

Analyse discriminante linéaire et quadratique

Marie Chavent

2015

On se place dans le cadre de la modélisation d'une variable Y **qualitative** à K modalités à partir de p variables explicatives X_1, \dots, X_p **quantitatives**. On suppose que l'on dispose d'un échantillon de taille n pour lequel les p variables explicatives et la variable à expliquer ont été mesurées simultanément. Généralement cet échantillon est appelé échantillon d'apprentissage. On veut définir à partir de cet échantillon d'apprentissage **une règle de classification** qui va permettre de prédire la valeur de Y pour un nouvel individu sur lequel on a mesuré uniquement les p variables explicatives. On parle de classification supervisée, chaque modalité de Y représentant une classe (un groupe) d'individus.

1 Les données

On dispose d'un échantillon de n observations de Y et de $X = (X_1, \dots, X_p)$: sur les n individus de l'échantillon, on a mesuré une variable qualitative à K modalités et p variables quantitatives. En notant y_i la valeur de la variable à expliquer mesurée sur les i ème individu, on obtient le vecteur $\mathbf{y} = (y_1, \dots, y_n)' \in \{1, \dots, K\}^n$. En notant x_{ij} la valeur de la j ème variable explicative mesurée sur le i ème individu, on obtient ainsi la matrice de données de dimension $n \times p$ suivante :

$$\mathbf{X} = \begin{array}{c|ccccccccc} & & & 1 & \dots & j & \dots & p \\ \hline 1 & & & & & & & & \\ \vdots & & & & & \vdots & & & \\ \mathbf{x} = & i & \dots & x_{ij} & \dots & & & & \\ \vdots & & & \vdots & & & & & \\ n & & & & & & & & \end{array}$$

Notons :

- $x_i = (x_{i1}, \dots, x_{ip})' \in \mathbb{R}^p$ une ligne de \mathbf{X} décrivant le i ème individu.
- $x^j = (x_{1j}, \dots, x_{nj})' \in \mathbb{R}^n$ une colonne de \mathbf{X} décrivant la j ème variable.
- E_k est le groupe des individus des l'échantillon qui possèdent la modalité k .
- $n_k = \text{card}(E_k)$ est le nombre d'individus qui possèdent la modalité k .

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

On a les relations suivantes classiques suivantes :

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

et

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

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

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

Exemple. On considère un jeu de données concernant $n = 101$ victimes d'infarctus du myocarde réparties en $K = 2$ groupes : le groupe G_1 des victimes décédées ($n_1 = 51$) et le groupe G_2 des victimes qui survivent ($n_2 = 50$). On veut construire un score de risque de décès applicable à une nouvelle victime. Pour chaque victime, 7 variables cliniques ont été mesurées :

- **FRCAR** : fréquence cardiaque,
- **INCAR** : index cardiaque,
- **INSYS** : index systolique,
- **PRDIA** : pression diastolique,
- **PAPUL** : pression artérielle pulmonaire,
- **PVENT** : pression ventriculaire,
- **REPUL** : resistance pulmonaire.

La variable qualitative (à expliquer) est donc la variable pronostics (1=décès et 2=survie).

```
load("infarctus.rda")
head(infarctus)

##      PRONO FRCAR INCAR INSYS PRDIA PAPUL PVENT REPUL
## 1 SURVIE    90  1.71  19.0    16  19.5  16.0   912
## 2 DECES     90  1.68  18.7    24  31.0  14.0  1476
## 3 DECES    120  1.40  11.7    23  29.0   8.0  1657
## 4 SURVIE    82  1.79  21.8    14  17.5  10.0   782
## 5 DECES     80  1.58  19.7    21  28.0  18.5  1418
## 6 DECES     80  1.13  14.1    18  23.5   9.0  1664
```

