Projet

Nadia GHERNAOUT
Philippine RENAUDIN

Table des matières

1 Introduction générale

L'idée générale du scoring est d'affecter une note (un score) globale à un individu à partir de plusieurs descripteurs, quantitatifs ou qualitatifs. À partir de cette note, on affecte l'individu à un groupe préexistant. Un score peut donc être définit comme un outil statistique ou probabiliste de détection de risque. Le scoring peut également être vu comme l'application au monde de l'entreprise de plusieurs techniques de classement. Nous en aborderons 2 dans ce rapport.

Nous pouvons déjà citer plusieurs types de score :

1. Les scores de risque :
   — risque de crédit ou credit scoring : prédire le retard de remboursement de crédit.
   — risque financier : prédire la bonne ou mauvaise santé d'une entreprise.
   — risque médical : prédire l'apparition d'une maladie chez un patient.
2. Les scores en marketing :
   — score d'attrition : 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.

La création d'un score se fait en fonction des objectifs recherchés et des moyens techniques disponibles. Par exemple, le développement d'un score comportemental nécessite de disposer de données sur au moins un an, si l'on a moins d'historique, il vaut mieux partir sur un score générique ou un score d'octroi.

Il faut aussi également choisir l'utilisation qui sera faite du score : outil d'aide à la décision ou outil de ciblage pour le marketing direct par exemple. C'est en fonction de l'utilisation que l'on en fera que la règle de décision sera ajustée.

Pour construire un score, il faut dans un premier temps disposer d'un échantillon suffisamment conséquent pour pouvoir tester plusieurs modèles prédictifs. De plus, pour éviter des problèmes de surestimation de la qualité du modèle, il est préférable de séparer l'échantillon d'étude en deux sous-échantillons : un échantillon d'apprentissage à partir duquel sera créé le modèle, et un échantillon test sur lequel sera testé la qualité du modèle par rapport à l'objectif recherché et au risque que l'on est prêt à prendre.

Ensuite, il faut élaborer un modèle prédictif à l'aide de techniques prédictives : analyse discriminante et régression logistique en l'occurrence.

Enfin, les notes de score sont découpées en plusieurs classes de valeur. Dans le domaine financier, on aura tendance à découper les notes de score en trois classes : faible, moyen, fortes. Dans le milieu médical, on préfèrera 2 classes : à risque, non à risque. La règle de classement (seuil comparatif du score) se décide en fonction du risque d'erreur que l'on souhaite prendre.

Nous présentons dans ce rapport 2 des techniques prédictives les plus utilisées en scoring : la régression logistique et l'analyse discriminante. Pour illustrer ce qu'est le scoring, nous avons utilisé ces 2 techniques sur 2 jeux de données différents. Nous présentons dans la suite la théorie de chaque technique ainsi que l'étude des données associée.

Première partie
Protocole

1 Validation croisée

Comme il a été mentionné en introduction, une habitude à prendre lors de toute analyse de données est de séparer les données en plusieurs sous échantillons pour éviter les problèmes de surestimation des capacités du modèle.

On appelle validation croisée la technique consistant à ajuster un modèle prédictif sur un échantillon d'apprentissage et à valider ce modèle sur un échantillon test. Ces échantillons peuvent provenir du même jeu de données auquel cas il est coutume que l'échantillon d'apprentissage représente entre 60% et 80% des données et que l'échantillon test représente 20% à 40%. Il est également possible, si l'on dispose de plusieurs jeux de données différents pour le sujet d'étude, de prendre un jeu de données comme échantillon d'apprentissage et de valider le modèle sur un deuxième jeu de données.

Nous avons choisi dans nos études de cas de réaliser une validation croisée à partir d'un seul jeu de données en le décomposant en un échantillon d'apprentissage représentant 80% du jeu de données et en un échantillon test représentant les 20% restant, et ce de manière aléatoire.

2 Sélection de modèles

Une fois le jeu de données séparé en deux échantillons, vient le moment de construire différents modèles prédictifs. Cependant, il n'est pas toujours évident de savoir quelles variables garder, quelles sont celles qui apportent le plus d'information, qui discriminent le mieux les groupes d'individus, etc... C'est pourquoi on s'appuie sur différents indicateurs, en plus de ceux implémentés par défaut dans les logiciels. Nous en évoquons 3 ici :

2.1 Les critères AIC et BIC

2.2 La déviance

Nous verrons plus tard dans ce rapport que les paramètres des différents modèles sont le plus souvent estimés par la méthode du maximum de vraisemblance.

Pour savoir quels modèles garder, il est donc courant d'utiliser le critère de la déviance. En effet, la déviance est égale à -2log-vraisemblance donc le modèle maximisant la vraisemblance est celui minimisant la déviance. On utilise plutôt cet indicateur car plus simple à calculer que les expressions du maximum de vraisemblance.

Il existe des fonctions dans Rstudio rendant le /les modèles minimisant la déviance, il n'est donc pas nécessaire de créer les modèles au préalable et de les comparer entre eux, le logiciel fait ce travail à notre place.

2.3 Le lambda de Wilks

Cet indicateur est propre à l'analyse discriminante et n'est pas utilisé en régression logistique. Le Lambda de Wilks est souvent utilisé dans les logiciels comme critère pour ne garder que les variables apportant de l'information sur l'appartenance ou non d'un individu à un groupe.

Le Lambda de Wilks est une approche paramétrique permettant de tester si plusieurs variables continues distinctes $X = (X_1, \cdots, X_p)$ sont liées à une variable qualitative Y à $K \geq 2$ groupes, lorsqu'elles sont consi-

dérées avec leurs différentes interactions multivariées.

Les hypothèses d'utilisation de ce test sont : $X|_{Y=1}, \cdots, X|_{Y=k}$ suivent une loi normale et leur matrice de covariance respective sont égales (homoscédasticité).

La statistique du test du Lambda de Wilks se définie de la manière suivante :

$$\Lambda = \frac{\det(SCR)}{\det(SCT)}$$

Cette statistique de test suit une loi de Wilks à $(P, n, K-1)$ degrés de liberté et l'hypothèse H_0 est : « Indépendance entre X et $Y|_{\mu_1=\cdots=\mu_k}$ ».

Une variable a un bon pouvoir discriminant si la dispersion intra-groupe est faible et si la dispersion intergroupe est forte. Donc plus le Lambda de Wilks sera faible, plus la variable considérée est discriminante. C'est ce critère qu'utilise la commande `greedy.wilks` de Rstudio que nous avons utilisée pour trouver les variables les plus discriminantes dans notre jeu de données et ainsi se focaliser sur un nombre de modèles plus réduit.

3 Qualité du modèle

3.1 Matrices de confusion

L'analyse factorielle discriminante probabiliste et la régression logistique sont utilisées pour affecter des scores à des individus. Mais donner un score à un individu sans contexte d'étude est assez absurde. C'est pourquoi, il faut évaluer le modèle choisi pour savoir si celui-ci permet une bonne prédiction des groupes d'appartenance.

Il est coutume de tester la qualité du modèle grâce aux matrices de confusions qui donnent le taux de bon et mauvais classement des individus dans chaque groupe. Nous construisons donc souvent plusieurs modèles et évaluons leurs capacités prédictives grâce à ces matrices.

