}$ on a :

$$\left\{ \begin{array}{l} \mathbf{u} = \mathbf{V}^{-1}(\mathbf{g}_1 - \mathbf{g}_2) \text{ est le facteur discriminant,} \\ \mathbf{a} = (\mathbf{g}_1 - \mathbf{g}_2) \text{ est l'axe discriminant,} \\ \mathbf{s} = \mathbf{X}\mathbf{V}^{-1}(\mathbf{g}_1 - \mathbf{g}_2) \text{ est la variable discriminante,} \\ \lambda = \frac{n_1 n_2}{n^2} d_{\mathbf{V}^{-1}}^2(\mathbf{g}_1, \mathbf{g}_2) \text{ est le pouvoir discriminant } \frac{Intra(\mathbf{s})}{Var(\mathbf{s})}. \end{array} \right.$$

où $d_{\mathbf{V}^{-1}}^2(\mathbf{g}_1, \mathbf{g}_2) = (\mathbf{g}_1 - \mathbf{g}_2)' \mathbf{V}^{-1}(\mathbf{g}_1 - \mathbf{g}_2)$ est le carré de la distance entre \mathbf{g}_1 et \mathbf{g}_2 .

Remarque : Il existe une infinité de vecteurs propres colinéaires à \mathbf{u} . Le carré de la \mathbf{V} -norme de $\mathbf{u} = \mathbf{V}^{-1}(\mathbf{g}_1 - \mathbf{g}_2)$ est :

$$\|\mathbf{V}^{-1}(\mathbf{g}_1 - \mathbf{g}_2)\|_{\mathbf{V}}^2 = (\mathbf{g}_1 - \mathbf{g}_2)' \mathbf{V}^{-1}(\mathbf{g}_1 - \mathbf{g}_2) = d_{\mathbf{V}^{-1}}^2(\mathbf{g}_1, \mathbf{g}_2).$$

On en déduit qu'en divisant \mathbf{u} , \mathbf{a} et \mathbf{s} par $d_{\mathbf{V}^{-1}}(\mathbf{g}_1, \mathbf{g}_2)$, le facteur discriminant est \mathbf{V} -normé à 1 ($\mathbf{u}'\mathbf{V}\mathbf{u} = 1$), l'axe discriminant est \mathbf{V}^{-1} -normé à 1 ($\mathbf{a}'\mathbf{V}^{-1}\mathbf{a} = 1$), et la variance de la variable discriminante est égale à 1 ($Var(\mathbf{s}) = 1$).

De la même manière, avec la métrique est $\mathbf{M} = \mathbf{W}^{-1}$ et le critère $\frac{\mathbf{u}'\mathbf{B}\mathbf{u}}{\mathbf{u}'\mathbf{W}\mathbf{u}} = \frac{Intra(\mathbf{s})}{Inter(\mathbf{s})}$ on a :

$$\left\{ \begin{array}{l} \mathbf{u} = \mathbf{W}^{-1}(\mathbf{g}_1 - \mathbf{g}_2) \text{ est le facteur discriminant,} \\ \mathbf{a} = (\mathbf{g}_1 - \mathbf{g}_2) \text{ est l'axe discriminant,} \\ \mathbf{s} = \mathbf{s} = \mathbf{X}\mathbf{W}^{-1}(\mathbf{g}_1 - \mathbf{g}_2) \text{ est la variable discriminante,} \\ \mu = \frac{n_1 n_2}{n^2} d_{\mathbf{W}^{-1}}^2(\mathbf{g}_1, \mathbf{g}_2) \text{ est le pouvoir discriminant } \frac{Intra(\mathbf{s})}{Inter(\mathbf{s})}. \end{array} \right.$$

où $d_{\mathbf{W}^{-1}}^2(\mathbf{g}_1, \mathbf{g}_2) = (\mathbf{g}_1 - \mathbf{g}_2)' \mathbf{W}^{-1}(\mathbf{g}_1 - \mathbf{g}_2)$ est le carré de la distance entre \mathbf{g}_1 et \mathbf{g}_2 .

Remarque : Il faut diviser \mathbf{u} , \mathbf{a} et \mathbf{s} par $d_{\mathbf{W}^{-1}}(\mathbf{g}_1, \mathbf{g}_2)$ pour que le facteur discriminant soit \mathbf{W} -normé à 1, l'axe discriminant soit \mathbf{W}^{-1} -normé à 1 et la variance intra-groupe de la variable discriminante soit égale à 1 ($Intra(\mathbf{s}) = 1$).

Finalement, dans le cas particulier de deux groupes, tout se calcule analytiquement à partir des centres de gravité \mathbf{g}_1 et \mathbf{g}_2 et l'axe discriminant engendre la droite reliant \mathbf{g}_1 et \mathbf{g}_2 . Ceci n'a rien de surprenant vu que l'on a préconisé que l'axe discriminant doit être tel qu'en projection sur Δ , les K centres de gravité doivent être aussi séparés que possible. Le choix de la métrique $\mathbf{M} = \mathbf{V}^{-1}$ (ou $\mathbf{M} = \mathbf{W}^{-1}$) a une interprétation géométrique. En effet, on projette le nuage des individus sur la droite Δ (passant par \mathbf{g}_1 et \mathbf{g}_2) en tenant compte de l'orientation des nuages par rapport à cette droite.

La règle de Fisher (1936). La première variable discriminante peut être utilisée comme score pour construire une règle de décision simple en fixant un seuil c . On fixe comme seuil c le milieu des scores des deux centres de gravité \mathbf{g}_1 et \mathbf{g}_2 sur la première (et unique...) variable discriminante (les coordonnées de leur projection sur l'axe discriminant). En d'autres termes, le seuil c est le milieu des moyennes des 2 groupes sur la variable discriminante $\mathbf{s} = \mathbf{X}\mathbf{u}$. On note \bar{s}_1 et \bar{s}_2 ces deux moyennes et on utilise ici la métrique $\mathbf{M} = \mathbf{W}^{-1}$ (mais ce serait équivalent d'utiliser $\mathbf{M} = \mathbf{V}^{-1}$). On a :

$$\begin{aligned}\bar{s}_1 &= \frac{1}{n_1} \sum_{i \in E_1} s_i = \mathbf{g}_1' \mathbf{W}^{-1} \mathbf{a} && \text{projection } \mathbf{W}^{-1} \text{ orthogonale de } \mathbf{g}_1 \text{ sur } \Delta, \\ \bar{s}_2 &= \frac{1}{n_2} \sum_{i \in E_2} s_i = \mathbf{g}_2' \mathbf{W}^{-1} \mathbf{a} && \text{projection } \mathbf{W}^{-1} \text{ orthogonale de } \mathbf{g}_2 \text{ sur } \Delta,\end{aligned}$$

avec ici $\mathbf{a} = (\mathbf{g}_1 - \mathbf{g}_2)$. On prend alors comme seuil :

$$\begin{aligned}c &= \frac{\bar{s}_1 + \bar{s}_2}{2} \\ &= \frac{1}{2}(\mathbf{g}_1' \mathbf{W}^{-1} \mathbf{a} + \mathbf{g}_2' \mathbf{W}^{-1} \mathbf{a}) \\ &= \frac{1}{2}(\mathbf{g}_1 + \mathbf{g}_2)' \mathbf{W}^{-1} (\mathbf{g}_1 - \mathbf{g}_2)\end{aligned}$$

On peut alors définir la règle de décision :

$$\begin{aligned}s_i &\geq c \Rightarrow \text{l'individu } i \text{ est affecté au groupe 1,} \\ s_i &< c \Rightarrow \text{l'individu } i \text{ est affecté au groupe 2.}\end{aligned}$$

Cette règle se réécrit donc :

$$\begin{aligned} \mathbf{x}_i' \mathbf{W}^{-1}(\mathbf{g}_1 - \mathbf{g}_2) - \frac{1}{2}(\mathbf{g}_1 + \mathbf{g}_2)' \mathbf{W}^{-1}(\mathbf{g}_1 - \mathbf{g}_2) &\geq 0 \Rightarrow \text{l'individu } i \text{ est affecté au groupe 1,} \\ \mathbf{x}_i' \mathbf{W}^{-1}(\mathbf{g}_1 - \mathbf{g}_2) - \frac{1}{2}(\mathbf{g}_1 + \mathbf{g}_2)' \mathbf{W}^{-1}(\mathbf{g}_1 - \mathbf{g}_2) &< 0 \Rightarrow \text{l'individu } i \text{ est affecté au groupe 2.} \end{aligned}$$

On retrouve la règle de Mahalanobis-Fisher dans le cas de deux groupes.

