Classification supervisée ou Analyse discriminante

1 Introduction

Contrairement à la classification non supervisée, on suppose ici avoir des observations déjà regroupées en classes (aussi appelées groupes). On dispose donc de n individus décrits par Y une variable qualitative à K modalités (avec K connu) indiquant la classe de chaque individu. Ces individus sont décrits également par p variables explicatives (aussi appelées prédicteurs).
L'objectif de l'analyse discriminate est :

- descriptif : trouver les combinaisons linéarires de variables qui permettent de séparer au mieux les n individus en K groupes ; comprendre le phénomène qui conduit à ce regroupement en classes
- décisionnel : un nouvel individu se présente pour lequel on observe les prédicteurs. Il s'agit de décider dans quel groupe l'affecter.

Exemples d'applications :

- Médecine : connaissant les symptômes présentés par un patient, peut-on prédire un risque médical (survie/mort ...) ?
- Finance : A partir des bilans d'une société, est-il possible d'estimer son risque de faillite à 3 ans ?
- Marketting : connaissant les caractéristiques d'un client, on cherche à prédire le risque qu'il passe à la concurrence ou qu'il achète tel ou tel produit.

Remarques. L'apprentissage supervisée regroupe plusieurs problèmes.

- la classification : la variables à expliquer Y est qualitative. Les variables explicatives peuvent être quantitatives et/ou qualitatives.
- la régression : la variable à expliquer est quantitative.

Méthodes discriminantes : nous verrons ces deux points de vue.

- Approche "géométrique" : Analyse factorielle discriminante (AFD)
- Approche probabiliste :
  - paramétrique : analyse disciminante linéaire et quadratique (LDA, QDA)
  - non paramétrique : méthode à noyaux, k-plus proches voisins, arbres de classification (CART)
  - "semi-paramétrique" : régression logistique.

Il existe également d'autres approches dont (cours apprentissage an prochain) :

- Réseaux de neurones et perceptron
- SVM (Support Vector Machines)

2 Analyse factorielle discriminante

Dans cette section, les variables explicatives sont quantitatives.
L'approche géométrique revient à chercher un "bon" axe discriminant, i.e qui permette de séparer au mieux les K classes. On est dans l'esprit de l'ACP.

2.1 Notations

Données. Sur les n individus de l'échantillon, on a observé Y une variable qualitative à K classes. On note donc $y_i \in \{1, \dots, K\}$ l'observation pour le i-ème individu. X est la matrice des variables explicatives quantitatives. Elle est de taille (n, p). Comme dans les précédents cours, la ligne x_i décrit le i-ème individu et la colonne x^j la j-ème variable.

Nuage de points. Comme en ACP on s'intéresse au nuage des points-individus, et comme en ACP on va travailler sur les données centrées : on suppose donc que $\bar{x}^j = 0 \quad \forall j$, i.e que le centre de gravité du nuage est l'origine 0.

On note S la matrice de variance-covariance empirique des données (ici le centre de gravité du nuage est $g = 0$) :

$$S = \frac{1}{n}\sum_{i=1}^{n}(x_i - g)(x_i - g)^T \quad \text{de taille } (p, p)$$

Géométrie d'un seul nuage. Si l'on avait un groupe au lieu de K, l'orientation du nuage serait décrite à partir d'une ACP au moyen du premier vecteur propre de la matrice S de façon à maximiser l'inertie projetée du nuage (i.e maximiser la dispersion du nuage projeté).

Géométrie de K nuages. On suppose maintenant que l'on a K groupes et donc que K nuages sont observés. On note

I_k les individus du groupe k et n_k son effectif.
$g_k \in \mathbb{R}^p$: le centre de gravité du groupe k
S_k la matrice variance-covariance empirique du nuage k :

$$S_k = \frac{1}{n_k}\sum_{i \in I_k}(x_i - g_k)(x_i - g_k)^T \quad \text{de taille } (p, p)$$

Définition. On définit

— la matrice de variance intra-classe (Within) comme la matrice pondérée des matrices de variance-covariances des K nuages :

$$W = \sum_{k=1}^{K}\frac{n_k}{n}S_k \quad \text{de taille } (p, p)$$

— la matrice de variance inter-classe (Between) comme la matrice pondérée des K centres de gravité. Elle décrit donc la géométrie du nuage formé par les K centres de gravité.

$$B = \frac{1}{n}\sum_{k=1}^{K} n_k(g_k - g)(g_k - g)^T \quad \text{de taille } (p, p)$$

Propriété 1. La matrice de variance des données se décompose ainsi :

$$S = W + B$$

Rappel : En classification non supervisé, on a déjà parlé d'inertie inter et intra classe. On a

$$I_T = Tr(S) \quad I_{inter} = Tr(B) \quad I_{intra} = Tr(W)$$

Remarque. En général W est inversible (c'est ce que l'on supposera dans la suite) mais B n'est pas inversible (cela vient du fait que les K centres de gravité vivent dans un sous-espace de dimension $K-1$ puisque $\sum_{k=1}^{K} n_k g_k = g = 0$)

2.2 Analyse factorielle discriminante

Elle consiste à chercher de nouvelles variables (les variables discriminantes) correspondant à des directions de $\mathbb{R}^p$ qui séparent le mieux possible en projection les K groupes.
Intuition :

On cherhe donc un axe qui :

1. sépare au mieux les K groupes ie tel que les centres de gravité projetés soient éloignés.
2. tout en regroupant bien les individus d'un même groupe, ie tel que les groupes projetés ne soient pas trop dispersés.

Rappel : en ACP l'inertie du nuage projeté sur l'axe dirigé par a (supposé unitaire) est : $a^T Sa$.
En analye discriminante, on cherche donc un axe a tel que :

1. l'inertie du nuage des centres de gravité g_k projeté sur a soit maximale :

$$\max_a a^T Ba$$

2. l'inertie intra-groupe projetée sur a soit minimale :

$$\min_a a^T Wa$$

La relation $S = W + B$ entraîne que $a^T Sa = a^T Wa + a^T Ba$. On va donc chercher à maximiser le rapport de l'inertie inter-classe à l'inertie totale (notation française), ou le rapport de l'inertie inter-classe à l'inertie intra-classe (notation anglo-saxone) :

$$\frac{a^T Ba}{a^T Sa} \text{ ou } \frac{a^T Ba}{a^T Wa}$$

Propriété 2. (notation française)
Le maximum est atteint pour u_1 vecteur propre de $S^{-1}B$ associé à la plus grande valeur propre λ_1, la valeur du maximum étant λ_1. $u_1 \in \mathbb{R}^p$ est appelé 1er axe discriminant et Xu_1 1ère variable discriminante. La valeur propre $\lambda_1 = \frac{u_1^T Bu_1}{u_1^T Su_1} \in [0, 1]$ donne une mesure du pouvoir discriminant d'un axe.
Et ainsi de suite : les vecteur propres de $S^{-1}B$ obtenus en ordonnant les valeurs propres par ordre décroissant correspondent aux axes discriminants de l'AFD. On peut construire $K-1$ axes discriminants car $rang(S^{-1}B) \leq min(p, K-1)$ avec en général $K - 1 < p$.

Cas particuliers :

— Cas $\lambda_1 = 1$
On a alors $u_1^T Bu_1 = u_1^T Su_1$ ie $u_1^T Wu_1 = 0$. Les k nuages sont donc chacun dans un hyperplan orthogonal au 1er axe discriminant. Il y a donc discrimination parfaite si les centres de gravité se projettent en des points différents.

Notons qu'on peut avoir $\lambda_1 < 1$ et des groupes tout de même bien séparés.

— Cas $\lambda_1 = 0$
On a alors $u_1^T Bu_1 = 0$: le meilleur axe ne permet pas de séparer les centres de gravité. C'est par exemple le cas lorsque les nuages sont concentriques et qu'aucune séparation linéaire n'est possible.

Notons qu'il peut cependant exister une discrimination non linéaire.

Remarques.

1. Il faut que la matrice de variances-covariances S soit inversible, et donc que la matrice des données X soit de plein rang. Il faudra donc, comme en régression multiple, faire attention aux problèmes de colinéarité entre variables.
2. Les vecteurs propres sont orthogonaux donc les axes discriminants sont orthogonaux.
3. Notation anglo-saxone : Le 1er axe discriminant est alors le vecteur propre de $W^{-1}B$ associé à la plus grande valeur propre μ_1. Or on peut montrer que $W^{-1}B$ et $S^{-1}B$ ont les mêmes vecteurs propres avec $\mu_1 = \frac{\lambda_1}{1-\lambda_1}$. Les axes discriminants sont donc les mêmes, mais au lieu d'avoir des valeurs propres dans $[0, 1]$, elles sont dans $[0, \infty[$. Moins facile à interpréter, mais c'est souvent ce que donnent les logiciels. C'est pourquoi, dans la suite du cours, on continue avec les notations anglo-saxones.