Dans le cas de deux groupes les matrices de confusions se représentent souvent de la façon suivante :

	$Y=0$	$Y=1$
$Y_{pred}=0$	TN	FN
$Y_{pred}=1$	FP	TP

Avec :
— 0 = négatif
— 1 = positif
— TN = vrai négatif
— FP = faux positif
— FN = faux négatif
— TP = vrai positif

En fonction de ce que l'on cherche à faire grâce au scoring, on préfèrera retenir le modèle minimisant les faux négatifs ou les faux positifs, ou encore le modèle maximisant les vrais positifs ou vrai négatif.

3.2 Courbes ROC et AUC

Une fois notre modèle choisi grâce aux comparaisons des différentes matrices de confusion, il est possible de visualiser graphiquement la qualité globale de ce modèle grâce à une courbe ROC.

Une courbe ROC (Receiver Operating Characteristic) est un graphique représentant les performances d'un modèle de classification pour tous les seuils de classification. Cette courbe trace le taux de vrais positifs en fonction du taux de faux positifs :

— Le taux de vrais positifs (TVP) (sensibilité) est défini comme : $TVP = \frac{VP}{VP + FN}$

— Le taux de faux positifs (TFP) (spécificité) est défini comme : $TFP = \frac{VN}{VN + FP}$

3.3 Histogrammes de scores

Le seuil utilisé dans l'algorithme de recherche du meilleur modèle (pour la problématique considérée) est par défaut fixé à 0.5 (dans le cas de 2 groupes) pour décider du groupe d'appartenance. Or toute l'essence du scoring est justement de trouver le seuil qui donnera un risque d'erreur le plus faible possible, tout en prenant les contraintes de coût financier en compte.

C'est pourquoi il est courant de représenter les histogrammes de score pour déterminer le seuil optimal.

On trace les histogrammes des scores des individus en fonction de leur vrai groupe d'appartenance. Dans le cas où les 2 histogrammes sont disjoints, alors il est possible de trouver un seuil qui annulera l'erreur prise, mais ce cas est plutôt rare. Les histogrammes sont toujours plus ont moins superposés et c'est donc le travail de l'analyste de choisir le seuil qui minimisera le plus possible le taux d'erreur pris, tout en prenant en compte encore une fois toutes les contraintes financières ou matérielles.

Deuxième partie

Régression logistique

1 Introduction

Les racine de la régression logistique plongent loin dans l'histoire de l'analyse des données. En effet c'est vers 1840 que P.F Verhulst introduit ce qu'il appelle "équation logistique" pour répondre à une problématique de dynamique des populations . La régression logistique consiste à expliquer une variable Y (variable cible), par une ou plusieurs variables explicatives X_j(qualitatives ou quantitatives). Cette méthode a été introduite en 1944 par Berkson [1] en biostatistiques.
La méthode de régression logistique est très appréciée pour sa généralité, son interprétabilité et sa robustesse.
La fonction logistique est utilisée dans de nombreux domaines :

- épidémiologie : la diffusion d'une épidémie
- marketing : ventes d'un nouveau produit
- psychologie : pour prédire des comportements
- technologie

Dans ce projet on se concentre au cas où la variable à expliquer est binaire. On suppose qu'il y a donc deux groupes à discriminer. Ainsi la variable à expliquer Y prend deux modalités 0 ou 1.
Quand le nombre de modalités de la variable à expliquer est supérieur à 2 on parle de régression logistique *polytomique* (scrutin a plus de deux candidats, degrés de satisfaction pour un produit, mention a un examen....)

Les avantages de cette méthode sont qu'il n'y a pas besoin d'hypothèses de multinormalité. On se place sous les hypothèses de normalité et égalité des matrices de variance covariance.

expliquer pourquoi on ne peut pas utiliser le modele de regression linéaire simple par un exemple introductif celui du CHD

Notations On note :

- Y la variable à expliquer : $Y = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$
- $X = \begin{pmatrix} 1 \\ X_1 \\ \vdots \\ X_p \end{pmatrix} \in \mathbb{R}^{p+1}$ un vecteur de variables explicatives $X_j \quad \forall j \in [\![1, p]\!]$
- Le vecteur des coefficients $\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} \in \mathbb{R}^{p+1}$ à estimer par maximum de vraisemblance
- $x = \begin{pmatrix} 1 \\ x_1 \\ \vdots \\ x_p \end{pmatrix} \in \mathbb{R}^{p+1}$ une réalisation de X, $y = \begin{pmatrix} y_1 \\ \vdots \\ y_p \end{pmatrix}$ est une réalisation de Y, y_i suit une loi de Bernoulli de paramètre $\pi_\beta(x_i)$.
- $(X_1, Y_1), \cdots (X_n, Y_n)$ est un n-échantillon aléatoire et de même loi que le couple (X, Y)
- $(x_1, y_1), \cdots, (x_n, y_n)$ une réalisation de $(X_1, Y_1) \cdots (X_n, Y_n)$

1. Joseph BERKSON (1899,1982) était un physicien, médecin statisticien américain. Il a introduit la notion de régression logistique dans son article *Are there two regressions ? (1950)*

L'objectif de la régression logistique est de modéliser l'espérance conditionnelle de Y par rapport à X : $\mathbb{E}[Y \mid X = x]$.
En régression linéaire, on a :

$$\mathbb{E}[Y \mid X = x] = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

Ceci ne convient pas lorsque Y est binaire (0 ou 1) puisque le terme ci dessus est non borné alors que $\mathbb{P}(Y \mid X = x)$ est dans l'intervalle $[0, 1]$. On a alors quand Y est binaire (0 ou 1) :

$$\mathbb{E}[Y \mid X = x] = 1 \times \mathbb{P}(Y = 1 \mid X = x) + 0 \times \mathbb{P}(Y = 0 \mid X = x)$$

Ainsi :

$$\begin{aligned}\mathbb{E}[Y \mid X = x] &= \mathbb{P}(Y = 1 \mid X = x) \\ &= f_\beta(x)\end{aligned}$$

Cette expression est la probabilité a posteriori d'appartenir au premier groupe. La fonction $f_\beta(x)$ est appelée fonction de transfert.

Exemple avec un cas simple (une variable explicative) On va tenter d'expliquer la présence de maladie cardiovasculaire par une seule variable explicative : l'âge du patient. Ici on va donc expliquer la variable `CHD` (0 si le patient est sain, 1 sinon) par la variable `AGE`. On dispose de 100 individus

```
ggplot(cardio, aes(x = AGE, y = CHD))
+ geom_point(aes(color=CHD))
+ labs(title = "Pr sence de maladie cardiovasculaire selon l' ge  du patient", x
    = "  ge  du patient", y="Pr sence de maladie")


ggplot(data_frame, aes(x= groupe_age,y = CHD_groupe))
+ geom_point(size = 2, aes(color=CHD_groupe))
+ labs(title = "Proportion de malades selon les groupes d' ge ", x = "Groupes d'
    ge ", y="Proportion de malades") + ylim(0, 1)
+ geom_line(aes( x = groupe_age, y = exp(-5.31+0.111*groupe_age)/(1 + exp(-5.31 +
    0.111*groupe_age)), color = CHD_groupe))
```

