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 deux 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

Régression logistique

1 Introduction

La Régression logistique consiste à expliquer une variable Y (variable cible), par une ou plusieurs variables explicatives X_j.
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.

expliquer pourquoi on ne peut pas utiliser le modele de regression linéaire simple

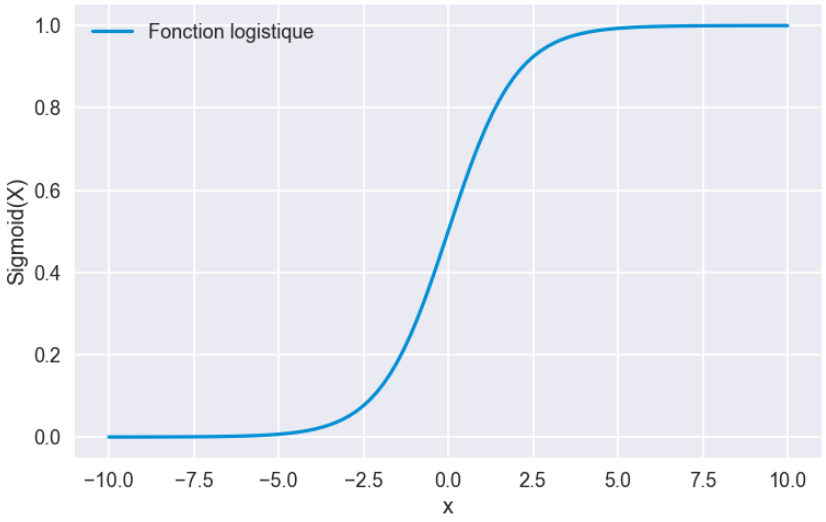
Quand le nombre de modalités de la variable à expliquer est supérieur à 2 on parle de régression logistique *polytomique* (scrutin a plus de deux candidats, degrés de satisfaction pour un produit, mention a un examen....)

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

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

Les avantages de cette méthode sont qu'il n'y a pas besoin d'hypothèses de multinormalité. Dans la régression logistique

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



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.

— $(X_1, Y_1), \cdots (X_n, Y_n)$ 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 $(X_1, Y_1) \cdots (X_n, Y_n)$

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.

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

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 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.

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

```
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 1 – 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

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

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

Démonstration. La vraisemblance s'écrit :

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

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

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

□

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

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

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

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

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

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 :

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

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 : $\widehat{\beta}$ existe et est unique.

Démonstration. La log vraisemblance s'écrit

$$\ell\ell_X(\beta) = \sum_{i=1}^{n} y_i \ln\left(\pi_\beta(x_i)\right) + (1-y_i)\ln\left(1-\pi_\beta(x_i)\right)$$

Les dérivées première et seconde de la fonction $\ln\left(\pi_\beta(x)\right)$ sont les suivantes :

$$\begin{aligned}
\frac{\partial \ln(\pi_\beta(x))}{\partial x} &= \frac{1}{\pi_\beta(x)} \times \frac{\partial \pi_\beta(x)}{\partial x} \\
&= \frac{1+e^{\beta x}}{e^{\beta x}} \times \frac{e^{\beta x}\left(1+e^{\beta x}\right) - e^{2\beta x}}{\left(1+e^{\beta x}\right)^2} \\
&= \frac{1}{1+e^{\beta x}} = \pi_\beta(-x) \\
\frac{\partial^2 \ln\left(\pi_\beta(x)\right)}{\partial x^2} &= \frac{\partial}{\partial x}\left(\frac{1}{1+e^{\beta x}}\right) \\
&= \frac{-e^{\beta x}}{\left(1+e^{\beta x}\right)^2} < 0
\end{aligned}$$

Les dérivées première et seconde de la fonction $\ln\left(1-\pi_\beta(x)\right)$ sont les suivantes :

$$\begin{aligned}
\frac{\partial \ln(1-\pi_\beta(x))}{\partial x} &= \frac{1}{1-\pi_\beta(x)} \times \frac{\partial (1-\pi_\beta(x))}{\partial x} \\
&= \left(1+e^{\beta x}\right) \times \frac{-e^{\beta x}}{\left(1+e^{\beta x}\right)^2} \\
&= \frac{-e^{\beta x}}{1+e^{\beta x}} = -\pi_\beta(x) \\
\frac{\partial^2 \ln\left(1-\pi_\beta(x)\right)}{\partial x^2} &= \frac{\partial}{\partial x}\left(\frac{-e^{\beta x}}{1+e^{\beta x}}\right) \\
&= \frac{-e^{\beta x}}{\left(1+e^{\beta x}\right)^2} < 0
\end{aligned}$$

Les fonctions $\ln\left(\pi_\beta(x)\right)$ et $\ln\left(1-\pi_\beta(x)\right)$ sont strictement concaves (car leurs dérivées secondes sont négatives). Alors la log-vraisemblance est strictement concave, ce qui garantit 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 β_0 □

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 $\|x_i\|^2 \geq 0$ et $\|x_i\|^2 = 0$ pour $x_i = 0$.

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

$$\begin{aligned}
H(\beta; Y, X) = \nabla^2_\beta \ell\ell(\beta; Y, X) &= \begin{pmatrix}
\sum_{i=1}^{n} \pi_i(1-\pi_i) & \sum_{i=1}^{n} x_{i1}\pi_i(1-\pi_i) & \cdots & \sum_{i=1}^{n} x_{ip}\pi_i(1-\pi_i) \\
\sum_{i=1}^{n} x_{i1}\pi_i(1-\pi_i) & \sum_{i=1}^{n} (x_{i1})^2\pi_i(1-\pi_i) & \cdots & \sum_{i=1}^{n} x_{i1}x_{ip}\pi_i(1-\pi_i) \\
\vdots & & \ddots & \vdots \\
\sum_{i=1}^{n} x_{ip}\pi_i(1-\pi_i) & \cdots & \cdots & \sum_{i=1}^{n} (x_{ip})^2\pi_i(1-\pi_i)
\end{pmatrix} \\
&= \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix}^T \begin{pmatrix} \pi_1(1-\pi_1) & & 0 \\ & \ddots & \\ 0 & & \pi_n(1-\pi_n) \end{pmatrix} \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix} \\
&= -X^T \Delta_\beta X
\end{aligned}$$

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 est définie.

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 $\widehat{\beta}$.

Existence du maximum de vraisemblance :

La concavité de la log vraisemblance ne garantit pas l'existence de solution pour le système d'équations du problème de maximisation.

Les cas où l'algorithme ne converge pas sont rares et cela traite du problème de **séparabilité des points** qui a été introduit par A.ALBERT et J. A. ANDERSON [1]

Le problème de maximisation de la fonction de log vraisemblance traite des différentes configurations des n points du jeu de données. On distingue trois types de séparabilité de points :

1. une séparation complète des points
2. une séparation quasi complète des points
3. un chevauchement des points

Pour résumer s'il y a une séparation complète ou quasi complète des points, le maximum de vraisemblance $\widehat{\beta}$ n'existe pas et le maximum de vraisemblance vaut 1.
S'il y a chevauchement des points, le maximum de vraisemblance existe et est unique.
Selon ces mêmes auteurs les problèmes rencontrés avec la séparation complète ou quasi complète des données est liée à la taille de l'échantillon de données qui est trop faible. En effet plus on augmente la taille de l'échantillon, plus la probabilité de trouver des points séparés tend vers zéro.

Pour illustrer cela on va générer un échantillon où les points sont complètement séparés