4. L'AFD peut être vue comme une ACP des centres de gravité en utilisant non pas la distance euclidienne, mais la distance basé sur S^{-1} (notation française) ou W^{-1} (notation anglo-saxone) appelée distance de Mahalanobis.

   Pour rappel en ACP, la distance entre individus est $d^2(x_i, x_{i'}) = (x_i - x_{i'})'M(x_i - x_{i'})$ avec :
   — Si $M = I$: distance euclidienne
   — Si $M = diag(1/s_1, \ldots, 1/s_p)$: distance euclidienne des données réduites (permet de traiter des varibles hétérogènes)
   — Si $M = W^{-1}$: distance de Mahalanobis (permet de tenir compte de la corrélation entre variables)

   Voici donc un autre intêret de la méthode géométrique AFD : comme en ACP, on peut projeter les individus sur les plans principaux et interpréter les variables discriminantes au moyen d'un cercle des corrélations. Ce qui permet d'expliciter le classement selon les groupes.

2.3 Cas particuliers de deux classes

En pratique on étudie souvent le cas de 2 classes (bon/mauvais client, sain/malade...)
On cherche donc un seul axe discriminant puisque $K - 1 = 1$. Cet axe discriminant est engendré par la droite reliant les 2 centres de gravité des 2 classes.

2.4 Régle géométrique de classement d'un nouvel individu

Ayant trouvé la meilleure représentation de la séparation en K classes des n individus, on peut alors chercher à affecter un nouvel individu à l'une des classes. Pour cela on calcule les distances de ce nouvel individu avec le centre de gravité de chacunes des classes, et on affecte l'individu à la classe la plus proche.
Quelle distance utiliser ? On utilise la distance utilisée pour faire l'ACP des centres de gravité, i.e la distance de Mahalanobis.

Propriété 3. Règle de Mahalanobis-Fisher
Affecter le nouvel individu x à la classe k^* telle que :

$$k^* = arg \min_{k=1,\ldots K} d^2(x, g_k)$$

avec

$$d^2(x, g_k) = (x - g_k)'W^{-1}(x - g_k)$$

Cela se réécrit, en développant la forme quadratique, W^{-1} étant symétrique :

$$k^* = arg \max_{k=1,\ldots K} x'W^{-1}g_k - \frac{1}{2}g_k'W^{-1}g_k$$

Cette quantité étant linéaire en x, les zones de séparations (ou frontière de décision) sont des hyperplans. (Voir TP : si $p = 2$, la frontière de décision est une droite)

Remarque. Cas de deux groupes : On affecte l'individu x au groupe 1 si

$$S(x) = x'W^{-1}(g_1 - g_2) - \frac{1}{2}(g_1 + g_2)'W^{-1}(g_1 - g_2) > 0$$

$S(x)$ s'appelle le score de Fisher.

2.5 Insuffisance de la règle géométrique

L'utilisation de la règle précédente conduit à des affectations incorrectes lorsque les dispersions (variances) des groupes sont très différentes entre elles, ou que les groupes ne sont pas du tout de la même taille (probabilité à priori différentes). D'où les modèles probabilistes (LDA et QDA. voir TP)

3 Méthodes probabilistes

En AFD on a défini une règle d'affectation d'un nouvel individu, fondée sur des arguments géométriques (affectation "à la classe la plus proche"). Dans cette section, on va définir une règle de décision bayésienne permettant d'affecter le nouvel individu à la classe "la plus probable".

3.1 Notion de modèle de mélange

Le modèle de mélange consiste à supposer que les données sont issues de plusieurs sous-populations. Chaque sous-population est modélisée par sa propre loi. La population totale est un mélange de ces sous-populations. Par exemple, on s'intéresse à la taille de la population française. La loi de X : "taille adulte en france" est une loi mélange entre la taille des hommes et celle des femmes.

Exemple. modèle de mélange gaussien univarié :

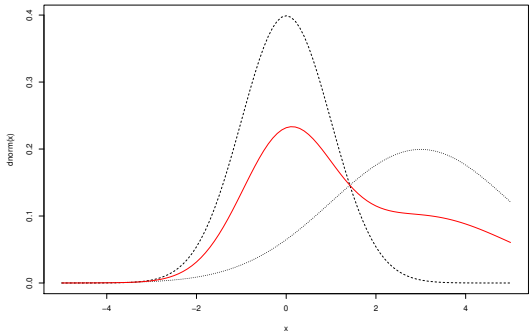


FIGURE 1 – $f_1(x) = dnorm(x, 0, 1), f_2(x) = dnorm(x, 3, 2), p_1 = 0.5, p_2 = 0.5$

- chaque sous population suit une loi gaussienne :
  individus de la 1ère sous-population (ie sachant que $Y = 1$) : $X|(Y = 1) \sim \mathcal{N}(0, 1)$ (densité f_1)
  individus de la 2ème sous-population (ie sachant que $Y = 2$) : $X|(Y = 2) \sim \mathcal{N}(3, 4)$ (densité f_2)
- on note p_1 et p_2 les proportions du mélange, i.e les probabilités d'appartenance aux 2 classes.
  ici on a choisi $p_1 = p_2$

La population totale est un mélange de deux gaussiennes. On note

$$f_X(x) = p_1 f_1(x) + p_2 f_2(x)$$

f_X est bien une densité : $\int_{\mathbb{R}} f_X(x)dx = 1$ car f_1 et f_2 sont des densités et que $p_1 + p_2 = 1$. C'est la densité "mélange" de la population.

On peut généraliser pour un mélange de K gaussiennes. On note

$$f_X(x) = \sum_{k=1}^{K} p_k f_k(x)$$

avec p_k les proportions du mélange telles que $\sum_{k=1}^{K} p_k = 1$.

Exemple. modèle de mélange gaussien p-varié :

Dans l'exemple précédent, X était une variable aléatoire. On peut généraliser la notion de modèle de mélange pour un vecteur aléatoire. Soit X le vecteur aléatoire défini par $X = (X^1, \ldots, X^p)$. X est une loi mélange si

$$f_X(x) = \sum_{k=1}^{K} p_k f_k(x)$$

avec p_k les proportions du mélange, et f_k les densités du mélange qui sont ici des densités multidimensionnelles :

$$(X^1, \ldots, X^p)|Y = k \sim \mathcal{N}_p(\mu_k, \Sigma_k)$$

i.e conditionnellement à $Y = k$, le vecteur aléatoire X est un vecteur gaussien $\mathcal{N}_p(\mu_k, \Sigma_k)$ avec μ_k vecteur de $\mathbb{R}^p$ et Σ_k matrice de variance-covariance de taille (p, p).

Remarque. Ici sont présentés les modèles de mélange gaussien, mais n'importe quelle autre densité peut-être utilisée.

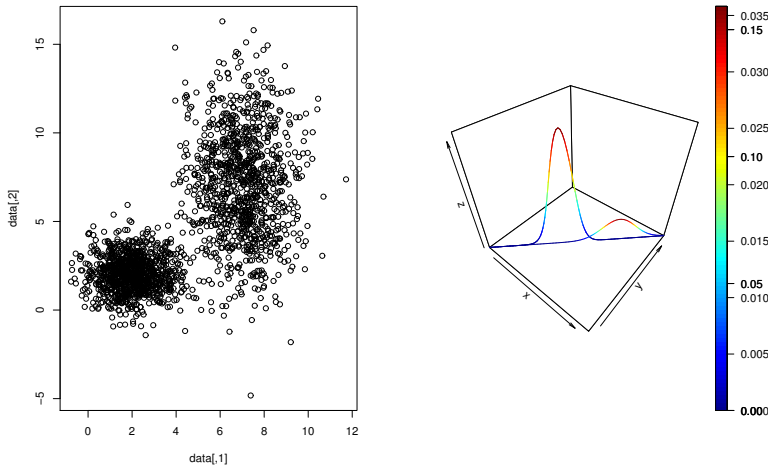


FIGURE 2 – Mélange gaussien bivarié $p = 2$

3.2 Modèle probabiliste et règle d'attribution de Bayes

On suppose que les variables en jeu sont des variables aléatoires notées Y et X^j $(j = 1, \dots, p)$, dont on observe des réalisations $y_1, \dots, y_n$ et $x_1^j, \dots x_n^j$ sur n individus.
On suppose dans cette section que les variables X^j sont quantitatives .