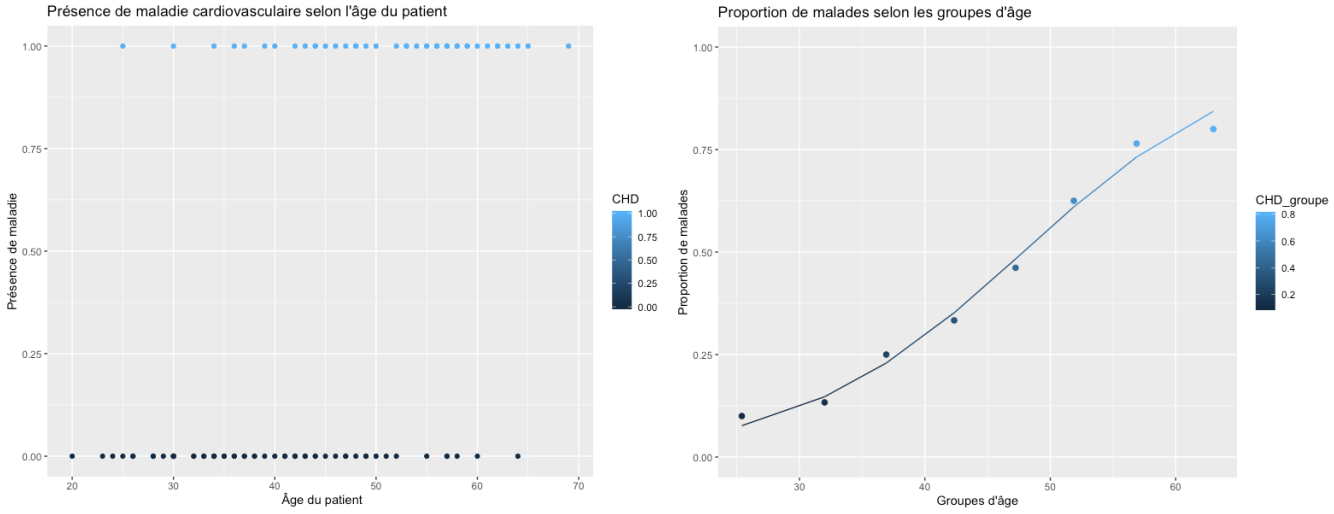

FIGURE 1 – Représentation des individus à gauche, proportion de malades selon des groupes d'âge à droite

Dans la régression logistique la fonction de transfert est la fonction $\pi_\beta(x)$

$$\mathbb{P}(G1|x) = \mathbb{P}(Y = 1|X = x) = \pi_\beta(x) = \frac{e^{\beta_0+\beta_1 x_1+\cdots+\beta_p x_p}}{1 + e^{\beta_0+\beta_1 x_1+\cdots+\beta_p x_p}}$$

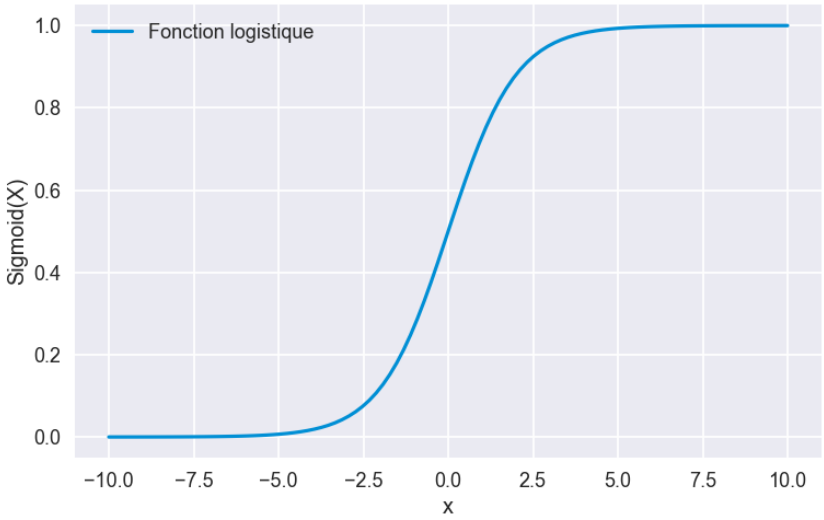

FIGURE 2 – Représentation graphique de la fonction logistique sur Python - Sigmoïde

Définition 1.1

La fonction $\pi(x)$ est appelée *fonction logistique*. Sa représentation graphique est une sigmoïde en fonction des modalités de x. La fonction $\pi(x)$ est comprise dans $]0,1[$, elle convient donc à une probabilité et donne souvent une bonne représentation des phénomènes.

$$\begin{aligned} \pi_\beta(x) : \quad & \mathbb{R}^{p+1} \longrightarrow]0,1[\\ & x \longmapsto \pi_\beta(x) = \frac{e^{\beta x}}{1+e^{\beta x}} \end{aligned}$$

On cherche à écrire l'espérance conditionnelle de la variable à expliquer Y comme combinaison linéaire de variables à expliquer X. On veut modéliser l'espérance conditionnelle $\mathbb{E}[Y|X=x]$.
On cherche la valeur moyenne de Y pour toute valeurs de X.

$$logit(\pi(x)) = \ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

Utilisée avec la fonction de logarithme néperien, logit est la réciproque de $f(x) = \dfrac{1}{1+e^{-x}}$ qui est utilisée pour linéariser les fonctions logistiques.

1.1 Interprétation avec des Odds-ratio

L'odds ratio permet de mesurer l'effet d'un facteur. L'odds ratio d'une variable explicative mesure lorsque X_j passe de x à $x+1$.

$$Odds = \frac{\pi(x)}{1-\pi(x)}$$

- Si $\beta \leq 0 \iff OR < 1$ cela indique que la variable explicative a une influence négative sur la variable à prédire.
- Si $\beta \geq 0 \iff OR > 1$ cela indique que la variable explicative a une influence positive sur la variable à prédire.

Quand la variable explicative X_j est binaire, on a :

$$Odds = \frac{\mathbb{P}(Y = 1 \mid X_j = 1)}{\mathbb{P}(Y = 0 \mid X_j = 1)}$$

On obtient un seul odds ratio qui est :

$$OR = \frac{\dfrac{\mathbb{P}(Y = 1 \mid X_j = 1)}{\mathbb{P}(Y = 0 \mid X_j = 1)}}{\dfrac{\mathbb{P}(Y = 1 \mid X_j = 0)}{\mathbb{P}(Y = 0 \mid X_j = 0)}} = e^{\beta_j}$$

C'est le facteur par lequel on multiplie la côte lorsque x passe de 0 à 1.

```r
cardio.glm = glm(CHD~AGE,family=binomial)
summary(cardio.glm)
```

```
Call:
glm(formula = CHD ~ AGE, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9718  -0.8456  -0.4576   0.8253   2.2859

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.30945    1.13365  -4.683 2.82e-06 ***
AGE          0.11092    0.02406   4.610 4.02e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 136.66  on 99  degrees of freedom
Residual deviance: 107.35  on 98  degrees of freedom
AIC: 111.35

Number of Fisher Scoring iterations: 4
```

FIGURE 3 – Sortie R de la fonction `glm`

On peut lire les coefficients $\beta_0 = -5.30945$ et $\beta_1 = 0.11092$. On va utiliser ces coefficients pour Les coefficients s'interprètent comme des logarithmes népériens d'odds ratio.

2 Estimation des paramètres

Pour estimer le vecteur de paramètres β on utilise la méthode de maximum de vraisemblance à partir d'un échantillon *iid* de n observations. En effet la variable Y à expliquer étant qualitative, on ne peut pas utiliser la méthode d'estimation par les moindres carrés habituelle.