```
matrice = matrix(nrow = 2, ncol = 100)

for (i in 1:50){
  matrice[1,i] = runif(1, min = -1, max = 0)  # ici le x_i
  matrice[2,i] = 0}                           # ici le y_i

for (i in 51:100){
  matrice[1,i] = runif(1, min = 0, max = 1)
  matrice[2,i] = 1

}
plot(matrice[1,], matrice[2,])
```

```
modele = glm(matrice[2,]~ matrice[1,], family = binomial(link="logit"))
```

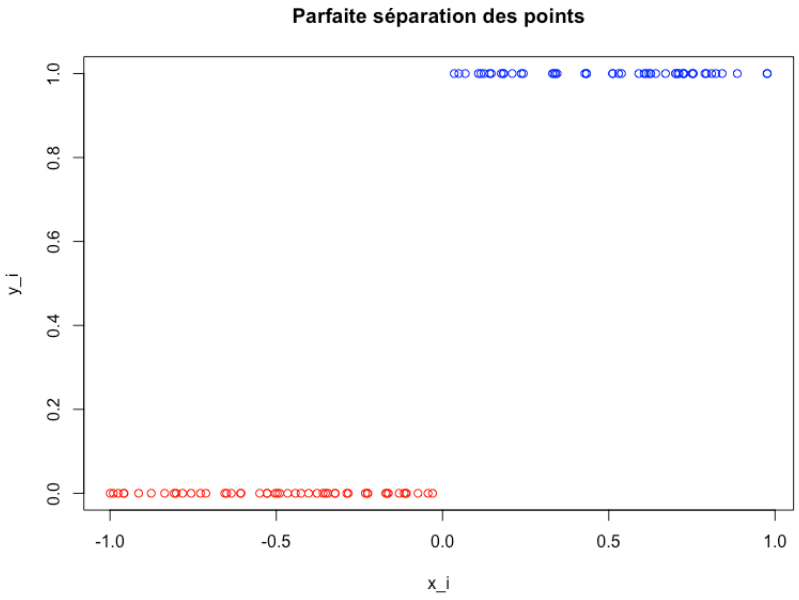


FIGURE 2 – Séparation parfaite des points

1. ALBERT et ANDERSON dans *On the existence of maximum likelihood estimates in logistic regression*

On a ici une séparation parfaite des points

```
Warning messages:
1: glm.fit: l'algorithme n'a pas convergé
2: glm.fit: des probabilités ont été ajustées numériquement à 0 ou 1
```

On obtient un message d'erreur informant que l'algorithme n'a pas convergé.

On peut changer la valeur de $y_1 = 1$