Notations et définitions :

- Y est une variable qualitative à K modalités. On note $p_k = \mathbb{P}(Y = k)$ la probabilité d'appartenance à la classe k. Ces propabilités $p_1, \dots, p_K$ sont dites probabilités **à priori**. Elles sont inconnues mais peuvent être estimés par $\frac{n_k}{n}$ la proportion d'individus dans chaque classe.
- On suppose que la loi du vecteur de variables aléatoires $X = (X^1, \dots, X^p)$ est une loi "mélange", i.e que conditionnellement à $Y = k$, le vecteur aléatoire X admet une densité f_k.

$$X = (X^1, \dots, X^p)|(Y = k) \quad \sim f_k \text{ ou autrement dit } f_{X|Y=k}(x) = f_k(x)$$

- On a une loi mélange et X est de densité :

$$f_X(x) = \sum_{k=1}^{K} p_k f_k(x)$$

Remarque. Les modèles de mélange servent également en classification non supervisée (algorithme EM pour Expectation-maximization, cours Apprentissage l'an prochain).

Propriété 4. Règle du MAP (maximum a posteriori) : on va affecter un nouvel individu dont on observe la réalisation $x = (x^1, \dots x^p)$ au groupe le plus probable au vue de cette réalisation, i.e au groupe :

$$k^* = arg \max_{k=1,\dots,K} \mathbb{P}(Y = k|X = x)$$

La probabilité conditionnelle $P(Y = k|X = x)$ s'appelle la probabilité **à postériori** du groupe k. D'après la formule de Bayes

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

La règle du MAP se réecrit donc, puisque le dénominateur ne dépend pas de k :

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

En pratique on a donc besoin des f_k et des p_k. Plusieurs approches sont possibles :

1. Approche paramétrique :
   (a) Approche la plus simple : Naive Bayes
   Supposer que les variables X^j sont indépendantes conditionnellement à Y, ie que

$$\mathbb{P}(X = x|Y = k) = \mathbb{P}(X^1 = x^1, \ldots, X^n = x^n|Y = k) = \Pi_{j=1}^n \mathbb{P}(X^j = x^j|Y = k)$$

   On se ramène alors à un problème paramétrique unidimensionnselle : on modélise chaque variable $X^j|(Y = k)$ par exemple avec une loi gaussienne unidimentionnelle $\mathcal{N}(\mu_{j,k}, \sigma^2_{j,k})$
   (b) Analyse discriminante : on suppose que f_k est une densité connue de $\mathbb{R}^p$.
   Exemple. : cas mélange gaussien p-dimensionnel avec f_k densité de la loi $\mathcal{N}_p(\mu_k, \Sigma_k)$ et on estime les paramètres inconus par maximum de vraisemblance.
2. Approche non paramétrique : Noyaux, K plus-proches voisins, CART
3. Approche "semi-paramétrique" : régression logistique

3.3 Analyse discriminante : cas gaussien

On suppose que dans chaque groupe k, la loi du vecteur $X = (X^1, \ldots, X^p)$ peut être modélisée par une loi gaussienne multidimensionnelle de densité :

$$x \in \mathbb{R}^p, \quad f_k(x) = \frac{1}{(2\pi)^{p/2}\sqrt{\det \Sigma_k}} \exp\left[-\frac{1}{2}(x - \mu_k)'\Sigma_k^{-1}(x - \mu_k)\right].$$

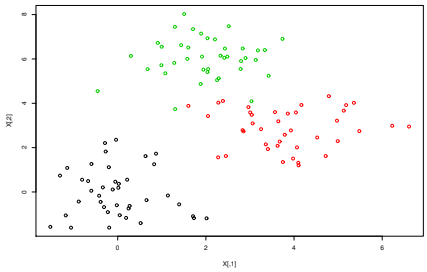
avec $\mu_k \in \mathbb{R}^p$ le vecteur des espérances et Σ_k la matrice (p, p) de variances-covariances (théorique) pour le groupe k. (cette densité est la généralisation de la densité d'une variable gaussienne).

La régle du MAP se réécrit alors (en passant au ln et en négligeant les termes ne dépendant pas de k) :

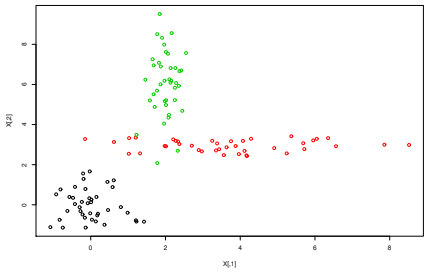
$$\begin{aligned} k^* &= arg \max_{k=1,\ldots,K} p_k f_k(x) \\ &= arg \max_{k=1,\ldots,K} ln(p_k) - \frac{1}{2} ln(\det \Sigma_k) - \frac{1}{2}(x - \mu_k)'\Sigma_k^{-1}(x - \mu_k) \end{aligned}$$

Les paramètres μ_k, Σ_k et pi_k sont inconnus. On les estime à partir de l'échantillon d'apprentissage par maximum de vraisemblance.

LDA et QDA
Si l'on peut supposer que toutes les matrices de variance-covariance sont égales $\Sigma_1 = \ldots = \Sigma_K = \Sigma$, on parle de l'analyse discriminante linéaire (LDA). Si les matrices de variance-covariance ne peuvent pas être supposées égales, on parle de l'analyse discriminante quadratique (QDA).



homoscédasticité



hétéroscédasticité

L'hypothèse d'homoscédasticité, lorqu'elle est vérifiée, permet de réduire très sensiblement le nombre de paramètres à estimer.
Il existe un test sur l'égalité des matrices de variance-covariance : c'est le test de Box'M (avec H_0 : homoscédasticité). C'est la version multidimensionnelle du test de Bartlett.
`https://cran.r-project.org/web/packages/biotools/biotools.pdf`
Lorsque l'on rejette l'égalité, il faudrait donc utiliser la régle QDA. Le problème est que dans ce cas, l'échantillon d'apprentissage doit être de taille importante car beaucoup de paramètres sont à estimer. Parfois, il vaut tout de même mieux utiliser la régle LDA et valider le classement effectuée avec la régle choisie (voir section 6).

3.3.1 Analyse discriminante linéaire ADL

Propriété 5. Les estimateurs de maximum de vraisemblance de p_k, μ_k et deΣ sont :

$$\begin{aligned}\hat{p}_k &= \frac{n_k}{n}\\ \hat{\mu_k} &= g_k \text{ le centre de gravité du groupe } k\\ \hat{\Sigma} &= W = \sum_{k=1}^{K} \frac{n_k}{n} S_k \text{ la matrice de variance intra-classe (ou son estimateur sans biais)} \frac{n}{n-k} W\end{aligned}$$

La version plug-in (ie en remplaçant les quantités inconnues par leur estimateur) de la règle du MAP se simplifie :

$$k^* = arg \max_{k=1,\ldots,K} L_k(x)$$

avec

$$L_k(x) = ln(\frac{n_k}{n}) - \frac{1}{2} g_k' W^{-1} g_k + x' W^{-1} g_k$$

Les fonctions discriminantes $L_k(x)$ sont des fonctions linéaires en x d'où le nom de LDA. Les frontières de décision entre classes sont alors des hyperplans (voir TP).

Propriété 6. Lien LDA et AFD.
Si on suppose que les probabilités a priori sont égales (on parle d'équiprobabilité) alors on retrouve la régle géométrique vue en AFD. Les approches géométriques et probabiliste (dans le cas homoscédastique et équiprobable) sont équivalentes.
Si les probabilités ne sont pas égales, alors dans la régle du MAP se trouve un terme supplémentaire $ln(\frac{n_k}{n})$.

Remarque. Différence dans le cas de deux classes : on affecte l'individu x au groupe 1 si

$$\begin{aligned}\text{en AFD} &: x' W^{-1}(g_1 - g_2) - \frac{1}{2}(g_1 + g_2)' W^{-1}(g_1 - g_2) > 0\\ \text{en LDA} &: x' W^{-1}(g_1 - g_2) - \frac{1}{2}(g_1 + g_2)' W^{-1}(g_1 - g_2) + ln(\frac{n_1}{n_2}) > 0\end{aligned}$$

La régle probabiliste va donc avoir tendance, par rapport à la régle géométrique, à pénaliser les classes de petits effectifs. Ou autrement dit à favoriser le pouvoir d'attraction des classes à grands effectifs (voir TP).