La vraisemblance La vraisemblance pour une observation (y_i, x_i) peut s'écrire :

$$\ell(\beta; y_i, x_i) = \pi_\beta(x_i)^{y_i} \times (1 - \pi_\beta(x_i))^{1-y_i}$$

Comme les observations sont *iid* on peut écrire que la vraisemblance du n-échantillon est égale au produit des vraisemblances par observation :

$$\ell(\beta; Y, X) = \prod_{i=1}^{p} \pi_\beta(x_i)^{y_i} \times (1 - \pi_\beta(x_i))^{1-y_i}$$

La log vraisemblance

Proposition 2.1

La log vraisemblance s'écrit

$$\beta \longrightarrow \ell\ell_X(\beta; Y, X) = \sum_{i=1}^{n} y_i \ln(\pi_\beta(x_i)) + (1 - y_i) \ln(1 - \pi_\beta(x_i))$$

Démonstration. La vraisemblance s'écrit :

$$\ell(\beta; Y, X) = \prod_{i=1}^{n} \pi_\beta(x_i)^{y_i} \times (1 - \pi_\beta(x_i))^{1-y_i}$$

Or $\pi_\beta(x_i) \in]0, 1[$. Donc la vraisemblance est strictement positive, on peut calculer la log vraisemblance.

$$\begin{aligned} \ell\ell(\beta) = \ln \ell(\beta) &= \sum_{i=1}^{n} \ln(\mathbb{P}(Y = y_i \mid X = x_i)) \\ &= \sum_{i=1}^{n} y_i \ln(\pi_\beta(x_i)) + (1 - y_i) \ln(1 - \pi_\beta(x_i)) \end{aligned}$$

□

Équations de vraisemblance Le vecteur gradient au point β est défini par :

$$\nabla_\beta \ell\ell(\beta) = \begin{pmatrix} \frac{\partial \ell\ell}{\partial \beta_0}(\beta) \\ \vdots \\ \frac{\partial \ell\ell}{\partial \beta_p}(\beta) \end{pmatrix}$$

Calculons $\dfrac{\partial \ell\ell}{\partial \beta_j}(\beta) \quad \forall j \in [\![0, p]\!]$. On a :

$$\begin{aligned} \ell\ell(\beta) &= \sum_{i=1}^{n} y_i \ln(\pi_\beta(x_i)) + (1 - y_i) \ln(1 - \pi_\beta(x_i)) \\ &= \sum_{i=1}^{n} y_i \ln\left(\frac{e^{\beta x_i}}{1 + e^{\beta x_I}}\right) + (1 - y_i) \ln\left(1 - \frac{e^{\beta x_i}}{1 + e^{\beta x_i}}\right) \\ &= \sum_{i=1}^{n} y_i \ln\left(\frac{e^{\beta x_i}}{1 + e^{\beta x_I}}\right) + (1 - y_i) \ln\left(\frac{1}{1 + e^{\beta x_i}}\right) \end{aligned}$$

Avec $\beta x_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_j x_{ij} + \cdots + \beta_p x_{ip}$

$$\begin{aligned}
\frac{\partial}{\partial\beta_j}\left[\ln\left(\frac{e^{\beta x_i}}{1+e^{\beta x_I}}\right)\right] &= \frac{x_{ij}e^{\beta x_i}\left(1+e^{\beta x_i}\right)-e^{\beta x_i}\left(x_{ij}e^{\beta x_i}\right)}{\left(1+e^{\beta x_i}\right)^2}\times\frac{1+e^{\beta x_i}}{e^{\beta x_i}}\\
&= \frac{x_{ij}e^{\beta x_i}}{\left(1+e^{\beta x_i}\right)^2}\times\frac{1+e^{\beta x_i}}{e^{\beta x_i}}\\
&= \frac{x_{ij}}{1+e^{\beta x_i}}\\
\frac{\partial}{\partial\beta_j}\left[\ln\left(\frac{1}{1+e^{\beta x_i}}\right)\right] &= -\frac{x_{ij}e^{\beta x_i}}{\left(1+e^{\beta x_i}\right)^2}\times\left(1+e^{\beta x_i}\right)\\
&= -\frac{x_{ij}e^{\beta x_i}}{1+e^{\beta x_i}}
\end{aligned}$$

$$\begin{aligned}
\frac{\partial\ell\ell}{\partial\beta_j}(\beta) &= \sum_{i=1}^{n} y_i\frac{x_{ij}}{1+e^{\beta x_i}}-(1-y_i)\frac{x_{ij}e^{\beta x_i}}{1+e^{\beta x_i}}\\
&= \sum_{i=1}^{n}\frac{x_{ij}\left(y_i-e^{\beta x_i}+y_ie^{\beta x_i}\right)}{1+e^{\beta x_i}}\\
&= \sum_{i=1}^{n} x_{ij}\frac{y_i\left(1+e^{\beta x_i}\right)-e^{\beta x_i}}{1+e^{\beta x_i}}\\
&= \sum_{i=1}^{n} x_{ij}\left(y_i-\frac{e^{\beta x_i}}{1+e^{\beta x_i}}\right)\\
&= \sum_{i=1}^{n} x_{ij}\left(y_i-\pi_\beta(x_i)\right) \qquad \forall j\in[\![0,p]\!] \text{ avec } x_{i0}=1
\end{aligned}$$

On obtient l'écriture générale :

$$\nabla_\beta\ell\ell(\beta)=\sum_{i=1}^{n} x_i(y_i-\pi_\beta(x_i)) \qquad \forall i\in[\![0,n]\!] \text{ avec } x_0=1$$

On peut également l'écrire sous forme matricielle :

$$X^T(Y-\Pi_\beta)$$

avec $X=\begin{pmatrix}1 & x_{11} & \cdots & x_{1p}\\ 1 & x_{21} & \cdots & x_{2p}\\ \vdots & \vdots & \ddots & \vdots\\ 1 & x_{n1} & \cdots & x_{np}\end{pmatrix}\in\mathbb{R}^{n\times(p+1)}$ $Y=\begin{pmatrix}y_1\\ \vdots\\ y_n\end{pmatrix}\in\mathbb{R}^n$ et $\Pi_\beta=\begin{pmatrix}\pi_\beta(x_1)\\ \vdots\\ \pi_\beta(x_n)\end{pmatrix}\in\mathbb{R}^n$

Recherche d'estimateur du maximum de vraisemblance :
Si l'estimateur de maximum de vraisemblance $\widehat{\beta}$ existe, il est solution de l'équation :

$$X^T\left(Y-\Pi_\beta\right)=0$$

Ainsi rechercher les solutions de cette équation revient à résoudre $p+1$ équations à $p+1$ inconnues $(\beta_0,\beta_1,\cdots,\beta_p)$:

$$\begin{cases} y_1 + \cdots + y_n = \pi_\beta(x_1) + \cdots + \pi_\beta(x_n) & j = 0 \\ x_{1j}y_1 + \cdots + x_{nj}y_n = x_{1j}\pi_\beta(x_1) + \cdots + x_{nj}\pi_\beta(x_n), & \forall j \in [\![1, p]\!] \end{cases}$$

$$\iff \begin{cases} y_1 + \cdots + y_n = \dfrac{e^{x_1^T\beta}}{1 + e^{x_1^T\beta}} + \cdots + \dfrac{e^{x_n^T\beta}}{1 + e^{x_n^T\beta}} & j = 0 \\ x_{1j}y_1 + \cdots + x_{nj}y_n = x_{1j}\dfrac{e^{x_1^T\beta}}{1 + e^{x_1^T\beta}} + \cdots x_{nj}\dfrac{e^{x_n^T\beta}}{1 + e^{x_n^T\beta}}, & \forall j \in [\![1, p]\!] \end{cases}$$

Ce système d'équations n'a pas de solution analytique et se résout par des procédures de calcul numérique

Théorème 2.1

Si X est de rang maximal, la log vraisemblance $\beta \longmapsto \ell\ell(\beta)$ est strictement concave : si $\widehat{\beta}$ existe il est unique.