Dans un premier temps $p = 3$ variables cliniques ont été retenues pour construire une règle de classification : fréquence cardiaque, index systolique, pression diastolique.

```
X <- subset(infarctus,select=c(FRCAR,INSYS,PRDIA))
head(X)

##   FRCAR  INSYS PRDIA
## 1    90  19.0   16
## 2    90  18.7   24
## 3   120  11.7   23
## 4    82  21.8   14
## 5    80  19.7   21
## 6    80  14.1   18

y <-infarctus[,1]
head(y)

## [1] SURVIE DECES  DECES  SURVIE DECES  DECES
## Levels: DECES SURVIE

#Description des donnees
source("LDA_procedures_chavent.R")
printStatDes(X,y)

## Taille de l'ensemble des groupes :
## -----
## [1] 101
##
## Moyennes pour l'ensemble des groupes :
## -----
## FRCAR  INSYS PRDIA
##  92.2  20.8  19.3
##
## Matrice de variances pour l'ensemble des groupes :
## -----
##          FRCAR  INSYS PRDIA
## FRCAR  267.22 -72.04  37.68
## INSYS -72.04  76.90 -24.47
## PRDIA  37.68 -24.47  33.41
##
## Matrice de correlations pour l'ensemble des groupes :
## -----
##          FRCAR  INSYS PRDIA
## FRCAR    1.0 -0.50  0.40
## INSYS   -0.5  1.00 -0.48
## PRDIA    0.4 -0.48  1.00
##
## Taille de l'échantillon du groupe ' DECES ' :
## -----
## [1] 51
```

```

## 
## Moyennes pour le groupe ' DECES ' :
## -----
## FRCAR INSYS PRDIA
##  95.9 15.0 22.0
##
## Matrice de variances pour le groupe ' DECES ' :
## -----
##          FRCAR  INSYS PRDIA
## FRCAR 316.83 -44.29 31.15
## INSYS -44.29  21.10 -8.33
## PRDIA  31.15 -8.33 25.92
##
## Matrice de correlations pour le groupe ' DECES ' :
## -----
##          FRCAR  INSYS PRDIA
## FRCAR  1.00 -0.54  0.34
## INSYS -0.54  1.00 -0.36
## PRDIA  0.34 -0.36  1.00
##
## 
## Taille de l'echantillon du groupe ' SURVIE ' :
## -----
## [1] 50
##
## Moyennes pour le groupe ' SURVIE ' :
## -----
## FRCAR INSYS PRDIA
##  88.3 26.8 16.5
##
## Matrice de variances pour le groupe ' SURVIE ' :
## -----
##          FRCAR  INSYS PRDIA
## FRCAR 187.74 -55.45 23.50
## INSYS -55.45  64.03 -8.54
## PRDIA  23.50 -8.54 26.02
##
## Matrice de correlations pour le groupe ' SURVIE ' :
## -----
##          FRCAR  INSYS PRDIA
## FRCAR  1.00 -0.51  0.34
## INSYS -0.51  1.00 -0.21
## PRDIA  0.34 -0.21  1.00

#Matrice de variance-covariance intra-groupe
var.p <- function(x){var(x)*(nrow(x)-1)/nrow(x)}
Xk <- split(X,y)
Vk <- lapply(Xk,function(x){var.p(x)})
W <- (51*Vk$DECES+50*Vk$SURVIE)/101
W

```

```

##          FRCAR     INSYS     PRDIA
## FRCAR 252.928 -49.8159 27.3647
## INSYS -49.816  42.3523 -8.4329
## PRDIA  27.365  -8.4329 25.9709

```

2 Règle géométrique de classification

On présente d'abord une approche purement géométrique (sans aucune hypothèse probabiliste). Notons x le vecteur des valeurs des p variables explicatives sur un nouvel individu dont que l'on veut classer. La règle géométrique consiste à calculer la distance de x à chacun des K centres de gravité $\mathbf{g}_1, \dots, \mathbf{g}_k$ et à affecter x au **groupe le plus proche**. Pour cela, il faut préciser la métrique à utiliser dans le calcul des distances. La règle la plus utilisée est celle de Mahalanobis-Fisher qui consiste à prendre la métrique \mathbf{W}^{-1} (ou \mathbf{V}^{-1} ce qui est équivalent). La distance du nouvel individu au groupe k est alors :

$$d^2(x, \mathbf{g}_k) = (x - \mathbf{g}_k)' \mathbf{W}^{-1} (x - \mathbf{g}_k). \quad (3)$$