3.3.2 Analyse discriminante quadratique QDA

Propriété 7. Les estimateurs de maximum de vraisemblance de p_k, μ_k et de Σ sont :

$$\begin{aligned}\hat{p}_k &= \frac{n_k}{n}\\ \hat{\mu_k} &= g_k \text{ le centre de gravité du groupe } k\\ \hat{\Sigma}_k &= S_k \text{ la matrice de variance covariance du groupe } k\end{aligned}$$

La régle du MAP se simplifie :

$$k^* = arg \max_{k=1,\ldots,K} Q_k(x)$$

avec

$$Q_k(x) = ln(\frac{n_k}{n}) - \frac{1}{2} ln(\det S_k) - \frac{1}{2}(x - g_k)' S_k^{-1}(x - g_k)$$

Les fonctions discriminantes $Q_k(x)$ sont des fonctions quadratiques d'où le nom de QDA. Les frontières de décision entre classes sont alors des hyper-surface dénifie par une equation quadratique (voir TP).

3.4 Approche non paramétrique

Dans la section précédente, on a supposé que le vecteur X suit, conditionnellement à la variable Y, une loi gaussienne. En pratique c'est rarement vérifié. Cependant, lorsque les variables X^j sont quantitatives et ont une distribution symétrique alors il est usuel de supposer que l'hypothèse de normalité est approximativement vérifiée. Dans ce cas en effet les méthodes basées sur des hypothèses de normalité sont généralement robustes et les résultats restent interprétables. Mais quand les variables sont de distribution fortement dissymétrique ou si elles sont qualitatives, il faut envisager d'autres approches.
Une méthode consiste à transformer les variables de façon à les rendre gausiennes. On peut essayer une transformation log par exemple.
Sinon, on estime la densité par une méthode non-paramétrique.

Idée : Estimateur par histogramme d'une densité
Supposons que l'on dispose de n observations $x_1, \ldots, x_n$ provenant de n variables aléatoires $X_1, \ldots, X_n$ i.i.d de densité f. f est inconnue et on cherche à l'estimer. Supposons pour simplifier que la densité est à support dans $[0, 1]$. On peut estimer f par un histogramme, i.e une fonction constante par morceaux. On découpe l'intervalle $[0, 1]$ en D intervalles I_j de longueur $h = 1/D$. On estime alors $f(x)$ par

$$\hat{f}(x) = \sum_{j=1}^{D} \frac{N_j}{nh} \mathbf{1}_{I_j}(x)$$

où N_j est le nombre d'observations parmi les $x_1, \ldots, x_n$, qui dans l'intervalle I_j.

On obtient bien une densité car

$$\int_0^1 \hat{f}(x) = \sum_{j=1}^{D} \frac{N_j}{nh} \int_0^1 \mathbf{1}_{I_j}(x) = \sum_{j=1}^{D} \frac{N_j}{nh} |I_j| = \sum_{j=1}^{D} \frac{N_j}{n} = 1$$

Le paramètre h (ou cela revient au même le paramètre D) contrôle le lissage de la densité estimée.

Idées de l'estimation par histogramme :
— tenir compte des données dans le voisinage de x
— besoin d'un paramètre de lissage
Voici en classification supervisée deux approches non paramétriques différentes pour délimiter ce voisinage local autour de x.

3.5 Estimateurs à noyau

Les estimateurs à noyau sont des estimateurs non paramétriques couramment utilisés en statistique. L'idée est celle de l'histogramme, mais on utilise un noyau au lieu d'une indicatrice, ce qui va donner un estimateur plus lisse de la densité. On estime une densité f par

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1\ldots n} \mathrm{K}\left(\frac{x - x_i}{h}\right)$$

où $h > 0$ joue le rôle du paramètre de lissage et $\mathrm{K} : \mathbb{R}^p \mapsto \mathbb{R}^+$ est un noyau i.e. une fonction symétrique, à valeurs positives ou nulles et d'intégrale 1.

Exemples de noyaux :
— le noyau $K(u) = \frac{1}{2}\ \mathbf{1}_{[-1,1]}(u)$ revient à estimer une densité par un histogramme.
— le noyau gaussien $K(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2}$

Le choix du noyau n'est pas le plus important en réalité. C'est surtout le choix du paramètre de lissage h qui importe. Les noyaux interviennent dans les SVM donc vous en reparlerez l'an prochain.

Dans le contexte de l'analyse discriminante, on estime donc pour chaque groupe k la densité f_k des variables explicatives par un noyau (il existe des noyaux multi-dimensionnel) et on utilise ensuite la régle du MAP comme pour les modèles de mélange gaussiens.

3.6 Estimateur des plus proches voisins

Voici une autre façon de définir un voisinage autour de x : on calcule les distances entre x et les individus du jeu d'apprentissage, et on garde les h individus les plus proches.

Il faut donc
- — choisir une distance entre individus. En général on prend la distance de Mahalanobis.
- — choisir une valeur de h. C'est la difficulté de cette méthode.

On estime alors la densité pour la classe k $f_k(x)$ par

$$\hat{f}_k(x) = \frac{N_k^h}{nV_x}$$

où N_k^h est le nombre d'individus de la classe k parmi les h plus proches voisins de x, et V_x est le volume de la boule autour de x.

On utilise ensuite la règle du MAP pour affecter le nouvel individu. Comme le volume V_x ne dépend pas de la classe k, la régle du MAP revient à classer x dans la classe qui réalise le maximum de $p_k N_k^h$ (dans le cas équiprobable, c'est donc la classe majoritaire parmi ses voisins).
La méthode des knn est simple mais coûteuse en place mémoire (calculer les distances de x à tous les voisins).

4 Arbre de Classification

C'est une approche non paramétrique qui donne lieu à une partition des données qui est présentée sous forme d'arbre et facile à interpréter.
Référence : Méthode CART : Classification and regression trees. L. Breiman, J. H. Friedman, R.A. Olshen, and C. J. Stone, Chapman Hall, 1984.

Dans ce cours on s'intéresse à de la classification ie avec la variable à expliquer Y qui est qualitative. On parle alors d'arbre de classification. On peut adapter ce que l'on va voir dans le cas d'une variable à expliquer quantitative. On parle alors d'abre de régression et de façon générale on parle d'arbre de décision.
Les variables explicatives peuvent dans les arbres de décision, être quantitatives et/ou qualitatives.

Exemple d'arbre de classification : Données d'après Philippe Preux (Lille 3)

```
> head(tennis)
           Ciel Temperature Humidite   Vent Jouer
1    Ensoleille        27.5       85 Faible   Non
2    Ensoleille        25.0       90   Fort   Non
3       Couvert        26.5       86 Faible   Oui
4         Pluie        20.0       96 Faible   Oui
5         Pluie        19.0       80 Faible   Oui
6         Pluie        17.5       70   Fort   Non
```