Matrice Hessienne Calculons la matrice Hessienne de la log vraisemblance

$$\nabla^2_\beta \ell\ell(\beta; Y, X) = \left(\frac{\partial^2 \ell\ell}{\partial\beta_i \partial\beta_j}(\beta; Y, X)\right)_{1 \le i,j \le p}$$

$$\begin{aligned} \nabla^2_\beta \ell\ell(\beta; Y, X) &= \nabla_\beta \left(\nabla_\beta \ell\ell(\beta; Y, X)\right) \\ &= \nabla_\beta \left(\sum_{i=1}^p x_i(y_i - \pi_\beta(x_i))\right) \\ &= -\sum_{i=1}^p x_i^T \nabla_\beta(\pi_\beta(x_i)) \\ &= -\sum_{i=1}^p x_i^T \frac{x_i e^{\beta x_i}}{(1 + e^{\beta x_i})^2} \\ &= -\sum_{i=1}^p x_i^T x_i \frac{e^{\beta x_i}}{(1 + e^{\beta x_i})^2} \\ &= -\sum_{i=1}^p x_i^T x_i \pi_\beta(x_i)(1 - \pi_\beta(x_i)) \end{aligned}$$

Or $\pi_\beta(x_i)(1 - \pi_\beta(x_i)) > 0$ car $\pi_\beta(x_i) \in]0, 1[$.

De plus $x_i^T x_i = \|x_i\|^2$ donc $\|x_i\|^2 \ge 0$ et $\|x_i\|^2 = 0$ pour $x_i = 0$.

Sous forme matricielle on a : Pour alléger les notations on va poser $\pi_i = \pi_\beta(x_i)$

$$H(\beta; Y, X) = \nabla^2_\beta \ell\ell(\beta; Y, X) = -\begin{pmatrix} \sum_{i=1}^n \pi_i(1-\pi_i) & \sum_{i=1}^n x_{i1}\pi_i(1-\pi_i) & \cdots & \sum_{i=1}^n x_{ip}\pi_i(1-\pi_i) \\ \sum_{i=1}^n x_{i1}\pi_i(1-\pi_i) & \sum_{i=1}^n (x_{i1})^2\pi_i(1-\pi_i) & \cdots & \sum_{i=1}^n x_{i1}x_{ip}\pi_i(1-\pi_i) \\ \vdots & & \ddots & \vdots \\ \sum_{i=1}^n x_{ip}\pi_i(1-\pi_i) & \cdots & \cdots & \sum_{i=1}^n (x_{ip})^2\pi_i(1-\pi_i) \end{pmatrix}$$

$$= -\begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix}^T \begin{pmatrix} \pi_1(1-\pi_1) & & 0 \\ & \ddots & \\ 0 & & \pi_n(1-\pi_n) \end{pmatrix} \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix}$$

$$= -X^T \Delta_\beta X$$

Δ_β est une matrice $n \times n$ diagonale où le k-ième terme est égal à $\pi_k(1-\pi_k) > 0$.
De plus si X est de rang maximal ($rg(X) = p+1$) alors X est injective et la matrice Δ_β est définie positive.

Ainsi la matrice hessienne de la log vraisemblance est définie négative, alors la log vraisemblance est strictement concave par rapport à β. Ceci garantit, s'il existe, l'unicité du maximum de cette fonction. Ainsi quel que soit le choix des conditions initiales ou de l'algorithme utilisé, les estimateurs du maximum de vraisemblance convergeront vers la vraie valeur $\widehat{\beta}$.

2.1 Comportement asymptotique des estimateurs du maximum de vraisemblance

Puisque on utilise l'estimateur du maximum de vraisemblance, il est possible de construire des intervalles de confiance asymptotiques

3 Tests et sélection des variables

3.1 Critères de l'AIC et BIC

Le critère AIC permet de comparer des modèles qui ne sont pas forcément emboîtés Pour cela on peut utiliser les fonctions `bestglm` ou `step`.

AIC (*Akaike Information criterion*)

$$AIC = -2\ell\ell(\widehat{\beta}) + 2k$$

L'objectif est de minimiser l'AIC

BIC (*Bayesian information criterion*)

$$BIC = -2\ell\ell(\widehat{\beta}) + \ln(n)k$$

Pour de grands échantillons le critère BIC aura tendance à favoriser les modèles avec moins de paramètres que le critère AIC

3.2 Test de Wald

3.3 Test du rapport des vraisemblances

4 Validation

4.1 Courbe ROC, AUC

Une courbe ROC (*Receiver Operating Characteristic*) est un graphique représentant les performances d'un modèle de classification pour tous les seuils de classification. Cette courbe trace le taux de vrais positifs en fonction du taux de faux positifs :

Le taux de vrais positifs (TVP) (sensibilité) est défini comme :

$$TVP = \frac{VP}{VP + FN}$$

Le taux de faux positifs (TFP) (spécificité) est défini comme :

$$TFP = \frac{VN}{VN + FP}$$

La courbe ROC résume les performances de toutes les règles de classement que l'on peut obtenir en faisant varier le seuil de décision. Les termes "positifs" et "négatifs" dépendent de ce que l'on aura choisi au préalable.

La *sensibilité* se définit comme le pourcentage de vrais positifs : $1 - \beta$: D'un point de vue médical cela veut dire être testé positif à un test détectant la présence de maladie quand on est bien malade .
La *spécificité* se définit quant à elle comme le pourcentage de vrais négatifs : $1-\alpha$. D'un point de vue médical cela signifie être testé négatif à un test détectant la présence de maladie, quand on est bien sain.

5 Lien avec le scoring

Une des applications les plus courantes de la régression logistique est la construction de scores

parler des seuils, déplacer la sous sectoin de la courbe ROC ici ? Avec les courbes ROC on peut faire varier le seuil s et voir ce qu'il se passe avec la sensibilité / spécificité etc. trouver la meilleure courbe ROC ?