4 Analyse discriminante géométrique

On va présenter ici une approche purement géométrique (sans aucune hypothèse probabiliste) de l'analyse discriminante décisionnelle. Ayant trouvé la meilleure représentation de la séparation en K groupes des n individus, on va chercher ici à affecter une nouvelle observation à l'un de ces groupes. Notons \mathbf{x} le vecteur des valeurs des p variables quantitatives mesurées sur ce nouvel individu. La règle géométrique consiste à calculer les distances de cette observation à chacun des K centres de gravité et à affecter naturellement cette observation au groupe le plus proche. Pour cela, il faut préciser la métrique à utiliser dans le calcul des distances. La règle la plus utilisée est celle de Mahalanobis-Fisher qui consiste à prendre la métrique \mathbf{W}^{-1} (ou \mathbf{V}^{-1} ce qui est équivalent). La distance du nouvel individu au groupe k est alors :

$$d^2(\mathbf{x}, \mathbf{g}_k) = (\mathbf{x} - \mathbf{g}_k)' \mathbf{W}^{-1}(\mathbf{x} - \mathbf{g}_k). \quad (10)$$

4.1 Fonctions linéaires discriminantes

La règle géométrique classe la nouvelle observation \mathbf{x} dans le groupe k^* tel que :

$$k^* = \arg \min_{k=1, \dots, K} d^2(\mathbf{x}, \mathbf{g}_k), \quad (11)$$

ce qui se réécrit :

$$k^* = \arg \max_{k=1, \dots, K} L_k(\mathbf{x}), \quad (12)$$

où

$$L_k(\mathbf{x}) = \mathbf{x}' \mathbf{W}^{-1} \mathbf{g}_k - \frac{1}{2} \mathbf{g}_k' \mathbf{W}^{-1} \mathbf{g}_k. \quad (13)$$

$L_k(\mathbf{x})$ est la fonction linéaire discriminante du groupe k (encore appelée fonction linéaire de classement).

Preuve : \mathbf{W}^{-1} étant symétrique on a :

$$\begin{aligned} d^2(\mathbf{x}, \mathbf{g}_k) &= (\mathbf{x} - \mathbf{g}_k)' \mathbf{W}^{-1} (\mathbf{x} - \mathbf{g}_k) \\ &= \underbrace{\mathbf{x}' \mathbf{W}^{-1} \mathbf{x}}_{\text{indépendant de } k} + \mathbf{g}_k' \mathbf{W}^{-1} \mathbf{g}_k - 2\mathbf{x}' \mathbf{W}^{-1} \mathbf{g}_k \end{aligned}$$

Donc minimiser d^2 est équivalent à maximiser $\frac{2\mathbf{x}' \mathbf{W}^{-1} \mathbf{g}_k - \mathbf{g}_k' \mathbf{W}^{-1} \mathbf{g}_k}{2} = L_k(\mathbf{x})$.

■

Chaque fonction linéaire discriminante définit une fonction score qui donne une “note” à l’observation \mathbf{x} dans chaque groupe. Cette observation est donc affectée au groupe pour lequel le score est le plus grand.

4.2 Le cas particulier de deux groupes

On peut dans ce cas particulier où $K = 2$, calculer la différence entre les fonctions linéaires discriminantes du groupe 1 et du groupe 2 pour définir une nouvelle fonction score que l’on comparera à zéro pour affecter une nouvelle observation. On définit ainsi la nouvelle fonction linéaire discriminante (fonction score) :

$$\begin{aligned} \Delta_{1/2}(\mathbf{x}) &= L_1(\mathbf{x}) - L_2(\mathbf{x}) \\ &= \mathbf{x}' \mathbf{W}^{-1} (\mathbf{g}_1 - \mathbf{g}_2) - \frac{1}{2} (\mathbf{g}_1 + \mathbf{g}_2)' \mathbf{W}^{-1} (\mathbf{g}_1 - \mathbf{g}_2) \end{aligned} \quad (14)$$

Preuve :

$$\begin{aligned} L_1(\mathbf{x}) - L_2(\mathbf{x}) &= \mathbf{x}' \mathbf{W}^{-1} \mathbf{g}_1 - \frac{1}{2} \mathbf{g}_1' \mathbf{W}^{-1} \mathbf{g}_1 - \left(\mathbf{x}' \mathbf{W}^{-1} \mathbf{g}_2 - \frac{1}{2} \mathbf{g}_2' \mathbf{W}^{-1} \mathbf{g}_2 \right) \\ &= \mathbf{x}' \mathbf{W}^{-1} (\mathbf{g}_1 - \mathbf{g}_2) - \frac{1}{2} \left(\underbrace{\mathbf{g}_1' \mathbf{W}^{-1} \mathbf{g}_1 - \mathbf{g}_2' \mathbf{W}^{-1} \mathbf{g}_2}_{(\mathbf{g}_1 + \mathbf{g}_2)' \mathbf{W}^{-1} (\mathbf{g}_1 - \mathbf{g}_2)} \right) \end{aligned}$$

■

On retrouve alors la règle de Mahalanobis-Fisher introduite en AFD :

$$\begin{aligned} \mathbf{x}' \mathbf{W}^{-1} (\mathbf{g}_1 - \mathbf{g}_2) - \frac{1}{2} (\mathbf{g}_1 + \mathbf{g}_2)' \mathbf{W}^{-1} (\mathbf{g}_1 - \mathbf{g}_2) &\geq 0 \Rightarrow \text{l'individu } i \text{ est affecté au groupe 1,} \\ \mathbf{x}' \mathbf{W}^{-1} (\mathbf{g}_1 - \mathbf{g}_2) - \frac{1}{2} (\mathbf{g}_1 + \mathbf{g}_2)' \mathbf{W}^{-1} (\mathbf{g}_1 - \mathbf{g}_2) &< 0 \Rightarrow \text{l'individu } i \text{ est affecté au groupe 2.} \end{aligned}$$

Dans le cas particulier de 2 groupes, la règle géométrique revient donc à projeter l'observation \mathbf{x} avec la métrique \mathbf{W}^{-1} sur le premier axe discriminant (calcul du score $\mathbf{x}'\mathbf{W}^{-1}(\mathbf{g}_1 - \mathbf{g}_2)$) et à classer \mathbf{x} selon un seuil c qui est le milieu des moyennes des groupes sur ce score ($c = \frac{1}{2}(\mathbf{g}_1 + \mathbf{g}_2)'\mathbf{W}^{-1}(\mathbf{g}_1 - \mathbf{g}_2)$).

De plus, $\Delta_{1/2}(\mathbf{x})$ s'écrit :

$$\Delta_{1/2}(\mathbf{x}) = (\mathbf{x} - \frac{\mathbf{g}_1 + \mathbf{g}_2}{2})'\mathbf{W}^{-1}(\mathbf{g}_1 - \mathbf{g}_2). \quad (15)$$

Donc la frontière entre les groupes 1 et 2 (aussi appelée frontière de décision) est l'hyperplan d'équation :

$$\langle \mathbf{x} - \frac{\mathbf{g}_1 + \mathbf{g}_2}{2}, \mathbf{g}_1 - \mathbf{g}_2 \rangle_{W^{-1}} = 0.$$

L'hyperplan séparateur est donc l'hyperplan W^{-1} -orthogonal à la droite reliant les points \mathbf{g}_1 et \mathbf{g}_2 et passant par le milieu de \mathbf{g}_1 et \mathbf{g}_2 .