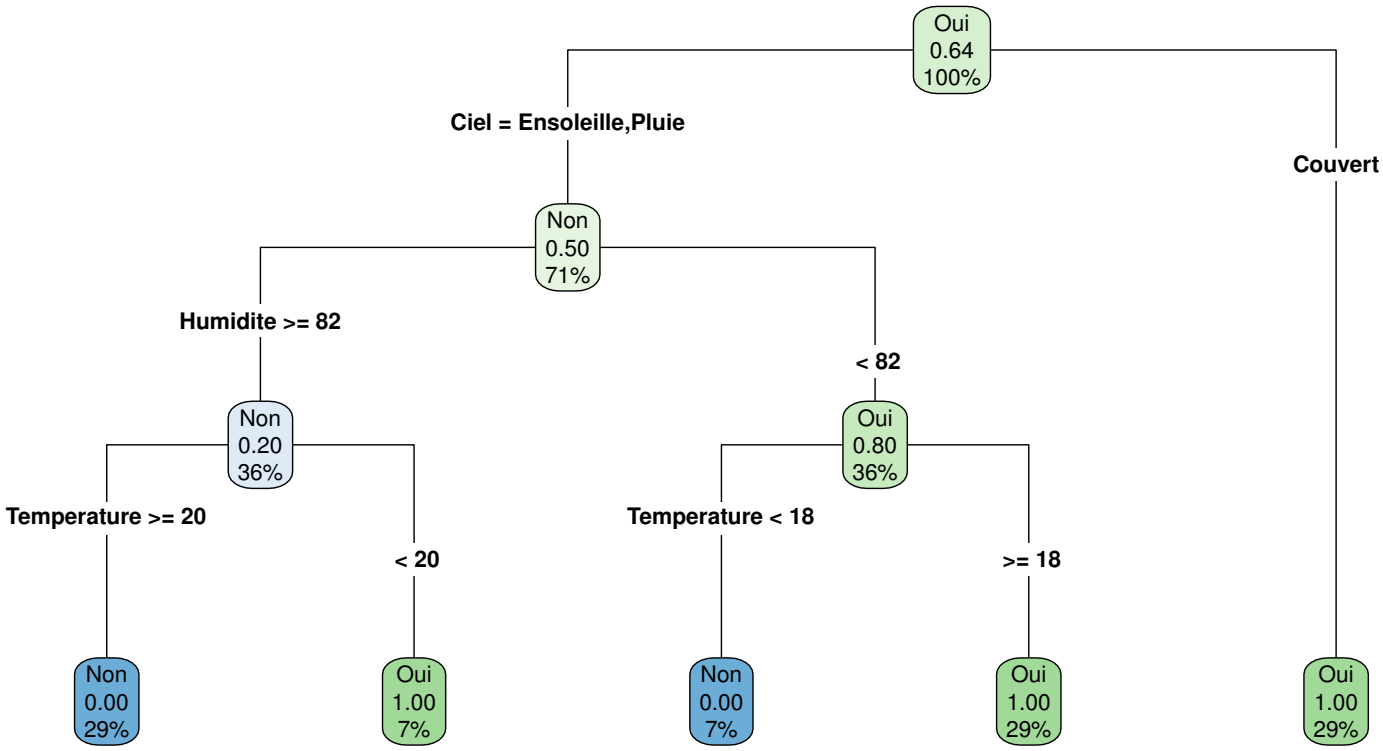


FIGURE 3 – Arbre de Classification

Construire un tel arbre consiste à définir une suite de noeuds, chaque noeud permettant de faire une partition des données en 2 groupes sur la base de l'une des variables explicatives.

4.1 Construction de l'arbre maximal

Principe : La méthode CART construit un arbre dit binaire (car 2 noeuds fils) :

1. Le noeud racine contient les n données.
2. A chaque étape, un noeud est divisé pour construire deux nouveaux noeuds ce qui conduit à une partition des données. Comment ? On va chercher une variable explicative, et une coupure de façon à "réduire de façon optimale le désordre des données".
3. L'arbre maximal est obtenu lorsqu'aucun noeud ne peut plus être divisé. Un noeud terminal (qui ne peut plus être divisé) est appelé une feuille.
4. Chaque feuille est alors associée à l'une des classes de la variable à expliquer.

Impureté d'un noeud :
On veut diviser un noeud t en deux sous-noeuds t_L (Left) et t_R (Right) qui soient les plus homogènes possibles (autrement dit les plus purs possibles) au sens de la variable à expliquer. Comment mesurer l'hétérogénéité ou impureté d'un noeud ?
Idée : on dira qu'un noeud est pur si toutes les observations du noeud appartiennent à la même classe, et que l'impureté d'un noeud est maximale lorsque toutes les classes de Y sont équiprobables dans le noeud.

Une mesure standard de l'impureté d'un noeud est donné par lindice de Gini :
On note $\hat{p}_k(t)$ la proportion d'observations Y_i de classe k dans le noeud t. Ainsi $\hat{p}_k(t)$ estime la probabilité d'appartenance à la classe k pour le noeud t (cf probabilité a priori). L'indice de Gini est donné par :

$$i(t) = \sum_{k=1}^{K} \hat{p}_k(t)(1 - \hat{p}_k(t))$$

Dans le cas de $K = 2$ classes, comme $\hat{p}_1(t) + \hat{p}_2(t) = 1$, on a $i(t) = 2\hat{p}_1(t)(1 - \hat{p}_1(t))$ et l'indice de Gini est maximum pour $\hat{p}_1(t) = 1/2$:

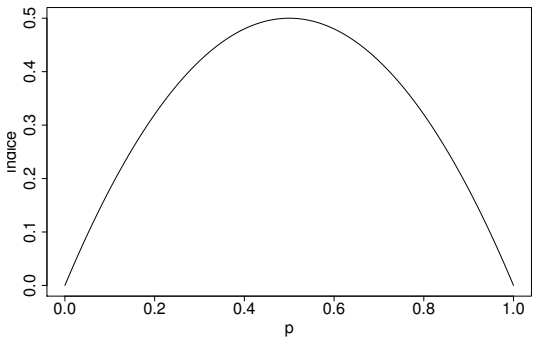


FIGURE 4 – Indice de Gini dans le cas de 2 classes

Lors de la division d'un noeud on cherche à réduire l'impureté du noeud t, ie on cherche la division (t_L, t_R) telle que

$$i(t) - \left(\frac{|t_L|}{|t|} i(t_L) + \frac{|t_R|}{|t|} i(t_R) \right)$$

soit maximale. Les proportions $\frac{|t_L|}{|t|}$ et $\frac{|t_R|}{|t|}$ sont respectivement les proportions de données du noeud t partant dans les noeuds t_L et t_R.

Comment diviser un noeud :
L'algorithme CART consiste à choisir parmi toutes les divisions (t_L, t_R) possibles celle qui maximise la réduction de l'impureté. Les divisions (t_L, t_R) d'un noeud t sont induites par une variable explicative et une question binaire (cf 2 noeuds fils) :

- Si la variable explicative X^j est quantitative, on lui associe un seuil de coupure c et :
  si $X_i^j \geq c$ alors l'observation i va dans le noeud fils gauche ;
  si $X_i^j < c$ alors l'observation i va dans le noeud fils droit ;
- Si la variable explicative X^j est qualitative, on lui associe une partition en deux groupes $\{A, \bar{A}\}$:
  si $X_i^j \in A$ alors l'observation i va dans le noeud fils gauche ;
  si $X_i^j \notin A$ alors l'observation i va dans le noeud fils droit ;

Diviser un noeud revient donc à choisir la meilleur variable X^j et la meilleure coupure (seuil c ou modalité(s) A).

4.2 Règle d'arret

Le critère d'arrêt consiste à ne pas découper un noeud pur. Afin d'éviter un découpage trop fin, on décide souvent de ne pas découper pas un noeud qui contient moins de n_{min} données avec souvent n_{min} compris entre 1 et 5.

4.3 Associer une classe à une feuille

A chaque noeud terminal (ie feuille), on associe la classe majoritaire des observations :

$$k_t = arg \max_{k=1,K} \hat{p}_k(t)$$

Pour une nouvelle observation x dont on veut prévoir la classe, la règle de classification consiste à faire "descendre" l'observation x jusqu'à une feuille (au vues des variables explicatives X pour cette observation) et à prédire par la classe associé à cette feuille. Ainsi la classe prédite pour l'observation x est :

$$k^*(x) = k_{\text{feuille}(x)}$$

avec feuille(x) la feuille dans laquelle tombe l'observation x
Exemple : pour la première observation du jeu de données tennis, on prédira comme classe : NON.

4.4 Elagage

Dans un second temps, l'arbre maximal est élagué afin d'éviter le sur-apprentissage et obtenir un modèle plus parcimonieux. On peut en effet voir l'arbre maximal comme ayant besoin de sélection de modèle. L'arbre optimal sera alors un sous-arbre élagué (sans séparer deux fils) minimisant un critère pénalisé mesurant un compromis entre la taille de l'arbre et le taux de mauvaises classifications :

$$crit_\alpha(T) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{k^*(X_i)\neq Y_i} + \alpha\frac{|T|}{n}$$

où $|T|$ est le nombre de feuilles T et $\sum_{i=1}^{n} \mathbf{1}_{k^*(X_i)\neq Y_i}$ est le nombre de mal classés de T.
Si $\alpha = 0$ le meilleur arbre est l'arbre maximal. Grace au terme de pénalité, on pénalise les "grands" arbres" : on va élaguer un arbre jusqu'à ce la division suivante devienne superflue.
Il s'agit alors de trouver la meilleur valeur de α, ce qui se fait par validation croisée.