Fonctions linéaires discriminantes. La règle géométrique classe la nouvelle observation x dans le groupe k^* tel que :

$$k^* = \arg \min_{k=1, \dots, K} d^2(x, \mathbf{g}_k), \quad (4)$$

ce qui se réécrit :

$$k^* = \arg \max_{k=1, \dots, K} L_k(x), \quad (5)$$

où

$$L_k(x) = x' \mathbf{W}^{-1} \mathbf{g}_k - \frac{1}{2} \mathbf{g}_k' \mathbf{W}^{-1} \mathbf{g}_k. \quad (6)$$

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

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

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

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

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

Exemple. Fonctions linéaires discriminantes pour les données infarctus.

```
res <- linear_func(X,y,type="geom")
#fonctions lineaires discriminantes
res$Lk

##           DECES      SURVIE
## constant -39.80843 -47.12602
## FRCAR     0.51556   0.56141
## INSYS    1.08019   1.37460
## PRDIA    0.63636   0.47770

#prediction pour x=(90,19,16) : premiere ligne de X
c(1,90,19,16) %*% res$Lk

##           DECES      SURVIE
## [1,] 37.297 37.161
```

3 Le cadre probabiliste

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

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

On notera :

- (π_1, \dots, π_K) la distribution de Y où $\pi_k = P(Y = k)$ est la proportion théorique de G_k encore appelée **probabilité à priori** de G_k .
- $f_k : \mathbb{R}^p \rightarrow [0, 1]$ la densité de X dans le groupe k .
- la densité de X est une densité de mélange :

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

La règle de classement optimale de Bayes affecte une nouvelle observation x au **groupe le plus probable** sachant x :

$$k^* = \arg \max_{k=1,\dots,K} P(G_k | x), \quad (7)$$

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

$$k^* = \arg \max_{k=1,\dots,K} \pi_k f_k(x). \quad (8)$$

On parle alors d'**approche générative**.

Preuve : On utilise le théorème de Bayes qui donne :

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

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

■

L'apprentissage de cette règle de classification supposera que l'on dispose d'un échantillon $(X_1, Y_1), \dots, (X_n, Y_n)$ i.i.d. de même loi que (X, Y) . Plusieurs approches sont alors possibles.

- a) 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.
- b) On peut supposer que les variables X_1, \dots, X_p sont indépendantes conditionnellement à Y avec $f_k(x) = \prod_{j=1}^p f_{k,j}(x_j)$. Il s'agit alors d'estimer les densités $f_{k,j}(x)$ soit de manière non paramétrique par un histogramme ou un estimateur de densité à noyau, soit en supposant une forme paramétrique à $f_{k,j}(x)$ et en estimant ces paramètres sur l'échantillon d'apprentissage. C'est la méthode dite du Bayesien naïf.
- c) Les approches non paramétriques : on cherche à estimer directement à partir des données les densités f_k avec des méthodes d'estimation de densité à noyau.

Ici on se place dans le cadre paramétrique gaussien.

3.1 Le cas gaussien

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

$$f_k(x) = \frac{1}{(2\pi)^{p/2}(\det(\Sigma_k))^{1/2}} \exp\left(-\frac{1}{2}(x - \mu_k)' \Sigma_k^{-1} (x - \mu_k)\right), \quad (9)$$

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

Dans ce cas, la règle de Bayes se réécrit :

$$k^* = \arg \min_{k=1,\dots,K} D_k^2(x), \quad (10)$$

où

$$D_k^2(x) = (x - \mu_k)' \Sigma_k^{-1} (x - \mu_k) - 2 \ln(\pi_k) + \ln(\det \Sigma_k), \quad (11)$$

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

Preuve : Maximiser $\pi_k f_k(x)$ est équivalent à maximiser $\ln(\pi_k f_k(x))$ et