4.3 Insuffisances de la règle géométrique

- L'utilisation de la règle d'affectation géométrique que l'on vient de présenter peut conduire à des affectations incorrectes lorsque les dispersions des groupes sont très différentes entre elles. En fait dans ce cadre-là, rien ne justifie l'utilisation de la même métrique pour les différents groupes.


Exemple.

La solution pour pallier ce problème peut être d'utiliser des métriques locales M_k à chaque sous-nuage E_k , proportionnelles à \mathbf{V}_k^{-1} et de calculer les distances $d_k^2(\mathbf{x}, g_l) = (\mathbf{x} - g_k)'\mathbf{M}_k(\mathbf{x} - \mathbf{g}_k)$.


- La question de l'optimalité de la règle géométrique d'affectation ne peut être résolue sans référence à un modèle probabiliste. En effet, une question est de savoir comment se comportera cette règle pour de nouvelles observations, ce qui va imposer de faire des hypothèses distributionnelles sur la répartition des individus dans l'espace.

Cette première approche de l'analyse discriminante décisionnelle atteint donc ici clairement ses limites, d'où la nécessité de passer à la section suivante dans laquelle seront présentées des méthodes probabilistes d'affectation qui conduiront à des règles optimales (en un certain sens).

5 Analyse discriminante probabiliste

 Dans la section précédente, on disposait d'un échantillon de taille n sur lequel étaient mesurées simultanément les p variables explicatives quantitatives et la variable qualitative à expliquer. Généralement cet échantillon est appelé échantillon d'apprentissage. On a vu que, lorsque l'on dispose d'un nouvel individu sur lequel on a mesuré uniquement les p variables explicatives (et dont on ne connaît pas sa valeur de la variable qualitative), on doit décider du groupe k auquel appartient ce nouvel individu. On a alors défini une règle de décision (d'affectation) fondée seulement sur des arguments géométriques (calcul des distances de cet individu aux centres de gravité de chaque groupe et affectation au groupe "le plus proche"). Dans cette section, on va définir des règles de décision bayésienne qui vont permettre d'affecter le nouvel individu à la classe "la plus probable". Pour cela, il va être nécessaire de faire des hypothèses probabilistes sur les données, d'où l'appellation de méthodes probabilistes (par opposition aux méthodes purement géométriques).


5.1 Le cadre probabiliste

 On suppose maintenant que l'échantillon d'apprentissage est issu d'une population en K groupes G_1, \dots, G_K et que :

- Y est une variable aléatoire qui prend ses valeurs dans $\{1, \dots, K\}$.
- $X = (X_1, \dots, X_p)$ est un vecteur de variables aléatoires réelles.

On notera :

- (p_1, \dots, p_K) la distribution de Y où $p_k = P(Y = k)$ est la proportion théorique de G_k encore appelée **probabilité à priori** de G_k .
- $f_k : \mathbb{R}^p \rightarrow [0, 1]$ la densité de X dans le groupe k .

 Exemple : 2 groupes et $p=1$

— la densité de X est une densité de mélange :

$$X \sim \sum_{k=1}^K p_k f_k(\mathbf{x})$$

On supposera que l'on dispose d'un échantillon i.i.d. $(X_1, Y_1), \dots, (X_n, Y_n)$ de même loi que (X, Y) .

5.2 Règle du maximum à posteriori

La règle du MAP (maximum à posteriori) affecte une nouvelle observation \mathbf{x} dans le groupe k^* le plus probable sachant \mathbf{x} :

$$k^* = \arg \max_{k=1, \dots, K} P(G_k | \mathbf{x}), \quad (16)$$

où $P(G_k | \mathbf{x}) = P(Y = k | X = \mathbf{x})$ est la probabilité conditionnelle appelée la **probabilité à posteriori** de G_k . Les probabilités à posteriori $P(G_k | \mathbf{x})$ sont parfois qualifiées de scores (notes) et on affecte donc une nouvelle observation au groupe pour lequel le score est le plus grand.

Cette règle se réécrit :

$$k^* = \arg \max_{k=1, \dots, K} p_k f_k(\mathbf{x}). \quad (17)$$

On parle alors souvent de règle de Bayes.

Preuve : On utilise le théorème de Bayes qui donne :

$$P(G_k | \mathbf{x}) = P(Y = k | X = \mathbf{x}) = \frac{f_k(\mathbf{x})P(Y = k)}{f_X(\mathbf{X})} = \frac{f_k(\mathbf{x})p_k}{f_X(\mathbf{X})}$$

Or $f_X(\mathbf{x})$ est indépendante de k donc il suffit de maximiser $f_k(\mathbf{x})p_k$.

■

Plusieurs approches sont possibles.

a) Les approches paramétriques :

- On peut supposer que $f_k(\mathbf{x})$ a une forme paramétrique et estimer les paramètres sur l'échantillon d'apprentissage. Par exemple, $f_k(\mathbf{x})$ est une densité $\mathcal{N}(\mu_k, \Sigma_k)$ pour les méthodes LDA et QDA.
- On peut supposer que la probabilité à posteriori $P(G_k | \mathbf{x})$ a une expression paramétrique et l'estimer directement. Par exemple $P(G_1 | \mathbf{x}) = \frac{\exp(\beta_0 + \beta' \mathbf{x})}{1 + \exp(\beta_0 + \beta' \mathbf{x})}$ pour la régression logistique (pour $K = 2$).

b) Les approches non paramétriques : on cherche à estimer directement à partir des données les densités f_k . On parle d'estimation non paramétrique (ou estimation fonctionnelle) lorsque le nombre de paramètres à estimer est infini. L'objet à estimer est alors une fonction que l'on supposera par exemple continue et dérivable. Cette approche très souple a donc l'avantage de ne pas nécessiter d'hypothèses particulières sur la densité f_k (seulement la régularité de f_k pour avoir de bonnes propriétés de convergence). En revanche, elle n'est applicable d'un point de vue pratique qu'avec des échantillons de grande taille d'autant plus que la dimension p augmente. Exemple : méthodes à noyaux.

Ici on se place dans le cadre paramétrique gaussien.

5.3 Le cas gaussien

On suppose maintenant que $X \sim \mathcal{N}(\mu_k, \Sigma_k)$ dans chaque groupe G_k :

$$f_k(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}(\det(\Sigma_k))^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x} - \mu_k)' \Sigma_k^{-1}(\mathbf{x} - \mu_k)\right), \quad (18)$$

avec $\mu_k \in \mathbb{R}^p$ le vecteur des moyennes théoriques et Σ_k la matrice $p \times p$ des variances-covariances théoriques.

Dans ce cas, la règle de Bayes se réécrit :

$$k^* = \arg \min_{k=1, \dots, K} D_k^2(\mathbf{x}), \quad (19)$$

où

$$D_k^2(\mathbf{x}) = (\mathbf{x} - \mu_k)' \Sigma_k^{-1}(\mathbf{x} - \mu_k) - 2 \ln(p_k) + \ln(\det \Sigma_k), \quad (20)$$

est appelé le carré de la distance de Mahalanobis théorique **généralisée dans SAS.** 

Preuve : Maximiser $p_k f_k(\mathbf{x})$ est équivalent à maximiser $\ln(p_k f_k(\mathbf{x}))$ et

$$\begin{aligned} \ln(p_k f_k(\mathbf{x})) &= \ln(p_k) + \ln(f_k(\mathbf{x})), \\ &= \ln(p_k) - \frac{p}{2} \ln(2\pi) - \frac{1}{2} \ln(\det(\Sigma_k)) - \frac{1}{2}(\mathbf{x} - \mu_k)' \Sigma_k^{-1}(\mathbf{x} - \mu_k). \end{aligned}$$

Donc avec $\frac{p}{2} \ln(2\pi)$ qui est indépendant de k , maximiser $\ln(p_k f_k(\mathbf{x}))$ est équivalent à minimiser $-2 \left(\ln(p_k) - \frac{1}{2} \ln(\det(\Sigma_k)) - \frac{1}{2}(\mathbf{x} - \mu_k)' \Sigma_k^{-1}(\mathbf{x} - \mu_k) \right) = D_k(\mathbf{x})^2$.