4.5 Limites des arbres de classification

1. Partitions binaires ? On pourrait envisager un partitionnement en plus de deux noeuds, mais en pratique cela s'avère moins bon car les données sont trop vite fragmentées.
2. Instabilité : c'est le principal inconvénient. Un faible changement dans les données peut donner un arbre très différent. Solution : les forêts aléatoires.

4.6 Forêts aléatoires

Les forêts aléatoires ont été introduites as Breiman. Il y a plusieurs algorithmes possibles.

1. Bagging (Breiman 1996) L'idée est d'améliorer les performances de Cart, en introduisant de la variabilité grâce au boostrap.

Pour une nouvelle observation x, on prédit la classe par aggrégation, ie :

$$k^*_{agg}(x) = \arg \max_{k=1,\dots,K} \sum_{b=1}^{B} \mathbf{1}_{k^*_b(x)=k}$$

où $k^*_b(x)$ est la classe prédite pour l'observation x dans l'arbre numéro b. Ainsi x est prédite par la classe la plus prédite parmi les B arbres CART.

2. Random Forest (Breiman 2001). C'est ce qui est codé dans le package R *randomForest*

Au lieu d'un arbre cart, on construit un arbre RI (Random Inputs). Ce qui change :

- pas d'élagage
- un arbre RI consiste à tirer aléatoirement à chaque noeud *mtry* variables, puis à chercher la meilleure coupure uniquement parmi les variables sélectionnées. *n "agite localement les données".*
  Le paramètre *mtry* est fixé par défaut à $\sqrt{p}$ en classification dans R avec p le nombre initial de variables. L'idée est qu'il faut *mtry* suffisamment grand pour que la probabilité de tirer une variable importante soit grande.

Avantage des Random Forest :

— L'aléa supplémentaire dû à *mtry* améliore les performances

— Lorsque p est grand, on gagne en temps de calcul car plus d'élagage et moins de variables en jeu à chaque noeud.

5 La régression logistique

La régression logistique est un modèle permettant d'expliquer et de prédire une variable qualitative Y à l'aide de variables quantitatives et/ou qualitatives.

Exemple. Etudier les facteurs de risque de survenue d'une maladie cardio-vasculaire (âge, sexe, tabac, ...)

5.1 Le modèle logistique

Dans ce cours, on se restreint au cas où Y n'a que deux classes (on dit qu'Y est dichotomique ou binaire). On notera ces deux valeurs : 1 ou 0 (qu'on appelera succès ou échec).
Pour une variable Y à $K > 2$ modalités, vous trouverez ici des cours très complets :
`https://eric.univ-lyon2.fr/~ricco/cours/cours/pratique_regression_logistique.pdf`
ou
`https://perso.univ-rennes2.fr/system/files/users/rouviere_l/poly_logistique_web.pdf`

5.1.1 Comment modéliser une variable Y binaire ?

Supposons que l'on dispose d'une seule variable explicative X.

Exemple. Risque de survenue d'une maladie cardio-vasculaire (chd) en fonction de l'âge. Pour n patients, on dispose de leur âge et on sait si oui, ou non ils ont une chd.

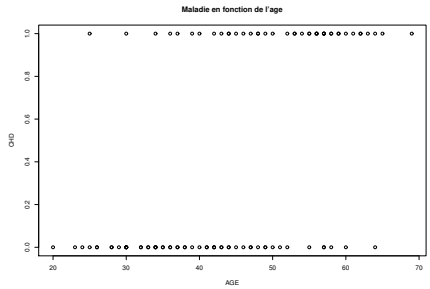


FIGURE 5 – présence ou absence d'une chd en fonction de l'âge des n patients de l'échantillon

Comment modéliser ces données ?
Si on regoupe les patients par classe d'âge et que l'on représente la proportion de chd par classe d'âge, voilà ce que l'on obtient :

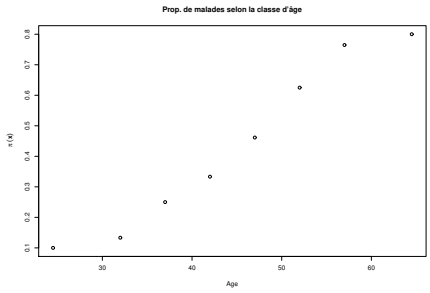


FIGURE 6 – proportion de chd par classe d'âge

Il apparaît que lorsque l'âge augmente, la proportion d'individus atteint par la maladie augmente.

Petit retour vers la régression simple

Dans le cas de la régression linéaire simple, si l'on considère que la variable explicative est fixe (i.e non aléatoire) le modèle s'écrit :

$$Y_i = a + bx_i + \epsilon_i$$

avec ϵ_i i.i.d $\sim \mathcal{N}(0, \sigma^2)$. Ainsi Y_i est gaussien d'espérance $E(Y_i) = a + bx_i$. C'est le cas étudié cette année avec par exemple x_i la quantité de blé semée sur la parcelle i. On parle de modèle de régression à design fixe.

Si l'on considère que la variable explicative X est aléatoire et que x_i en est une réalisation, alors on parle de modèle à design alétoire et le modèle revient à dire que le loi de Y conditionnellement à $X = x_i$ est gaussienne, d'espérance

$$E(Y|X = x_i) = a + bx_i$$

En régression logistique, comme Y peut prendre 2 valeurs, on va modéliser la loi de Y conditionnellement à $X = x$ par une loi de Bernoulli. De quelle paramètre ? Notons $\pi(x) = \mathbb{P}(Y = 1|X = x)$ son paramètre (c'est ce que l'on a déjà appelé probabilité à posteriori).

Comme $Y|X = x$ suit une loi de Bernoulli, on a $E(Y|X = x) = P(Y = 1|X = x)$. Mais ca n'a pas de sens de modéliser cette espérance par une droite (comme en régression simple) puisqu'une probabilité doit être comprise entre 0 et 1, alors qu'une droite n'est pas bornée.

Sur le graphe de la proportion en fonction de l'âge, on observe une relation dite sigmoidale (en forme de S). Ce phénomène s'observe dans de nombreuses situations où l'utilisation d'un modèle logistique est justifié. On va donc modéliser $\mathbb{P}(Y = 1|X = x)$ par la fonction sigmoide :

$$\pi(x) = \frac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}}$$

avec β_0 et β_1 des paramètres à estimer à partir des données. La fonction sigmoide, elle, a bien ses valeurs $\in [0, 1]$.

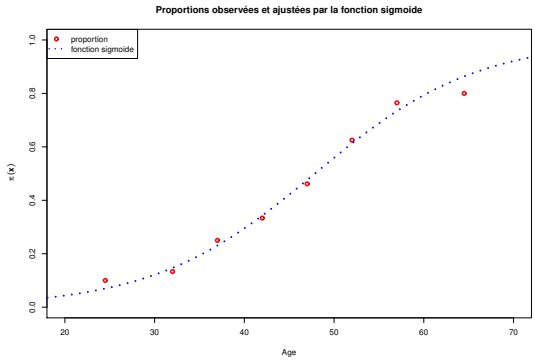


FIGURE 7 – présence ou absence d'une chd en fonction de l'âge des n patients de l'échantillon

5.1.2 Le modèle logistique simple

Définition. On appelle *logit* la fonction bijective suivante :

$$\begin{aligned} logit :]0, 1[&\rightarrow \mathbb{R} \\ p &\rightarrow log(\frac{p}{1-p}) \end{aligned}$$

La fonction logit est une transformation qui permet d'avoir des valeurs dans $\mathbb{R}$.
La fonction logit a pour inverse $logit^{-1}(y) = \frac{e^y}{1+e^y} = \frac{1}{1+e^{-y}}$.

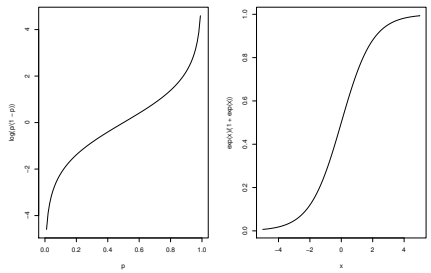


FIGURE 8 – a gauche : fonction logit, à droite : son inverse la fonction sigmoide

Définition. On note $\pi(x) = \mathbb{P}(Y = 1|X = x)$ la probabilité à posteriori. Le modèle logistique simple s'écrit :

$$\pi(x) = \frac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}}$$

ou encore

$$logit(\pi(x)) = \beta_0 + \beta_1 x$$

Ainsi on suppose que la fonction $logit(\pi(x))$ peut etre modéliser par une droite.