6 Des Exemples avec R

6.1 Fonction GLM

On utilise avec le logiciel R la fonction `glm` modèle linéaire généralisé

6.2 Exemple médical sur les données diabète

Cette base de données contient des observations de 768 individus. Tous les patients ici sont des femmes d'au moins 21 ans d'origine indienne Pima. Le peuple Pima est connu pour être une des communautés comportant le plus grand pourcentage d'obèses et de diabétiques au monde, et à ce titre est un sujet d'études pour les scientifiques.
L'objectif est de prédire si la patiente est diabétique ou non.
Le but est d'expliquer la variable Y ici `Outcome` par les variables explicatives quantitatives suivantes :

- `Pregnancies` : Nombre de grossesses de la patiente.
- `Glucose` : Concentration de glucose plasmatique après 2 heures par un test de tolérance au glucose par voie orale.
- `BloodPressure` : Pression artérielle diastolique (mm Hg)
- `SkinThickness` : Épaisseur du pli cutané au niveau du triceps (mm)
- `Insulin` : mesure de l'insuline 2h après une injection d'insuline (mu U/ml)

— `BMI` : Indice de masse corporelle : $\frac{\text{poids en kg}}{(\text{ taille en m })^2}$
— `DiabetesPedigreeFunction` : score qui représente la probabilité d'être diabétique selon les antécédents familiaux
— `Age` : âge de la patiente au moment du diagnostic

```
> head(diabete)
```

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI
1	6	148	72	35	0	33.6
2	1	85	66	29	0	26.6
3	8	183	64	0	0	23.3
4	1	89	66	23	94	28.1
5	0	137	40	35	168	43.1
6	5	116	74	0	0	25.6

	DiabetesPedigreeFunction	Age	Outcome
1	0.627	50	1
2	0.351	31	0
3	0.672	32	1
4	0.167	21	0
5	2.288	33	1
6	0.201	30	0

FIGURE 4 – Représentation des 6 premières patientes

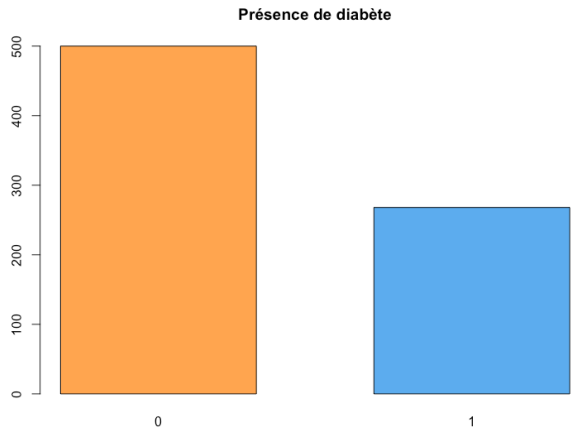

FIGURE 5 – Présence de diabète ou non chez les patientes

On voit que dans cet échantillon 268 sont diabétiques contre 500 non diabétiques

6.2.1 Choix des modèles

```
modele_complet = glm(Outcome ~ .  , data = diabete, family = binomial(link = "
    logit"))
summary(modele_complet)
```

Nous décidons à partir de ces résultats de créer un nouveau modèle : ce nouveau modèle comportera les variables les plus significatives

Modèle 2 : Variables significatives

6.3 Exemple économique avec des données sur les exploitations fermières

```
  Pregnancies        Glucose      BloodPressure    SkinThickness      Insulin          BMI        DiabetesPedigreeFunction
Min.   : 0.000   Min.   :  0.0   Min.   :  0.00   Min.   : 0.00   Min.   :  0.0   Min.   : 0.00   Min.   :0.0780
1st Qu.: 1.000   1st Qu.: 99.0   1st Qu.: 62.00   1st Qu.: 0.00   1st Qu.:  0.0   1st Qu.:27.30   1st Qu.:0.2437
Median : 3.000   Median :117.0   Median : 72.00   Median :23.00   Median : 30.5   Median :32.00   Median :0.3725
Mean   : 3.845   Mean   :120.9   Mean   : 69.11   Mean   :20.54   Mean   : 79.8   Mean   :31.99   Mean   :0.4719
3rd Qu.: 6.000   3rd Qu.:140.2   3rd Qu.: 80.00   3rd Qu.:32.00   3rd Qu.:127.2   3rd Qu.:36.60   3rd Qu.:0.6262
Max.   :17.000   Max.   :199.0   Max.   :122.00   Max.   :99.00   Max.   :846.0   Max.   :67.10   Max.   :2.4200
      Age        Outcome
Min.   :21.00   0:500
1st Qu.:24.00   1:268
Median :29.00
Mean   :33.24
3rd Qu.:41.00
Max.   :81.00
```

FIGURE 6 – Variables

Troisième partie

Analyse Factorielle Discriminante

1 Théorie de la méthode probabiliste

Le but premier de cette méthode est de prédire au mieux les valeurs d'une variable Y qualitative à K modalités, à partir de p variables explicatives $X = (X_1, \cdots, X_p)$ quantitatives.

Dans cette section, nous allons définir des règles de décision bayésiennes qui vont permettre d'affecter un nouvel individu à la classe "la plus probable" et non pas au groupe « le plus proche » comme c'est le cas pour l'analyse discriminante géométrique. Pour cela, il est nécessaire de faire des hypothèses probabilistes sur les données, d'où le nom de la méthode.

On suppose maintenant que les données sont issues d'une population regroupant des individus de K groupes prédéfinis différents $G_1, \cdots, G_k$ et que :

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

On notera :

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

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

1.1 Règle du maximum à posteriori

La règle du maximum à posteriori affecte une nouvelle observation x dans le groupe k^* le plus probable sachant x :

$$k^* = \underset{k=1,\cdots,K}{\arg\max}\ \mathbb{P}(G_k \mid x)$$

où $\mathbb{P}(G_k \mid x) = \mathbb{P}(Y = k \mid X = x)$ est la probabilité conditionnelle appelée la **probabilité à posteriori** de G_k. Les probabilités à posteriori $\mathbb{P}(G_k \mid x)$ 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 :

$$k^* = \underset{k=1,\cdots,K}{\arg\max}\ p_k f_k(x)$$

On parle alors souvent de **règle de Bayes**.

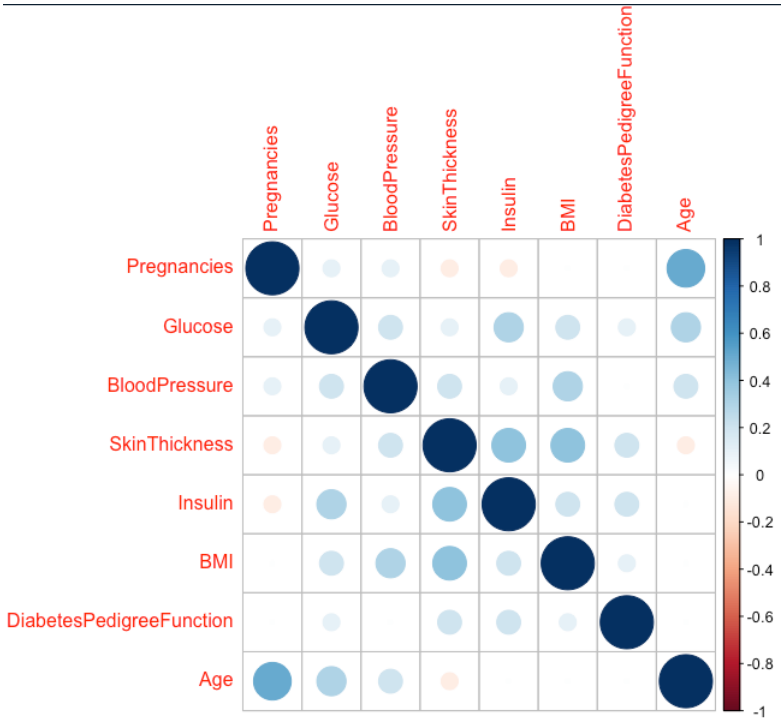

FIGURE 7 – Corrélation entre variables

Démonstration. On utilise le théorème de Bayes qui donne :

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

Or $f_X(x)$ est indépendante de k donc il suffit de maximiser $f_k(x)p_k$. □

Plusieurs approches sont possibles.