■

5.3.1 Estimation des paramètres

A partir de l'échantillon d'apprentissage, on veut estimer le paramètre

$$\theta = (p_1, \dots, p_K, \mu_1, \dots, \mu_K, \Sigma_1, \dots, \Sigma_K).$$

La méthode du maximum de vraisemblance peut être utilisée. La vraisemblance s'écrit :

$$L(\theta) = \prod_{i=1}^n f_X(\mathbf{x}_i) = \prod_{k=1}^K \prod_{\mathbf{x}_i \in E_k} p_k f_k(\mathbf{x}_i),$$

et on en déduit que la log-vraisemblance s'écrit :

$$\ln(L(\theta)) = \sum_{k=1}^K \sum_{\mathbf{x}_i \in E_k} \left(\ln(p_k) - \frac{p}{2} \ln(2\pi) - \frac{1}{2} \ln(\det(\Sigma_k)) - \frac{1}{2} (\mathbf{x}_i - \mu_k)' \Sigma_k^{-1} (\mathbf{x}_i - \mu_k) \right).$$

On obtient alors les estimateurs du maximum de vraisemblance suivant :

$$\begin{aligned} \hat{p}_k &= \frac{n_k}{n} \\ \hat{\mu}_k &= \frac{1}{n_k} \sum_{i \in E_k} \mathbf{x}_i \\ \hat{\Sigma}_k &= \begin{cases} \hat{\Sigma} = \frac{1}{n} \sum_{k=1}^K \sum_{i \in E_k} (\mathbf{x}_i - \mu_k)(\mathbf{x}_i - \mu_k)' & \text{dans le cas homoscédastique,} \\ \hat{\Sigma}_k = \frac{1}{n_k} \sum_{i \in E_k} (\mathbf{x}_i - \mu_k)(\mathbf{x}_i - \mu_k)' & \text{dans le cas hétéroscédastique.} \end{cases} \end{aligned}$$

Les estimateurs de $\hat{\Sigma}_k$ étant biaisé, on a les estimateurs sans biais suivants :

$$\begin{aligned} \hat{\Sigma} &= \frac{1}{n - K} \sum_{k=1}^K \sum_{i \in E_k} (\mathbf{x}_i - \mu_k)(\mathbf{x}_i - \mu_k)', \\ \hat{\Sigma}_k &= \frac{1}{n_k - 1} \sum_{i \in E_k} (\mathbf{x}_i - \mu_k)(\mathbf{x}_i - \mu_k)'. \end{aligned}$$

5.3.2 Analyse discriminante quadratique (QDA en anglais)

On se place dans le cas où $\exists k \neq k^*$ tel que $\Sigma_k \neq \Sigma_{k^*}$ appelé cas hétéroscédastique. On estime alors les paramètres sur l'échantillon d'apprentissage et en reprenant les notations de la section 2 :

- μ_k est estimée par $g_k = \frac{1}{n_k} \sum_{i \in E_k} \mathbf{x}_i$,
- Σ_k est estimée par $\mathbf{V}_k = \frac{1}{n_k} \sum_{i \in E_k} (\mathbf{x}_i - \mathbf{g}_k)(\mathbf{x}_i - \mathbf{g}_k)'$ ou encore par sa version sans biais :

$$\mathbf{V}_k = \frac{1}{n_k - 1} \sum_{i \in E_k} (\mathbf{x}_i - \mathbf{g}_k)(\mathbf{x}_i - \mathbf{g}_k)',$$

- p_k est estimée par $\pi_k = \frac{n_k}{n}$.

Avec ces estimateurs, la règle de Bayes se réécrit :

$$k^* = \arg \min_{k=1,\dots,K} Q_k(\mathbf{x}), \quad (21)$$

où

$$Q_k(\mathbf{x}) = (\mathbf{x} - \mathbf{g}_k)' \mathbf{V}_k^{-1} (\mathbf{x} - \mathbf{g}_k) - 2 \ln(\pi_k) + \ln(\det(\mathbf{V}_k)), \quad (22)$$

est la fonction quadratique discriminante du groupe k (encore appelée fonction quadratique de classement). Chaque fonction quadratique discriminante définit une fonction score et une nouvelle observation sera affectée au groupe pour lequel le score sera le plus **petit**.

5.3.3 Analyse discriminante linéaire (LDA en anglais)

On se place dans le cas où $\Sigma_1 = \dots = \Sigma_K = \Sigma$ appelé cas homoscedastique. Dans ce cas, la règle de Bayes se réécrit :

$$k^* = \arg \max_{k=1,\dots,K} \mathbf{x}' \Sigma^{-1} \mu_k - \frac{1}{2} \mu_k' \Sigma^{-1} \mu_k + \ln(p_k). \quad (23)$$

Preuve : $\Sigma_k = \Sigma$ pour tout $k = 1, \dots, K$ donc

$$\begin{aligned} D_k^2(\mathbf{x}) &= (\mathbf{x} - \mu_k)' \Sigma^{-1} (\mathbf{x} - \mu_k) - 2 \ln(p_k) + \ln(\det \Sigma), \\ &= \underbrace{\mathbf{x}' \Sigma^{-1} \mathbf{x}}_{\text{indép. de } k} - \underbrace{2 \mathbf{x}' \Sigma^{-1} \mu_k}_{\text{car } \Sigma^{-1} \text{ sym.}} + \mu_k' \Sigma^{-1} \mu_k - 2 \ln(p_k) + \underbrace{\ln(\det(\Sigma))}_{\text{indép. de } k}. \end{aligned}$$

Donc minimiser $D_k^2(\mathbf{x})$ est équivalent à maximiser $-\frac{1}{2} (2 \mathbf{x}' \Sigma^{-1} \mu_k + \mu_k' \Sigma^{-1} \mu_k - 2 \ln(p_k))$. ■

On estime alors les paramètres sur l'échantillon d'apprentissage et en reprenant les notations de la section 2 :

- μ_k est estimée par $g_k = \frac{1}{n_k} \sum_{i \in E_k} \mathbf{x}_i$,
- la matrice de variances-covariances Σ commune aux différents groupes est estimée par $\mathbf{W} = \frac{1}{n} \sum_{k=1}^K \sum_{i \in E_k} (\mathbf{x}_i - \mathbf{g}_k)(\mathbf{x}_i - \mathbf{g}_k)'$ ou encore par la version sans biais :

$$\mathbf{W} = \frac{1}{n - K} \sum_{k=1}^K \sum_{i \in E_k} (\mathbf{x}_i - \mathbf{g}_k)(\mathbf{x}_i - \mathbf{g}_k)',$$

- p_k est estimée par $\pi_k = \frac{n_k}{n}$.

Avec ces estimateurs, la règle de Bayes se réécrit :

$$k^* = \arg \max_{k=1,\dots,K} L_k(\mathbf{x}), \quad (24)$$

où

$$L_k(\mathbf{x}) = \mathbf{x}'\mathbf{W}^{-1}\mathbf{g}_k - \frac{1}{2}\mathbf{g}_k'\mathbf{W}^{-1}\mathbf{g}_k + \ln(\pi_k), \quad (25)$$

est la fonction linéaire discriminante du groupe k (encore appelée fonction linéaire de classement). Chaque fonction linéaire discriminante définit une fonction score et une nouvelle observation sera affectée au groupe pour lequel le score sera le plus **grand**.

Remarque : On retrouve la fonction linéaire discriminante (13) de l'analyse discriminante géométrique avec le terme $\ln(\pi_k)$ en plus. Dans le cas où l'on fait l'hypothèse d'égalité des probabilités à priori ($p_1 = \dots = p_K$), la règle de l'analyse discriminante linéaire (LDA) est équivalente à la règle de l'analyse discriminante géométrique.