Remarques. Pour aller plus loin...

Dans le modèle logistique, on modélise la loi de Y conditionnellement à $X = x$:

1. par une loi de Bernoulli de paramètre $\pi(x)$
2. et on choisit $\pi(x)$ tel que $logit(\pi(x)) = \beta_0 + \beta_1 x$

Au lieu d'utiliser la fonction logit pour modéliser $\pi(x)$, on peut utiliser la fonction probit, basée sur l'inverse de la fonction de répartition d'une loi normale. Mais dans la pratique, cela donne des résultats peu différents (les graphes logit et probit sont relativement proches) et le modèle logistique permet une interprétation facile en terme d'odds ratio (voir section suivante) et est donc plus souvent utilisé.

Au lieu de modéliser la loi de Y conditionnellement à $X = x$ par une loi de Bernoulli, on peut choisir d'autres lois, ce qui donne d'autres modèles. Par exemple, on peut utiliser une loi de Poisson dans le cas où Y est une variable de comptage (on parle alors de modèle log-linéaire), ou utiliser une loi normale (ce qui revient à un modèle de régression à design aléatoire). Tout ces modèles font partie de ce que l'on appelle le modèle linéaire généralisé GLM.

Remarque. Lien avec le modèle probabiliste et les approches LDA et QDA.

Dans le modèle probabiliste, on a modélisé les lois de $X|Y = k$ pour avoir l'expression de $\mathbb{P}(Y = k|X = x)$ avec la formule de Bayes. En régression logistique, on modélise directement

$$logit(\pi(x)) = log(\frac{\pi(x)}{1 - \pi(x)}) = log(\frac{\mathbb{P}(Y = 1|X = x)}{\mathbb{P}(Y = 0|X = x)})$$

C'est pourquoi on parle en général de modèle "semi-paramétrique".
De plus, comme en régression logistique, on ne fait pas d'hypothèses sur la loi de $X|Y = k$, la régression logistique est moins restrictive.

Quelle méthode choisir ? Si les hypothèses de la LDA ou QDA sont adaptées aux données étudiées, l'approche probabiliste dispose de plus d'information que la régression logistique pour estimer les paramètres du modèle. Mais sinon, la régression logistique est plus souple. En pratique, on peut comparer les performances des deux méthodes en étudiant leur erreurs de classement respectifs (voir dernière section)

5.1.3 Le modèle logistique multiple

On peut généraliser le modèle logistique simple au cas où on a p régresseurs, ie. $X = (X^1, \ldots, X^p)$.

Définition. Le modèle logistique multiple s'écrit

$$\pi(x) = \mathbb{P}(Y = 1|X = x) = \frac{e^{\beta_0+\beta_1 x_1+\ldots+\beta_p x_p}}{1 + e^{\beta_0+\beta_1 x_1+\ldots+\beta_p x_p}}$$

ou encore

$$logit(\pi(x)) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$$

Exemple. Survenue d'une chd en fonction de 2 variables : l'âge (quantitative) et le sexe (qualitative binaire).

On propose un modèle additif (sans interaction entre l'âge et le sexe) :

$$logit(\pi(x)) = \beta_0 + \beta_1 age + \beta_2 sexe$$

Remarque. Comme en régression multiple, on peut modéliser des termes d'intéraction entre variables explicatives.

5.2 Interprétation des coefficients du modèle en terme d'odds-ratio

Dans certains domaines, l'explication est bien plus importante que la prédiction. Par exemple, dans le domaine de la santé, on cherche certes à détecter automatiquement les personnes qui développent une maladie particulière, mais il est peut être plus important que l'on comprenne pourquoi ils la développent pour qu'on puisse l'anticiper.

Définition. On appelle "côte" de la variable Y le rapport $\frac{\mathbb{P}(Y=1|X=x)}{\mathbb{P}(Y=0|X=x)} = \frac{\mathbb{P}(Y=1|X=x)}{1-\mathbb{P}(Y=1|X=x)}$.
Si la côte est supérieure à 1, alors un succès est plus probable qu'un échec.

Définition. On appelle Odds-ratio le rapport des côtes obtenus lorsque la variable explicative est modifiée.

— Si X est qualitative binaire ($X = 1$ ou 0)

$$OR = \frac{\mathbb{P}(Y=1|X=1)}{1-\mathbb{P}(Y=1|X=1)} / \frac{\mathbb{P}(Y=1|X=0)}{1-\mathbb{P}(Y=1|X=0)}$$

L'OR mesure donc l'évolution de la côte lorsque X passe de 0 à 1.

— Si X est qualitative avec plus de 2 modalités, une modalité est prise comme référence et plusieurs OR mesurent l'évolution de la côte lorsque l'on passe d'une modalité à une autre.

— Si X est quantitative : on considère des odds-ratio pour un accroissement unitaire de la variable X :

$$OR = \frac{\mathbb{P}(Y=1|X=x+1)}{1-\mathbb{P}(Y=1|X=x+1)} / \frac{\mathbb{P}(Y=1|X=x)}{1-\mathbb{P}(Y=1|X=x)}$$

L'OR mesure donc l'évolution de la côte lorsque X passe de x à $x+1$ unités. L'OR dépend donc de l'unité de la variable X. Dans notre exemple sur les chd, l'odds-ratio mesure l'évolution du risque de survenue d'une chd lorsque l'âge accroit d'une unité.

Cas d'un modèle logistique simple :

Propriété 8. Dans le modèle de régression logistique simple,

$$OR = e^{\beta_1}$$

β_1 s'interprète donc à partir d'un odds-ratio. En particulier si $\beta_1 > 0$, la courbe est croissante et l'âge augmente le risque de MCV : on a $OR > 1$. A l'inverse si $\beta_1 < 0$, la courbe est décroissante et l'âge diminue le risque de MCV : on a $OR < 1$.

Remarques.

— A partir de l'estimation de β_1 (voir section suivante), on calcule l'OR et on peut l'interpréter.
Dans notre exemple sur les chd, l'odds-ratio mesure l'évolution du risque de survenue d'une chd lorsque l'âge accroit d'une unité. Par exemple si l'âge est mesurée en dizaine d'année, et que $OR = 2$, cela signifie que "10 ans de plus" multiplie par 2 le risque de maladie.

— Attention : il semble évident que de passer de 20 à 30 ans n'a pas la même signification que de passer de 40 à 50 ans. Or nous avons supposé que le logit de $\pi(x)$ est linéaire par rapport à la variable explicative. C'est une hypothèse un peu forte. Il faut donc vérifier cette hypothèse de linéarité avant d'utiliser le modèle et manipuler les OR avec prudence.

— Pour une variable explicative quantiative, OR mesure l'évolution de la côte lorsque X passe de x à $x+1$ unités. Si l'on augmente de u unités la variable explicative, l'odds-ratio devient alors $e^{u\times\beta_1}$.

— Si OR=1, il y a indépendance entre Y et X. Ici cela voudrait dire que l'âge n'est pas un facteur de risque d'une chd. On peut constuire un test de $H_0 : OR = 1$, ce qui revient à tester $H_0 : \beta_1 = 0$ (voir section test).

Cas d'un modèle logistique multiple
On parle d'OR partiel lorsqu'on considère l'odds-ratio pour un accroissement d'une variable tout en laissant fixes les autres variables ("toutes choses égales par ailleurs").

Exemple. Dans le modèle $logit(\pi(x)) = \beta_0 + \beta_1 age + \beta_2 sexe$, on a

$$OR_{age} = e^{\beta_1}; OR_{sexe} = e^{\beta_2}$$

On peut donc interpréter l'effet de l'âge pour un sexe donné, ou l'effet du sexe sur la chd à un âge donné.

Remarque. L'interprétation des odds-ratio est difficile pour les modèles avec interaction. Ils sont plus ou moins liés avec les coeffcients de la régression, nous devons tenir compte des valeurs prises par les autres explicatives.

5.3 Estimateurs des paramètres

Les variables aléatoires $Y|X = x$ suivent une loi de Bernoulli de paramètre $\pi(x)$ avec, par exemple dans le modèle logistique simple, $\pi(x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}$ où les paramètres β_0 et β_1 sont inconnus.
On cherche à estimer les paramètres inconnus du modèle. Ecrivons la vraisemblance :

$$L = \Pi_{i=1}^n \mathbb{P}(Y = y_i|X = x_i) = \Pi_{i=1}^n \pi(x_i)^{y_i}(1 - \pi(x_i))^{1-y_i}$$

La méthode de maximum de vraisemblance ne permet pas d'obtenir de solution analytique et les estimateurs de maximum de vraisemblance sont obtenues à l'aide d'un algorithme d'optimisation (Newton-Raphson par exemple).