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

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

■

3.2 Estimation des paramètres

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

$$\theta = (\pi_1, \dots, \pi_K, \mu_1, \dots, \mu_K, \Sigma_1, \dots, \Sigma_K).$$

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

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

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

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

On obtient alors les estimateurs du maximum de vraisemblance suivant :

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

Ces estimateurs de Σ_k sont biaisés et on a les estimateurs sans biais suivants :

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

4 Analyse discriminante quadratique (QDA en anglais)

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

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

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

On obtient ainsi la **règle de classification d'analyse discriminante quadratique** :

$$k^* = \arg \min_{k=1,\dots,K} Q_k(x), \quad (12)$$

où

$$Q_k(x) = (x - \mathbf{g}_k)' \mathbf{V}_k^{-1} (x - \mathbf{g}_k) - 2 \ln(\pi_k) + \ln(\det(\mathbf{V}_k)), \quad (13)$$

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

5 Analyse discriminante linéaire (LDA en anglais)

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

$$k^* = \arg \max_{k=1,\dots,K} x' \Sigma^{-1} \mu_k - \frac{1}{2} \mu'_k \Sigma^{-1} \mu_k + \ln(\pi_k). \quad (14)$$

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

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

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

■

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

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

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

On obtient ainsi la **règle de classification d'analyse discriminante linéaire** :

$$k^* = \arg \max_{k=1,\dots,K} L_k(x), \quad (15)$$

où

$$L_k(x) = x' \mathbf{W}^{-1} \mathbf{g}_k - \frac{1}{2} \mathbf{g}'_k \mathbf{W}^{-1} \mathbf{g}_k + \ln(\hat{\pi}_k), \quad (16)$$

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. Lorsqu'on suppose que $\pi_1 = \dots = \pi_K$ (égalité des probabilités à priori), la règle de l'analyse discriminante linéaire (LDA) est équivalente à la règle de classement géométrique qui consiste à affecter une nouvelle observation x au groupe k dont le centre de gravité g_k est le plus proche en terme de distance. La distance utilisée est alors la distance de Mahalanobis correspondant à la métrique W^{-1} .

Exemple. Fonctions linéaires discriminantes pour les données infarctus.

```
linear_func(X,y,type="geom")$Lk

##           DECES      SURVIE
## constant -39.80843 -47.12602
## FRCAR     0.51556   0.56141
## INSYS    1.08019   1.37460
## PRDIA    0.63636   0.47770

linear_func(X,y,type="prob")$Lk

##           DECES      SURVIE
## constant -40.49172 -47.82912
## FRCAR     0.51556   0.56141
## INSYS    1.08019   1.37460
## PRDIA    0.63636   0.47770
```

Avec la fonction `lda` du package MASS.

```
library(MASS)
mod <- lda(X,y)
mod

## Call:
## lda(X, y)
##
## Prior probabilities of groups:
##   DECES      SURVIE
## 0.50495 0.49505
##
## Group means:
##          FRCAR  INSYS  PRDIA
## DECES  95.902 14.996 21.961
## SURVIE 88.340 26.752 16.504
##
## Coefficients of linear discriminants:
##          LD1
## FRCAR  0.022980
## INSYS  0.147572
## PRDIA -0.079528
```

```

predict(mod)$class[1:5]

## [1] DECES DECES DECES SURVIE DECES
## Levels: DECES SURVIE

predict(mod)$posterior[1:5,]

##      DECES SURVIE
## 1 0.53890 0.46110
## 2 0.81959 0.18041
## 3 0.88497 0.11503
## 4 0.35001 0.64999
## 5 0.76881 0.23119

```

Avec toutes les variables explicatives, un échantillon test et les probabilité à priori égales.

```

train <- sample(1:101,80)
mod <- lda(PRONO~.,infarctus,prior = c(1,1)/2,subset=train)
pred <- predict(mod)$class
length(pred)

## [1] 80

prob <- predict(mod)$posterior
dim(prob)

## [1] 80  2

```