5.3.4 Estimation des probabilités à posteriori



On peut montrer que dans le cas gaussien, on a :

$$P(G_k | \mathbf{x}) = \frac{\exp(-\frac{1}{2}D_k^2(\mathbf{x}))}{\sum_{l=1}^K \exp(-\frac{1}{2}D_l^2(\mathbf{x}))} \quad (26)$$

où $D_k^2(\mathbf{x})$ est le carré de la distance de Mahalanobis théorique généralisée du logiciel SAS définie en (20).

Preuve : D'après le théorème de Bayes :

$$P(G_k | \mathbf{x}) = \frac{p_k f_k(\mathbf{x})}{\sum_{l=1}^K p_l f_l(\mathbf{x})}, \quad (27)$$

et on a $\ln(p_k f_k(\mathbf{x})) = -\frac{1}{2}D_k^2(\mathbf{x}) - \frac{p}{2}\ln(2\pi)$, d'où :

$$\begin{aligned} p_k f_k(\mathbf{x}) &= \exp(-\frac{1}{2}D_k^2 - \frac{p}{2}\ln(2\pi)) \\ &= \exp(-\frac{1}{2}D_k^2) \times (2\pi)^{\frac{p}{2}}. \end{aligned}$$

Donc en mettant $(2\pi)^{\frac{p}{2}}$ en facteur au dénominateur de (27), le terme se simplifie au numérateur et on retrouve (26).



En pratique, on estime les probabilités à posteriori par :

$$\hat{P}(G_k | \mathbf{x}) = \frac{\exp(-\frac{1}{2}\hat{D}_k^2(\mathbf{x}))}{\sum_{l=1}^K \exp(-\frac{1}{2}\hat{D}_l^2(\mathbf{x}))}, \quad (28)$$

où

$$\hat{D}_k^2(\mathbf{x}) = (\mathbf{x} - \hat{\mu}_k)'\hat{\Sigma}_k^{-1}(\mathbf{x} - \hat{\mu}_k) + g_1(\mathbf{x}) + g_2(\mathbf{x}), \quad (29)$$

et

$$g_1(\mathbf{x}) = \begin{cases} \ln(\hat{\Sigma}_k) & \text{dans le cas hétéroscédastique (quadratique),} \\ 0 & \text{dans le cas homoscdastique (linéaire).} \end{cases}$$

$$g_2(\mathbf{x}) = \begin{cases} -2 \ln(\hat{p}_k) & \text{si toutes les probabilités à priori ne sont pas égales,} \\ 0 & \text{si elles sont toutes égales (equiprobabilité).} \end{cases}$$

Preuve :

Dans le cas homoscdastique, $\hat{\Sigma}_k = \hat{\Sigma}$ est indépendant de k , donc $\exp(\ln(\det(\hat{\Sigma}))) = \det(\hat{\Sigma})$ se met en facteur au dénominateur et s'annule avec le numérateur dans (27). ■

On prendra $\hat{\mu}_k = \mathbf{g}_k$, $\hat{p}_k = \frac{n_k}{n}$ et $\hat{\Sigma}_k = \begin{cases} \mathbf{V}_k & \text{dans le cas hétéroscédastique (quadratique),} \\ \mathbf{W} & \text{dans le cas homoscdastique (linéaire).} \end{cases}$



Remarques. Dans le logiciel SAS :

- la matrice \mathbf{W} (sa version corrigée) est appelée “pooled covariance matrix”,
- $-\frac{1}{2}\hat{D}_k^2(\mathbf{x})$ est appelé “score discriminant”.

5.3.5 Cas particulier de deux groupes

On se place dans le cadre linéaire (homoscdastique) et dans le cas où $K = 2$. On définit alors la nouvelle fonction linéaire discriminante (fonction score) :

$$\begin{aligned} \Delta_{1/2}(\mathbf{x}) &= L_1(\mathbf{x}) - L_2(\mathbf{x}) \\ &= \mathbf{x}'\mathbf{W}^{-1}(\mathbf{g}_1 - \mathbf{g}_2) - \frac{1}{2}(\mathbf{g}_1 + \mathbf{g}_2)'\mathbf{W}^{-1}(\mathbf{g}_1 - \mathbf{g}_2) + \ln\left(\frac{\pi_1}{\pi_2}\right). \end{aligned} \quad (30)$$

que l'on compare à 0 pour affecter un nouvel individu à l'un des deux groupes. La règle de Bayes (version estimée) se réécrit alors :

$$\begin{aligned} \mathbf{x}'\mathbf{W}^{-1}(\mathbf{g}_1 - \mathbf{g}_2) - \frac{1}{2}(\mathbf{g}_1 + \mathbf{g}_2)'\mathbf{W}^{-1}(\mathbf{g}_1 - \mathbf{g}_2) + \ln\left(\frac{\pi_1}{\pi_2}\right) &\geq 0 \Rightarrow \text{l'individu } i \text{ est affecté au groupe 1,} \\ \mathbf{x}'\mathbf{W}^{-1}(\mathbf{g}_1 - \mathbf{g}_2) - \frac{1}{2}(\mathbf{g}_1 + \mathbf{g}_2)'\mathbf{W}^{-1}(\mathbf{g}_1 - \mathbf{g}_2) + \ln\left(\frac{\pi_1}{\pi_2}\right) &< 0 \Rightarrow \text{l'individu } i \text{ est affecté au groupe 2.} \end{aligned}$$

Remarques. Dans le cas particulier de deux groupes avec hypothèse d'égalité des matrices de variances-covariances des groupes :

- La règle de Bayes revient à projeter l'observation \mathbf{x} avec la métrique \mathbf{W}^{-1} sur le premier axe discriminant (calcul du score $\mathbf{x}'\mathbf{W}^{-1}(\mathbf{g}_1 - \mathbf{g}_2)$) et à classer \mathbf{x} selon un seuil c qui est le milieu des moyennes des groupes sur ce score - $\ln(\frac{\pi_1}{\pi_2})$. Cela veut dire

qu'on décale le seuil vers la moyenne du groupe 2 si $\pi_1 > \pi_2$ et vers la moyenne du groupe 1 dans le cas contraire.

Exemple :

- On peut montrer que

$$\hat{P}(G_1 | \mathbf{x}) = \frac{\exp(\Delta_{1/2}(\mathbf{x}))}{1 + \exp(\Delta_{1/2}(\mathbf{x}))} \quad (31)$$

On dit que $\hat{P}(G_1 | \mathbf{x})$ est une fonction logistique du score $\Delta_{1/2}(\mathbf{x})$ qui s'écrit aussi :

$$\Delta_{1/2} = \hat{\beta}_0 + \hat{\beta}'\mathbf{x}, \quad (32)$$

où

$$\hat{\beta}_0 = -\frac{1}{2}(\mathbf{g}_1 + \mathbf{g}_2)'\mathbf{W}^{-1}(\mathbf{g}_1 - \mathbf{g}_2) + \ln\left(\frac{\pi_1}{\pi_2}\right) \in \mathbb{R}, \quad (33)$$

et

$$\hat{\beta} = (\hat{\beta}_1, \dots, \hat{\beta}_p)' = \mathbf{W}^{-1}(\mathbf{g}_1 - \mathbf{g}_2) \in \mathbb{R}^p. \quad (34)$$

5.4 Régression logistique

On suppose que l'échantillon d'apprentissage est issu d'une population en 2 groupes G_1 et G_2 et que :

- Y est une variable aléatoire binaire ($Y = 1$ ou $Y = 0$).
- $X = (X_1, \dots, X_p)$ est un vecteur de variables réelles supposées cette fois non aléatoires.

Le modèle de régression logistique est alors :

$$Y = f(\beta_0 + \beta'X) + \epsilon, \quad (35)$$

et f est la fonction logistique $f(x) = \frac{\exp(x)}{1 + \exp(x)}$ de \mathbb{R} dans $[0, 1]$.