5.4 Intervalle de confiance et tests dans le modèle logistique

Comme en régression, on peut vouloir effectuer l'un des tests suivants :
- tester l'apport d'au moins une des variables explicatives.
- tester la contribution individuelle d'une variable
- tester la contribution d'un bloc de plusieurs variables.

Ils existent pour cela deux types de tests :

1. Le test basé sur le rapport de vraisemblance. Il consiste à comparer les vraisemblances des modèles sous H_0 et sous H_1. C'est une procédure très courante en statistique qui est utilisée dans d'autres modèles qu'en régression logistique. En particulier, le test de student (t.test) est un test basé sur le rapport de vraisemblance, même s'il n'a pas été présenté tel quel. La statistique de test est

$$\text{LR} = -2\ln(L_0/L_1) = -2\ln(L_0) - (-2\ln(L_1)) := D_0 - D_1$$

où L_0 et L_1 sont les vraisemblances des modèles M_0 (modèle sous H_0) et M_1 (modèle sous H_1). D_0 et D_1 sont les déviances des modèles : elles sont définies par -2 fois la log-vraisemblance du modèle.
On rejette H_0 si LR est suffisamment grande, ce qui signifie que le modèle sous H_1 est plus vraisemblable que celui sous H_0. Le seuil à partir duquel on rejette dépend de la dimension des deux modèles. En effet on peut montrer que sous H_0, LR suit asymptotiquement (pour n grand) une loi du $\mathcal{X}^2(\dim(M_1)-\dim(M_0))$.
$\rightarrow$ Sous R, ce test permet de tester l'apport d'au moins une des variables explicatives, et la contribution d'un bloc de plusieurs variables. C'est aussi le test mis en oeuvre pour faire de la selection de modèle automatique (voir TP).
2. Le test de Wald basé sur la normalité asymptotiques des estimateurs. C'est un test analogue au test de Student en régression. Pour tester l'apport d'une variable X^j, le test de Wald s'appuie sur la statistique :

$$W_j = \frac{\hat{\beta}_j}{\sqrt{\widehat{var}(\hat{\beta}_j)}} \underset{H_0}{\sim} \mathcal{N}(0,1) \text{ asymptotiquement}$$

La statistique W_j permet également de construire des intervalles de confiance pour les paramètres β_j du modèle. Notons que $W_j^2 \underset{H_0}{\sim} \mathcal{X}^2(1)$ asymptotiquement.

$\rightarrow$ Sous R, ce test permet de tester la contribution individuelle d'une variable

5.5 Prédiction avec le modèle

Une fois les paramètres du modèle estimées, on dispose de l'estimation des probabilités à posteriori $\widehat{\pi(x)}$. Par exemple dans le modèle logistique simple :$\widehat{\pi(x_i)} = logit^{-1}(\widehat{\beta_0} + \widehat{\beta_1} x_i)$.

Supposons un nouvel individu, pour lequel on observe la (ou les) variable explicative x. On applique la régle du MAP et on affecte l'individu à la classe la plus probable. En général on utilise la régle suivante :

$$\begin{aligned} \text{Si } \widehat{\pi(x)} > 1/2 \quad &: \quad \hat{Y} = 1 \text{ i.e le nouvel individu est affecté dans la classe 1} \\ \text{Sinon} \quad &: \quad \hat{Y} = 1 \text{ i.e le nouvel individu est affecté dans la classe 0} \end{aligned}$$

On peut bien sûr changer le seuil 1/2. Plus $\widehat{\pi(x)}$ est proche de 1 ou de 0, plus la prédiction est sûre. Donc on peut adapter le seuil en fonction du domaine d'application (est-ce-qu'une mauvaise affectation est dramatique ?)

On peut de plus construire un intervalle de confiance de la probabilité à posteriori $\pi(x)$. En effet, comme en régression simple, on peut construire un intervalle de prédiction de $logit(\pi(x)) = \beta_0 + \beta_1 x$. En inversant la fonction logit on en déduit un intervalle de prédiction de $\pi(x)$.

6 Evaluation de la qualité d'une méthode discriminante

6.1 Sur quelles données évaluer la méthode ?

Pour évaluer la qualité d'une méthode discriminante, on peut mesurer le taux d'erreur de classement. L'échantillon d'apprentissage a permis de construire une règle d'affectation. Si on teste cette règle d'affectation sur l'échantillon d'apprentissage, comme on utilise les mêmes observations pour construire la règle et la valider, on va sous-estimer le taux d'erreur de classement (on a un biais d'"optimisme").

Validation simple

Une partie des données (de l'ordre de 80%) sert d'échantillon d'apprentissage et le reste des données (20%) sert d'échantillon-test.

Cependant, il faut suffisamment de données. Si n est petit, couper l'échantillon en 2 implique avoir un trop petit échantillon d'apprentissage pour espérer avoir un modèle de bonne qualité, et/ou un trop petit échantillon test pour avoir une évaluation fiable de la méthode.

Validation croisée

Lorsque les effectifs sont faibles, nous avons intérêt à construire le modèle sur la totalité des données, puis à utiliser des techniques de ré-échantillonnage pour en mesurer les performances L'intérêt est double. Nous utilisons la totalité des données pour construire le classiffieur. Et nous pouvons obtenir une évaluation (plus ou moins) faiblement biaisée de son erreur de prédiction.

- Lorsque n est très faible, on peut utiliser la méthode du "leave one out" : pour chaque observation i on considère l'échantillon d'apprentissage constitué en éliminant la i-ème observation. Puis l'observation i sert d'échantillon test.
- Pour des échantillons plus grands, la méthode du "leave one out" devient trop lourde en temps de calcul et on lui préfère la validation croisée K-fold. (Attention, ce K n'a rien à voir avec le nombre de classes !)
  - Découper aléatoirement l'échantillon en K parties de tailles égales
  - pour $k = 1$ à K, on met la k-ème partie des observations de côté pour former l'échantillon test et on utiliser le reste des observations comme échantillon d'apprentissage.

6.2 Indicateurs de qualité

Taux d'erreur en analyse discriminante

Dans le cas d'une variable Y à deux classes, on peut construire la matrice de confusion :

	observation		total
	$Y=1$	$Y=0$	
prédision $\hat{Y}=1$	TP	FP	$\hat{P}$
$\hat{Y}=0$	FN	TN	$\hat{N}$
total	P	N	n

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

A partir de ce tableau, on peut calculer différents indicateurs, dont :

- Le taux de bon classement aussi appelé accuracy $= \frac{TP+TN}{n}$
- Le taux de vrais positifs ou sensibilité $= \frac{TP}{P}$
- Le taux de vrais négatifs ou spécificité $= \frac{TN}{N}$
  $\rightarrow$ Une bonne régle de classification doit avoir une grande sensibilité et une grande spécificité.
- Le FDR (false discovery rate) : $= \frac{FP}{\hat{P}}$.
  C'est un indicateur surtout utilisée dans la théorie des test mutliples, puisqu'il regarde le nombre de détection à tort ($\hat{Y}=1$ alors que $Y=0$) sur le nombre de détection ($|\hat{Y}=1| = \hat{P}$)

Courbe ROC

On classe un individu dans la classe 1 si $\widehat{\pi(x)} > 1/2$. Modifions le seuil 1/2, et classons un individu dans la classe 1 si $\widehat{\pi(x)} > s$. L'idée de la courbe ROC est de faire varier ce seuil. Si s augmente, alors on aura moins d'individu prédit dans la classe 1, et donc la sensibilité va diminuer et la spécificité augmenter.

La courbe ROC représente l'évolution de la sensibilité en fonction de 1 - spécificité quand on fait varier le seuil s. C'est une courbe croissante entre le point $(0,0)$ et le point $(1,1)$. Meilleure est la prédiction, plus la courbe est "haute", alors qu'une prédiction random (jouer à pile ou face) donnerait la première bissectrice.

L'aire sous la courbe ROC (AUC, Area Under the Curve) donne un indicateur de la qualité de la prédiction (1 pour une prédiction idéale, 0.5 pour une prédiction random).