1. Les approches paramétriques :
   — On peut supposer que $f_k(x)$ a une forme paramétrique et estimer les paramètres sur l'échantillon d'apprentissage. Par exemple, $f_k(x)$ est une densité $\mathcal{N}(\mu_k, \Sigma_k)$ pour les méthodes LDA et QDA.
   — On peut supposer que la probabilité à posteriori $\mathbb{P}(G_k \mid x)$ a une expression paramétrique et l'estimer directement. Par exemple $\mathbb{P}(G_1 \mid x) = \dfrac{e^{\beta^T x}}{1 + e^{\beta^T x}}$ pour la régression logistique avec $K = 2$.
2. 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.

1.2 Le cas gaussien

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

$$f_k(x) = \frac{1}{(2\pi)^{\frac{p}{2}} \left(\det(\Sigma_k)\right)^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x - \mu_k)^T \Sigma_k^{-1}(x - \mu_k)\right)$$

```
Call:
glm(formula = Outcome ~ ., family = binomial(link = "logit"),
    data = diabete)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5566  -0.7274  -0.4159   0.7267   2.9297

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)             -8.4046964  0.7166359 -11.728  < 2e-16 ***
Pregnancies              0.1231823  0.0320776   3.840 0.000123 ***
Glucose                  0.0351637  0.0037087   9.481  < 2e-16 ***
BloodPressure           -0.0132955  0.0052336  -2.540 0.011072 *
SkinThickness            0.0006190  0.0068994   0.090 0.928515
Insulin                 -0.0011917  0.0009012  -1.322 0.186065
BMI                      0.0897010  0.0150876   5.945 2.76e-09 ***
DiabetesPedigreeFunction 0.9451797  0.2991475   3.160 0.001580 **
Age                      0.0148690  0.0093348   1.593 0.111192
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 993.48  on 767  degrees of freedom
Residual deviance: 723.45  on 759  degrees of freedom
AIC: 741.45

Number of Fisher Scoring iterations: 5
```

FIGURE 8 – Variables

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 :

$$k^* = \underset{k=1,\cdots,K}{\arg\min}\, D_k^2(x)$$

où

$$D_k^2(x) = (x-\mu_k)^T \Sigma_k^{-1}(x-\mu_k) - 2\ln(p_k) + \ln\left(\det \Sigma_k\right)$$

est appelé le **carré de la distance de Mahalanobis théorique.**

Démonstration. Maximiser $p_k f_k(x)$ est équivalent à maximiser $\ln\left(p_k f_k(x)\right)$ et

$$\begin{aligned}\ln\left(p_k f_k(x)\right) &= \ln(p_k) + \ln(f_k(x)) \\ &= \ln(p_k) - \frac{p}{2}\ln(2\pi) - \frac{1}{2}\ln\left(\det(\Sigma_k)\right) - \frac{1}{2}(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k)\end{aligned}$$

Donc avec $\frac{p}{2}\ln(2\pi)$ qui est indépendant de k, maximiser $\ln(p_k f_k(x))$ est équivalent à minimiser :

$$-2\left(\ln(p_k) - \frac{1}{2}\ln(\det(\Sigma_k)) - \frac{1}{2}(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k)\right) = D_k(x)^2$$

□

1.2.1 Estimation des paramètres

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

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

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

$$\ell(\theta) = \prod_{i=1}^{n} f_X(x_i) = \prod_{k=1}^{K} \prod_{x_i \in E_k} p_k f_k(x_i)$$

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

$$\ell\ell(\theta) = \ln(\ell(\theta)) = \sum_{k=1}^{K} \sum_{x_i \in E_k} \left(\ln(p_k) - \frac{p}{2}\ln(2\pi) - \frac{1}{2}\ln(\det(\Sigma_k)) - \frac{1}{2}(x_i - \mu_k)^T \Sigma_k^{-1}(x_i - \mu_k) \right)$$

On obtient alors les estimateurs du maximum de vraisemblance suivant :

$$\widehat{p_k} = \frac{n_k}{n}$$

$$\widehat{\mu_k} = \frac{1}{n_k} \sum_{i \in E_k} x_i$$

$$\widehat{\Sigma_k} = \begin{cases} \widehat{\Sigma} = \dfrac{1}{n} \displaystyle\sum_{k=1}^{K} \sum_{i \in E_k} (x_i - \mu_k)(x_i - \mu_k)^T & \text{dans le cas homoscédastique} \\ \widehat{\Sigma_k} = \dfrac{1}{n_k} \displaystyle\sum_{i \in E_k} (x_i - \mu_k)(x_i - \mu_k)^T & \text{dans le cas hétéroscédastique} \end{cases}$$

Les estimateurs de $\widehat{\Sigma_k}$ étant biaisés, on a les estimateurs sans biais suivants :

$$\widehat{\Sigma} = \frac{1}{n-K} \sum_{k=1}^{K} \sum_{i \in E_k} (x_i - \mu_k)(x_i - \mu_k)^T$$

$$\widehat{\Sigma_k} = \frac{1}{n_k - 1} \sum_{i \in E_k} (x_i - \mu_k)(x_i - \mu_k)^Y$$

1.2.2 Analyse discriminante quadratique (QDA)

On se place dans le cas où il existe $k \neq k^*$ 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 précédente (*à rappeler*) :

— μ_k est estimée par $g_k = \dfrac{1}{n_k} \displaystyle\sum_{i \in E_k} x_i$

— Σ_k est estimée par $V_k = \dfrac{1}{n_k} \displaystyle\sum_{i \in E_k} (x_i - g_k)(x_i - g_k)^T$ ou encore par sa version sans biais=

$$V_k = \frac{1}{n_k - 1} \sum_{i \in E_k} (x_i - g_k)(x_i - g_k)^T$$

— p_k est estimée par $\pi_k = \dfrac{n_k}{n}$

Avec ces estimateurs, la règle de Bayes se réécrit :

$$k^* = \underset{k=1,\cdots,K}{\arg\min}\, Q_k(x)$$

où

$$Q_k(x) = (x - g_k)^T V_k^{-1}(x - g_k) - 2\ln(\pi_k) + \ln(\det(V_k))$$

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

1.2.3 Analyse discriminante linéaire (LDA)

On se place dans le cas où $\Sigma_1 = \cdots = \Sigma_K = \Sigma$ appelé homoscédastique. Dans ce cas, la règle de Bayes se réecrit :

$$k^* = \underset{k=1,\cdots,K}{\arg\max}\, x^T\Sigma^{-1}\mu_k - \frac{1}{2}\mu_k^T + \ln(p_k)$$

Démonstration. $\Sigma_k = \Sigma$ pour tout $k = 1, \cdots, K$ donc

$$\begin{aligned} D_k^2(x) &= (x-\mu_k)^T\Sigma^{-1}(x-\mu_k) - 2\ln(p_k) + \ln(\det(\Sigma)) \\ &= \underbrace{x^T\Sigma^{-1}x}_{\text{indépendant de } k} - \underbrace{x^T\Sigma^{-1}\mu_k}_{\text{car } \Sigma^{-1} \text{ symétrique}} + \mu_k^T\Sigma^{-1}\mu_k - 2\ln(p_k) + \underbrace{\ln(\det(\Sigma))}_{\text{indépendant de } k} \end{aligned}$$