En supposant que le terme d'erreur ϵ est centré, on aura $E(\epsilon) = 0$ et donc

$$E(Y | X) = \frac{\exp(\beta_0 + \beta'X)}{1 + \exp(\beta_0 + \beta'X)}.$$

Et comme

$$E(Y | X) = 1 \times P(Y = 1 | X = \mathbf{x}) + 0 \times P(Y = 0 | X = \mathbf{x}) = P(G_1 | \mathbf{x}),$$

on modélise la probabilité à postériori par :

$$P(G_1 | \mathbf{x}) = \frac{\exp(\beta_0 + \beta'X)}{1 + \exp(\beta_0 + \beta'X)}. \quad (36)$$

Les paramètres β_0 et β sont estimés par maximum de vraisemblance sur un échantillon i.i.d. pour obtenir $\hat{\beta}_0$ et $\hat{\beta}$ et donc $\hat{P}(G_1 | \mathbf{x}) = \frac{\exp(\hat{\beta}_0 + \hat{\beta}'X)}{1 + \exp(\hat{\beta}_0 + \hat{\beta}'X)}$.

Les avantages : Les variables explicatives peuvent être qualitatives. On considère alors les variables indicatrices des modalités. De plus, les coefficients s'interprètent comme des logarithmes népériens d'odds ratio.

En régression logistique, on utilisera généralement la règle de classement :

$$\begin{aligned} \hat{P}(G_1 | \mathbf{x}) \geq 0.5 &\Rightarrow \mathbf{x} \text{ est affecté au groupe 1,} \\ \hat{P}(G_1 | \mathbf{x}) < 0.5 &\Rightarrow \mathbf{x} \text{ est affecté au groupe 2.} \end{aligned}$$

6 Mesures d'efficacité

6.1 Qualité d'une règle de décision

On suppose maintenant que l'on a un prédicteur (classifieur) c'est à dire une règle de décision de type régression logistique, analyse discriminante, classification nonparamétrique.... On mesure souvent l'efficacité d'une règle de décision en terme de **taux d'erreurs de classement**. On va décrire ici les trois méthodes courantes d'estimation de ce taux d'erreur.

6.1.1 La méthode de resubstitution

Cette méthode consiste à appliquer la règle d'affectation choisie sur l'échantillon d'apprentissage. On calcule ensuite le taux de mal-classés à partir de la matrice de confusion. Ce taux s'appelle le *taux apparent d'erreurs*.

Cette méthode présente l'avantage d'être peu coûteuse en temps de calcul. Mais, elle possède un grave défaut : elle sous-estime souvent (voire systématiquement) le taux d'erreurs vu que l'on utilise les mêmes observations pour le calculer que celles qui ont servi à trouver les fonctions discriminantes.

Remarque. Le taux d'erreur est d'autant plus faible que le modèle est complexe (sur-paramétrisation) et que la taille de l'échantillon est faible. Cette méthode est donc peu recommandée.

6.1.2 La méthode de l'échantillon-test

Cette méthode consiste à partager l'échantillon en deux parties :

- une partie (de l'ordre de 80%) sert d'échantillon d'apprentissage de la règle de décision,
- l'autre partie sert d'échantillon-test et permet de tester la règle d'affectation et donc de calculer le taux d'erreurs.

Cette méthode est plus fiable mais nécessite un échantillon plus important.

Remarque En effectuant plusieurs tirages aléatoires d'échantillons d'apprentissage (et donc d'échantillons-test), on peut améliorer l'estimation du taux d'erreurs en calculant la moyenne (et l'écart-type) des valeurs des taux d'erreurs obtenues à chaque tirage.

6.1.3 La méthode de validation croisée

Cette méthode est la plus lourde en terme de temps de calcul mais convient mieux aux petits échantillons. Pour tout $i = 1, \dots, n$, on va considérer les n échantillons d'apprentissage constitués en éliminant la i ème observation. La règle de décision qui en découle est utilisée pour affecter cette i ème observation. A l'issue de ces n analyses discriminantes, le taux d'erreurs est estimé en divisant le nombre de mal-classés par n . On parle de méthode LOO (Leave One Out).

Pour des échantillons plus important, on peut également utiliser la méthode de validation croisée K -fold qui consiste à découper aléatoirement le jeu de données en K sous-échantillons de même taille. Pour tout $k = 1, \dots, K$, on va considérer comme échantillon d'apprentissage les $K - 1$ sous-échantillons constitués en éliminant le k ème. La règle de décision qui en découle est utilisée pour affecter les individus du k ème sous-échantillon. A l'issue de ces K analyses discriminantes, le taux d'erreurs est estimé en divisant le nombre de mal-classés par n .

La encore, en effectuant plusieurs tirages aléatoires de découpages en K échantillons aléatoires, on peut améliorer l'estimation du taux d'erreurs en calculant la moyenne (et l'écart-type)des valeurs des taux d'erreurs obtenues à chaque tirage.

6.1.4 Indicateurs naifs

Il semble raisonnable de comparer les taux d'erreurs obtenu avec un prédicteur au taux d'erreurs de classement obtenu avec une règle naive de décision :

- Le critère MCC (maximum chance criterion) est le taux d'erreur que l'on obtient lorsque la règle de décision consiste à affecter une observation au groupe la plus probable à priori. En pratique, on estime ces probabilités par les fréquences des groupes et on affecte donc toutes observations au même groupe. Dans le cas de deux groupes, le taux apparent d'erreur est $1 - \frac{n_{k^*}}{n}$ où k^* est la classe la plus fréquente.
- Le critère PCC (proportional chance criterion) est le taux d'erreur que l'on obtient lorsque la règle consiste à affecter aléatoirement une observation à l'un des groupes selon les probabilités à priori des groupes. La probabilité d'affecter une observation au groupe k est estimée par $\frac{n_k}{n}$ et le nombre de bien classés dans le groupe k est $\frac{n_k^2}{n}$. Dans le cas de deux groupes, le taux d'erreur apparent est $1 - \frac{n_1^2 + n_2^2}{n^2}$.

Exemple.

6.1.5 Indicateurs dans le cas binaire

On suppose ici que l'on a construit (généralement sur un échantillon d'apprentissage) un règle de décision (ou encore un prédicteur) d'une variable Y prenant deux valeurs 0 et 1. Par convention on note généralement :

$$\begin{aligned} 1 &= \text{“positif”} = \text{“oui”} = \text{“groupe 1”}, \\ 0 &= \text{“négatif”} = \text{“non”} = \text{“groupe 2”}. \end{aligned}$$

On dispose donc que pour chaque individu i observé dans l'échantillon (généralement l'échantillon test), une valeur observée sur y_i et une valeur prédite \hat{y}_i . On peut alors construire la matrice de confusion qui est simplement le tableau de contingence suivant :

		valeur observée		
		$Y = 1$	$Y = 0$	total
valeur prédite	$\hat{Y} = 1$	TP	FP	\hat{P}
	$\hat{Y} = 0$	FN	TN	\hat{N}
total		P	N	n

où TP désigne les vrais positifs (true positive), TN les vrais négatifs (true negative), FP désigne les faux positifs (false positive) et FN désigne les faux négatifs (false negative).

A partir de ce tableau de contingence, une batterie d'indicateurs permettant de juger de la qualité du prédicateur peuvent être construits :

1. Le taux de bon classement aussi appelé accuracy (ACC) :

$$\text{ACC} = \frac{\text{TP} + \text{TN}}{n},$$

ou encore le taux de mauvais classement aussi appelé l'erreur de classement (err) :

$$\text{err} = \frac{\text{TP} + \text{TN}}{n} = 1 - \text{ACC}.$$

2. Les indicateurs du tableau des fréquences ligne sont :

	$Y = 1$	$Y = 0$
$\hat{Y} = 1$	$\text{PPV} = \frac{\text{TP}}{\hat{P}}$	$\text{FDR} = \frac{\text{FP}}{\hat{P}}$
$\hat{Y} = 0$	$\text{FOR} = \frac{\text{FN}}{\hat{N}}$	$\text{NPV} = \frac{\text{TN}}{\hat{N}}$
tous	$\frac{P}{n}$	$\frac{N}{n}$

où PPV désigne la valeur prédictive positive (positive predictive value), NPV désigne la valeur prédictive négative (negative predictive value), FDR correspond à *false discovery rate* et FOR correspond à *false omission rate*. La valeur prédictive positive VPP, encore appelée "Précision", doit être comparée à $\frac{P}{n}$ (appelée la prévalence en épidémiologie).