```
matrice[2,1] <- 1
modele = glm(matrice[2,]~ matrice[1,], family = binomial(link="logit"))
summary(modele)
```

```
Call:
glm(formula = matrice[2, ] ~ matrice[1, ], family = binomial(link = "logit"))

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.45331  -0.02120   0.00006   0.00441   2.71136

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)    0.6934     0.6984   0.993  0.32080
matrice[1, ]  20.1227     6.4537   3.118  0.00182 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 138.589  on 99  degrees of freedom
Residual deviance:  15.358  on 98  degrees of freedom
AIC: 19.358

Number of Fisher Scoring iterations: 9
```

2.1 Calcul des estimateurs : deux algorithmes

Deux algorithmes sont souvent utilisées par les logiciels pour calculer les estimateurs du maximum de vraisemblance .

Méthode de Newton Raphson

Algorithme IRLS On peut également utiliser l'algorithme IRLS (*Iteratively Reweighted Least Squares*)

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

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 3 – Représentation des 6 premières patientes

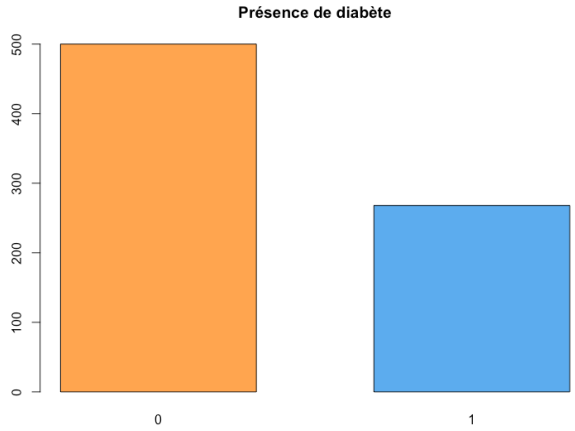


FIGURE 4 – Présence de diabète ou non chez les patientes

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)
```

```
  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 5 – Variables

```
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
```

6.3 Exemple économique avec des données sur les exploitations fermières

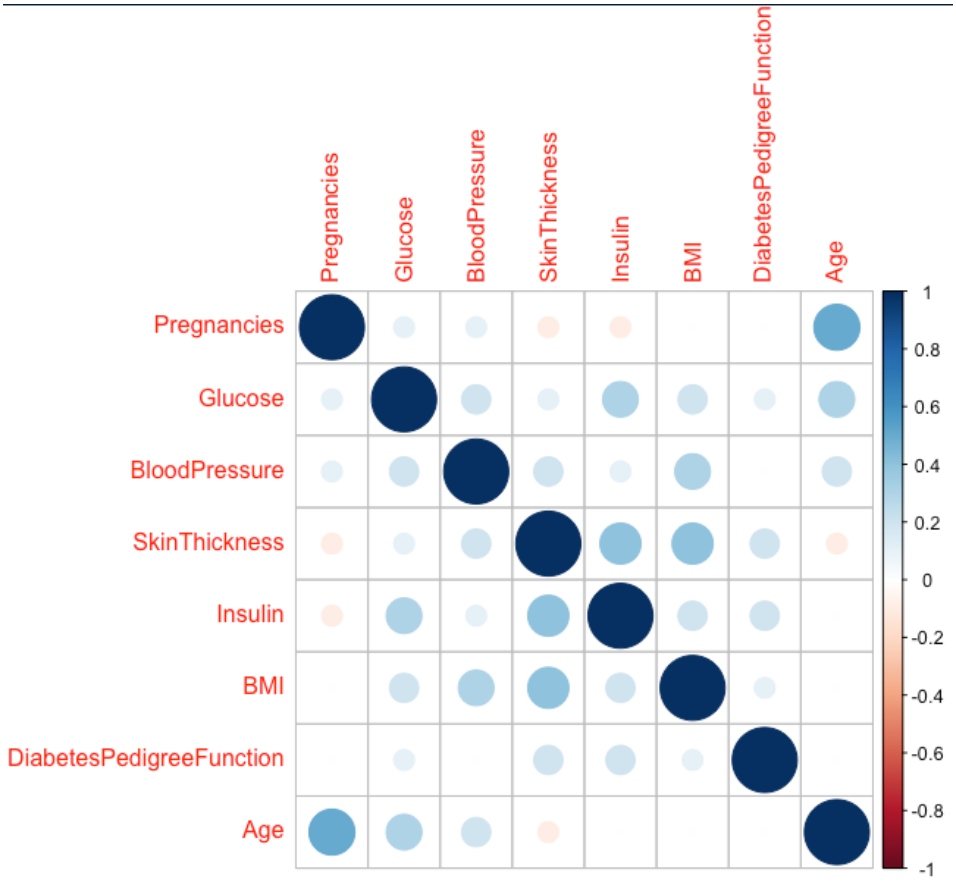


FIGURE 6 – Corrélation entre variables

Deuxième partie

Analyse Factorielle Discriminante

1 Théorie de la méthode probabiliste

2 ACP préliminaire (présentation du jeu de données)

3 Exemple sur R