Donc minimiser $D_k^2(x)$ est équivalent à maximiser $-\frac{1}{2}\left(2x^T\Sigma^{-1}\mu_k + \mu_k^T\Sigma^{-1}\mu_k - 2\ln(p_k)\right)$ □

On estime alors les paramètres sur l'échantillon d'apprentissage et en reprenant les notations de la section précédente (*à rappeler donc*) :

— μ_k est estimée par $g_k = \dfrac{1}{n_k}\displaystyle\sum_{i\in E_k} x_i$

— La matrice de variances-covariances Σ commune aux différents groupes est estimée par

$$V_k = \frac{1}{n}{}_{k=1}\sum_{i\in E_k}(x_i - g_k)(x_i - g_k)^T$$

ou encore par sa version sans biais :

$$W = \frac{1}{n-K}\sum_{k=1}\sum_{i\in E_k}(x_i - g_k)(x_i - g_k)^T$$

— p_k est estimée par $\pi_k = \dfrac{n_k}{n}$

Avec ces estimateurs, la règle de Bayes se réécrit :

$$k^* = \underset{k=1,\cdots,K}{\arg\max}\, L_k(x)$$

où

$$L_k(x) = x^TW^{-1}g_k - \frac{1}{2}g_k^TW^{-1}g_k + \ln(\pi_k)$$

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

Remarque 1. On retrouve la fonction linéaire discriminante

$$L_k(x) = x^TW^{-1}g_k - \frac{1}{2}g_k^TW^{-1}g_k$$

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 ($p_1 = \cdots = p_K$), la règle de l'analyse discriminante linéaire (LDA) est équivalente à la méthode de l'analyse discriminante géométrique.

En pratique, on estime les probabilités à posteriori par :

$$\widehat{\mathbb{P}}(G_k \mid x) = \frac{\exp\left(-\frac{1}{2}\widehat{D_k^2}(x)\right)}{\sum_{i=1}^{K}\exp\left(-\frac{1}{2}D_l^2(x)\right)}$$

où

$$g_1(x) = \begin{cases} \ln(\widehat{\Sigma_k}) & \text{dans le cas hétéroscédastique (quadratique)} \\ 0 & \text{dans le cas homoscédastique (linéaire)} \end{cases}$$

$$g_2(x) = \begin{cases} -2\ln(\widehat{p_k}) & \text{si toutes les probabilités à priori ne sont pas égales} \\ 0 & \text{si elles sont toutes égales (équiprobabilité)} \end{cases}$$

Démonstration. Dans le cas homoscédastique, $\widehat{\Sigma_k} = \widehat{\Sigma}$ est indépendant de k, donc $\exp\left(\ln(\det(\widehat{\Sigma}))\right) = \det(\widehat{\Sigma})$se met en facteur au dénominateur et s'annule avec le numérateur dans

$$\mathbb{P}(G_k \mid x) = \frac{p_k f_k(x)}{\sum_{l=1}^{K} p_l f_l(x)}$$

□

On prendra :

$$\widehat{\mu_k} = g_k \qquad \widehat{p_k} = \frac{n_k}{n} \qquad \widehat{\Sigma_k} = \begin{cases} V_k & \text{dans le cas hétéroscédastique (quadratique)} \\ W & \text{dans le cas homoscédastique (linéaire)} \end{cases}$$

1.2.4 Cas particulier de deux groupes

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

$$\begin{aligned} \Delta_{\frac{1}{2}}(x) &= L_1(x) - L_2(x) \\ &= x^T W^{-1}(g_1 - g_2) - \frac{1}{2}(g_1 + g_2)^T W^{-1}(g_1 - g_2) + \ln\left(\frac{\pi_1}{\pi_2}\right) \end{aligned}$$

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 :

$$x^T W^{-1}(g_1 - g_2) - \frac{1}{2}(g_1 + g_2)^T W^{-1}(g_1 - g_2) + \ln\left(\frac{\pi_1}{\pi_2}\right) \geq 0 \Longrightarrow \text{ l'individu } i \text{ est affecté au groupe 1}$$

$$x^T W^{-1}(g_1 - g_2) - \frac{1}{2}(g_1 + g_2)^T W^{-1}(g_1 - g_2) + \ln\left(\frac{\pi_1}{\pi_2}\right) < 0 \Longrightarrow \text{ l'individu } i \text{ est affecté au groupe 2}$$

Remarque 2. 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 x avec la métrique W^{-1} sur le premier axe discriminant (calcul du score $x^T W^{-1}(g_1 - g_2)$) et à classer x selon un seuil s qui est le milieu des moyennes des groupes sur ce score $-\ln\left(\frac{\pi_1}{\pi_2}\right)$

1.3 Sélection des variables d'entrée et qualité du modèle

1.3.1 Qualité du modèle

L'analyse factorielle discriminante probabiliste est donc utilisée pour affecter des scores à des individus. Mais donner un score à un individu sans contexte d'étude est assez absurde. C'est pourquoi, il faut évaluer le modèle choisi pour savoir si celui-ci permet une bonne prédiction des groupes d'appartenance.

Il est coutume de tester la qualité du modèle grâce aux matrices de confusions qui donnent le taux de bon et mauvais classement des individus dans chaque groupe.

Nous construisons donc souvent plusieurs modèles et évaluons leurs capacités prédictives grâce à ces matrices. Or si l'on a un très grand nombre de variables explicatives, cela peut être extrêmement fastidieux et coûteux de tester tous les modèles possibles et imaginables. C'est pourquoi il est coutume d'utiliser des outils de pré-sélection de variables les plus discriminantes.

1.3.2 Sélection des variables

Le Lambda de Wilks est souvent utilisé dans les logiciels comme critère pour ne garder que les variables apportant de l'information sur l'appartenance ou non d'un individu à un groupe. Le Lambda de Wilks est une approche paramétrique permettant de tester si plusieurs variables continues distinctes $X = (X_1, \cdots, X_p)$ sont liées à une variable qualitative Y à $K \geq 2$ groupes, lorsqu'elles sont considérées avec leurs différentes interactions multivariées.

Les hypothèses d'utilisation de ce test sont : $X_{|Y=1}, \cdots X_{|Y=K}$ suivent une loi normale et leur matrice de covariance respective sont égales (homoscédasticité).

La statistique du test du Lambda de Wilks se définie de la manière suivante :

$$\Lambda = \frac{\det(SCR)}{\det(SCT)}$$

Où SCR est la matrice de variance-covariance intragroupe et SCT la matrice de variance-covariance globale. Cette statistique de test suit une loi de Wilks à $(P, n, K-1)$ degrés de liberté et l'hypothèse H_0 est : « Indépendance entre X et $Y_{|\mu_1=\cdots=\mu_k}$ ».

Une variable a un bon pouvoir discriminant si la dispersion intra-groupe est faible et si la dispersion intergroupe est forte. Donc plus le Lambda de Wilks sera faible, plus la variable considérée est discriminante. C'est ce critère qu'utilise la commande `greedy.wilks` de Rstudio que nous avons utilisée pour trouver les variables les plus discriminantes dans notre jeu de données et ainsi se focaliser sur un nombre de modèles plus réduit.

2 ACP préliminaire (présentation du jeu de données)

3 Exemple sur R