Exemple.

3. Les indicateurs du tableau des fréquences colonne sont :

	$Y = 1$	$Y = 0$
$\hat{Y} = 1$	$\text{TPR} = \frac{\text{TP}}{\text{P}}$	$\text{FPR} = \frac{\text{FP}}{\text{N}}$
$\hat{Y} = 0$	$\text{FNR} = \frac{\text{FN}}{\text{P}}$	$\text{TNR} = \frac{\text{TN}}{\text{N}}$

où TPR désigne le taux de vrais positifs (true positive rate), FPR désigne le taux de faux positifs (false positive rate), FNR désigne le taux de faux négatifs (false negative rate) et TNR désigne le taux de vrais négatifs (true negative rate). En pratique, on s'intéresse plus particulièrement aux deux indicateurs TPR et TNR avec :

- TPR, aussi appelé la sensibilité (Se) ou encore “Recall”, qui indique la proportion de 1 bien prédits,
- TNR, aussi appelé la spécificité (Sp) qui indique la proportion de 0 bien prédits,
- $\text{FPR} = 1 - \text{TPR} = 1 - \text{Spécificité}$ est aussi appelé “Fall-out”.

Une bonne règle de décision doit être à la fois sensible et spécifique.

Exemple.

Quelques liens :

http://en.wikipedia.org/wiki/F1_score#Diagnostic_Testing

http://cedric.cnam.fr/~saporta/Sensibilite_specificiteSTA201.pdf

6.2 Qualité d'un score

La prédiction \hat{Y} est généralement obtenue en comparant le score d'un individu à un seuil.

Dans le cas binaire ($K = 2$), nous avons vu deux types de fonction score S qui permettent de donner une note à une observation $\mathbf{x} \in \mathbb{R}^p$:

- Le score $S(\mathbf{x}) = \hat{P}(G_1 | \mathbf{x})$ que l'on compare généralement au seuil $s = 0.5$. Ces probabilités a posteriori peuvent être estimée par régression logistique, par les méthodes LDA, QDA....
- Le score $S(\mathbf{x}) = \hat{\beta}_0 + \hat{\beta}'\mathbf{x}$ que l'on compare généralement au seuil $s = 0$. Par exemple, $\hat{\beta}_0$ et $\hat{\beta}$ sont définis en (33) et (34) pour la méthode LDA ou encore obtenus par maximum de vraisemblance en régression logistique.

La règle de décision est alors :

$$\hat{y}_i = \begin{cases} 1 & \text{si } S(\mathbf{x}_i) \geq s, \\ 0 & \text{si } S(\mathbf{x}_i) < s, \end{cases}$$

Si l'on modifie le seuil s , on modifie la règle de décision, la matrice de confusion, et donc tous les indicateurs présentés précédemment (taux d'erreur, spécificité, sensibilité...).

On mesure souvent visuellement de l'efficacité d'un score S indépendamment du choix du seuil à partir de la courbe ROC (Receiver Operating Characteristic) et numériquement à partir de l'AUC (area under the curve).

6.2.1 Courbe ROC et AUC

La sensibilité et la la spécificité dépendent donc du seuil s choisi. On peut donc définir une fonction de sensibilité $Se(s) = TPR(s)$ et une fonction de spécificité $Sp(s) = TNR(s)$ et la courbe ROC représente l'antispécificité ($1 - Sp(s)$) en abscisse et la sensibilité en ordonnée $Se(s)$. La courbe ROC est alors la courbe

$$\{(1 - Sp(s), Se(s))_s\}$$

Si cette courbe coïncide avec la diagonale, c'est que le score n'est pas plus performant qu'un modèle aléatoire (où on attribue la classe au hasard). Plus la courbe ROC s'approche du coin supérieur gauche, meilleur est le modèle, car il permet de capturer le plus possible de vrais positifs avec le moins possible de faux positifs. En conséquence, l'aire sous la courbe ROC (AUC) peut être vu comme une mesure de la qualité du score. Ce critère AUC varie entre 0 (cas le pire) et 1 (cas le meilleure...).

Exemple.

Quelques liens :

<http://rocr.bioinf.mpi-sb.mpg.de/>

6.2.2 Choix d'un seuil

On utilise parfois la courbe ROC pour choisir un seuil. En pratique, on peut prendre le seuil correspondant au point de la courbe la plus éloigné de la première bissectrice et le plus prêt du point supérieur gauche $(0, 1)$. Ou encore le seuil correspondant au point où la pente de la courbe est la plus proche de 0.

Mais on peut également choisir le seuil qui minimise le taux d'erreur (ou un autre indicateur) sur l'ensemble d'apprentissage.

Exemple.

6.2.3 Courbe lift et indice de Gini

7 Annexe

Exercice 1. Démontrez (1) et (2).

Preuve de (1) : $\sum_{k=1}^K \frac{n_k}{n} \mathbf{g}_k = \sum_{k=1}^K \frac{n_k}{n} \sum_{i \in E_k} \frac{1}{n_k} \mathbf{x}_i = \mathbf{g}$.

Preuve de (2) :

$$\begin{aligned} \mathbf{V} &= \frac{1}{n} \sum_{i=1}^n (\mathbf{x}_i - \mathbf{g})(\mathbf{x}_i - \mathbf{g})' = \frac{1}{n} \sum_{k=1}^K \sum_{i \in E_k} (\mathbf{x}_i - \mathbf{g})(\mathbf{x}_i - \mathbf{g})', \\ &= \frac{1}{n} \sum_{k=1}^K \sum_{i \in E_k} (\mathbf{x}_i - \mathbf{g}_k + (\mathbf{g}_k - \mathbf{g}))(\mathbf{x}_i - \mathbf{g}_k + (\mathbf{g}_k - \mathbf{g}))', \\ &= \frac{1}{n} \left[\sum_{k=1}^K \sum_{i \in E_k} (\mathbf{x}_i - \mathbf{g}_k)(\mathbf{x}_i - \mathbf{g}_k)' + \sum_{k=1}^K \sum_{i \in E_k} (\mathbf{g}_k - \mathbf{g})(\mathbf{g}_k - \mathbf{g})' + 2 \sum_{k=1}^K \sum_{i \in E_k} (\mathbf{x}_i - \mathbf{g}_k)(\mathbf{g}_k - \mathbf{g})' \right]. \end{aligned}$$

et

$$\begin{aligned}
\frac{1}{n} \sum_{k=1}^K \sum_{i \in E_k} (\mathbf{x}_i - \mathbf{g}_k)(\mathbf{x}_i - \mathbf{g}_k)' &= \mathbf{W} \\
\frac{1}{n} \sum_{k=1}^K \sum_{i \in E_k} (\mathbf{g}_k - \mathbf{g})(\mathbf{g}_k - \mathbf{g})' &= \frac{1}{n} \sum_{k=1}^K n_k (\mathbf{g}_k - \mathbf{g})(\mathbf{g}_k - \mathbf{g})' = \mathbf{B} \\
\sum_{k=1}^K \sum_{i \in E_k} (\mathbf{x}_i - \mathbf{g}_k)(\mathbf{g}_k - \mathbf{g})' &= 0
\end{aligned}$$

Exercice 2. Démontrez (3) (4) et (5).

Preuve de (3) : la variable $\mathbf{s} = \mathbf{X}\mathbf{u}$ est une combinaison linéaire des colonnes de \mathbf{X} centrées donc \mathbf{s} centrée et $\bar{\mathbf{s}} = 0$. Donc

$$\begin{aligned}
Var(\mathbf{s}) &= \frac{1}{n} \sum_{i=1}^n s_i^2 \\
&= \frac{1}{n} \mathbf{s}' \mathbf{s} = \frac{1}{n} \mathbf{u}' \mathbf{X}' \mathbf{X} \mathbf{u} \\
&= \mathbf{u}' \mathbf{V} \mathbf{u}
\end{aligned}$$

Preuve de (4) :

$$\begin{aligned}
\mathbf{u}' \mathbf{W} \mathbf{u} &= \mathbf{u}' \left(\sum_{k=1}^K \frac{n_k}{n} \mathbf{V}_k \right) \mathbf{u} = \sum_{k=1}^K \frac{n_k}{n} \mathbf{u}' \mathbf{V}_k \mathbf{u}, \\
&= \sum_{k=1}^K \frac{n_k}{n} \mathbf{u}' \left[\sum_{i \in E_k} \frac{1}{n_k} (\mathbf{x}_i - \mathbf{g}_k)(\mathbf{x}_i - \mathbf{g}_k)' \right] \mathbf{u}, \\
&= \sum_{k=1}^K \frac{n_k}{n} \sum_{i \in E_k} \frac{1}{n_k} \mathbf{u}' (\mathbf{x}_i - \mathbf{g}_k)(\mathbf{x}_i - \mathbf{g}_k)' \mathbf{u}, \\
&= \sum_{k=1}^K \frac{n_k}{n} \sum_{i \in E_k} \frac{1}{n_k} [\mathbf{x}_i' \mathbf{u} - \mathbf{g}_k' \mathbf{u}]^2, \\
&= \sum_{k=1}^K \frac{n_k}{n} \sum_{i \in E_k} \frac{1}{n_k} (s_i - \bar{s}_k)^2 = Intra(\mathbf{s}).
\end{aligned}$$

Exercice 2. Montrer que le rapport $\frac{\mathbf{u}' \mathbf{A} \mathbf{u}}{\mathbf{u}' \mathbf{B} \mathbf{u}}$ est maximal pour u_1 vecteur propre de $\mathbf{B}^{-1} \mathbf{A}$ associé à la plus grande valeur propre notée λ_1 , la valeur du maximum étant λ_1 .

Preuve : $f(\mathbf{u}) = \frac{\mathbf{u}' \mathbf{A} \mathbf{u}}{\mathbf{u}' \mathbf{B} \mathbf{u}} \in \mathbb{R}$ et

$$\frac{\partial f}{\partial \mathbf{u}} = \frac{\frac{\partial \mathbf{u}' \mathbf{A} \mathbf{u}}{\partial \mathbf{u}} (\mathbf{u}' \mathbf{B} \mathbf{u}) - (\mathbf{u}' \mathbf{A} \mathbf{u}) \frac{\partial \mathbf{u}' \mathbf{B} \mathbf{u}}{\partial \mathbf{u}}}{(\mathbf{u}' \mathbf{B} \mathbf{u})^2}.$$

$$\text{On a } \frac{\partial \mathbf{u}' \mathbf{M} \mathbf{u}}{\partial \mathbf{u}} = \begin{cases} 2\mathbf{M} \mathbf{u} \text{ si } \mathbf{M} \text{ est symétrique,} \\ (\mathbf{M} + \mathbf{M}') \mathbf{u} \text{ sinon.} \end{cases}$$

D'où :

$$\frac{\partial f}{\partial \mathbf{u}} \Big|_{\mathbf{u}=\mathbf{u}_1} = 0,$$

$$2\mathbf{A} \mathbf{u} (\mathbf{u}' \mathbf{B} \mathbf{u}) - (\mathbf{u}' \mathbf{A} \mathbf{u}) 2\mathbf{B} \mathbf{u} = 0,$$

$$\mathbf{A} \mathbf{u} = \frac{\mathbf{u}' \mathbf{A} \mathbf{u}}{\mathbf{u}' \mathbf{B} \mathbf{u}} \mathbf{B} \mathbf{u},$$

$$\mathbf{B}^{-1} \mathbf{A} \mathbf{u} = \frac{\mathbf{u}' \mathbf{A} \mathbf{u}}{\mathbf{u}' \mathbf{B} \mathbf{u}} \mathbf{u}.$$

Exercice 3. Démontrez :

a) $\mathbf{B} = \frac{n_1 n_2}{n^2} (\mathbf{g}_1 - \mathbf{g}_2)(\mathbf{g}_1 - \mathbf{g}_2)'$.

b) $\mathbf{V}^{-1}(\mathbf{g}_1 - \mathbf{g}_2)$ est vecteur propre de $\mathbf{V}^{-1} \mathbf{B}$,

$\lambda = \frac{n_1 n_2}{n^2} (\mathbf{g}_1 - \mathbf{g}_2)' \mathbf{V}^{-1} (\mathbf{g}_1 - \mathbf{g}_2)$ est la valeur propre de $\mathbf{W}^{-1} \mathbf{B}$.

Preuves :

a) On sait que $\mathbf{g} = \frac{n_1}{n} \mathbf{g}_1 + \frac{n_2}{n} \mathbf{g}_2$ donc $\mathbf{g}_2 = -\frac{n_1}{n} \mathbf{g}_1$ et $\mathbf{g}_1 - \mathbf{g}_2 = \mathbf{g}_1 + \frac{n_1}{n_2} \mathbf{g}_1 = \frac{n_1 + n_2}{n_2} \mathbf{g}_1 = \frac{n}{n_2} \mathbf{g}_1$.

D'où $\frac{n_1 n_2}{n^2} (\mathbf{g}_1 - \mathbf{g}_2)(\mathbf{g}_1 - \mathbf{g}_2)' = \frac{n_1 n_2}{n^2} \frac{n}{n_2} \mathbf{g}_1 \frac{n}{n_2} \mathbf{g}_1' = \frac{n_1}{n_2} \mathbf{g}_1 \mathbf{g}_1'$.

De plus on a :

$$\begin{aligned} \mathbf{B} &= \frac{n_1}{n} \mathbf{g}_1 \mathbf{g}_1' + \frac{n_2}{n} \mathbf{g}_2 \mathbf{g}_2' = \frac{n_1}{n} \mathbf{g}_1 \mathbf{g}_1' + \frac{n_2}{n} \frac{n_1}{n_2} \mathbf{g}_1 \frac{n_1}{n_2} \mathbf{g}_1' \\ &= \frac{n_1}{n} \mathbf{g}_1 \mathbf{g}_1' + \frac{n_1^2}{n n_2} \mathbf{g}_1 \mathbf{g}_1' \\ &= \frac{n_1 n_1 + n_1^2}{n n_2} \mathbf{g}_1 \mathbf{g}_1' = \frac{n_1 (n_1 + n_2)}{n n_2} \mathbf{g}_1 \mathbf{g}_1' \\ &= \frac{n_1}{n_2} \mathbf{g}_1 \mathbf{g}_1' \end{aligned}$$

b) On a :

$$\begin{aligned} \mathbf{V}^{-1} \mathbf{B} \mathbf{V}^{-1} (\mathbf{g}_1 - \mathbf{g}_2) &= \mathbf{V}^{-1} \frac{n_1 n_2}{n^2} (\mathbf{g}_1 - \mathbf{g}_2)(\mathbf{g}_1 - \mathbf{g}_2)' \mathbf{V}^{-1} (\mathbf{g}_1 - \mathbf{g}_2) \\ &= \mathbf{V}^{-1} (\mathbf{g}_1 - \mathbf{g}_2) \underbrace{\frac{n_1 n_2}{n^2} (\mathbf{g}_1 - \mathbf{g}_2)' \mathbf{V}^{-1} (\mathbf{g}_1 - \mathbf{g}_2)}_{\text{scalaire}} \end{aligned}$$