

# Data Science for Actuaries (ACT6100)

Arthur Charpentier

Supervisé # 1 (Concepts Fondamentaux - 4)

automne 2020

 <https://github.com/freakonometrics/ACT6100/>

# Bayes Classifier

- **Bayes classifieur** est le modèle qui maximise la probabilité de classifier correctement une observation:

$$\hat{Y} = \operatorname{argmax}_{y \in \mathcal{Y}} \{ \mathbb{P}(Y = y | \mathbf{X} = \mathbf{x}) \} = m^*(\mathbf{x}).$$

- La distribution de  $Y$  n'est pas connue: il n'est pas possible, en pratique, d'utiliser le classifieur de Bayes (sans hypothèses supplémentaires).
- On a

$$\begin{aligned} \mathbb{P}(\hat{Y} \neq Y) &= \mathbb{E}[\mathbb{P}(\hat{Y} \neq Y | \mathbf{X})] \\ &= \mathbb{E}\left[1 - \max_{g \in \mathcal{G}} \mathbb{P}(Y = g | \mathbf{X})\right] \\ &= 1 - \mathbb{E}\left[\max_{g \in \mathcal{G}} \mathbb{P}(Y = g | \mathbf{X})\right]. \end{aligned}$$

- Le classifieur de Bayes est le "meilleur" possible, son taux d'erreur est une borne minimale.

# Bayes Classifier

- ▶ the **classification risk** (or **error rate**) of  $m$  is

$$\mathcal{R}(m) = \mathbb{P}[m(\mathbf{X}) \neq Y]$$

- ▶ the **empirical classification risk** (or **training error rate**) of  $m$  is

$$\hat{\mathcal{R}}(m) = \frac{1}{n} \sum_{i=1}^n \mathbf{1}(m(\mathbf{x}_i) \neq y_i)$$

- ▶ the classifier that minimizes  $\mathcal{R}$  is Bayes classifier  $m^*$

**Proof:** let us proof that  $\mathcal{R}(m) - \mathcal{R}(m^*) \geq 0$

$$\mathcal{R}(m) = \int \mathbb{P}[m(\mathbf{X}) \neq Y | \mathbf{X} = \mathbf{x}] d\mathbb{P}_{\mathbf{X}}(\mathbf{x})$$

$$\mathbb{P}[m(\mathbf{X}) \neq Y | \mathbf{X} = \mathbf{x}] = 1 - \mathbb{P}[m(\mathbf{X}) = Y | \mathbf{X} = \mathbf{x}]$$

# Bayes Classifier

and  $\mathbb{P}[m(\mathbf{X}) = Y|\mathbf{x}]$  can be written

$$\mathbb{P}[m(\mathbf{X}) = 1|\mathbf{x}]\mathbb{P}[Y = 1|\mathbf{x}] + \mathbb{P}[m(\mathbf{X}) = 0|\mathbf{x}]\mathbb{P}[Y = 0|\mathbf{x}]$$

Let  $r(\mathbf{x}) = \mathbb{E}[Y|\mathbf{X} = \mathbf{x}]$ , so that

$$\mathbb{P}[m(\mathbf{X}) \neq Y|\mathbf{X} = \mathbf{x}] = 1 - [r(\mathbf{x})m(\mathbf{x}) + (1 - r(\mathbf{x}))(1 - m(\mathbf{x}))]$$

and  $\mathbb{P}[m(\mathbf{X}) \neq Y|\mathbf{X} = \mathbf{x}] - \mathbb{P}[m^*(\mathbf{X}) \neq Y|\mathbf{X} = \mathbf{x}]$  is equal to

$$2\left[r(\mathbf{x}) - \frac{1}{2}\right][m^*(\mathbf{x}) - m(\mathbf{x})], \quad \text{with } m^*(\mathbf{x}) = \mathbf{1}\left(r(\mathbf{x}) \geq \frac{1}{2}\right)$$

so when  $r(\mathbf{x}) \geq 1/2$ ,  $m^*(\mathbf{x}) - m(\mathbf{x}) = 1 - m(\mathbf{x}) \geq 0$

when  $r(\mathbf{x}) < 1/2$ ,  $m^*(\mathbf{x}) - m(\mathbf{x}) = -m(\mathbf{x}) \leq 0$

# Bayes Classifier

- ▶ let  $\pi_y = \mathbb{P}[Y = y]$ , so that we can write

$$m^*(\mathbf{x}) = \mathbf{1}\left(r(\mathbf{x}) \geq \frac{1}{2}\right) = \mathbf{1}\left(\frac{\mathbb{P}[\mathbf{X} = \mathbf{x} | Y = 1]}{\mathbb{P}[\mathbf{X} = \mathbf{x} | Y = 0]} > \frac{1 - \pi_1}{\pi_1}\right)$$

# Oracle Classifier

- ▶ let  $\mathcal{M}$  denote a set of classifiers,

$$m_0 = \operatorname{argmin}_{m \in \mathcal{M}} \{ \mathcal{R}(m) \}$$

is called the **oracle classifier**

- ▶ For any  $m \in \mathcal{M}$ ,

$$\mathcal{R}(m) - \mathcal{R}(m^*) = \underbrace{\mathcal{R}(m) - \mathcal{R}(m_0)} + \underbrace{\mathcal{R}(m_0) - \mathcal{R}(m^*)}$$

# Analyse discriminante de Fisher

On dispose de données en dimension 2  $\{\mathbf{x}_1, \dots, \mathbf{x}_n\}$  dans deux classes,  $y_i \in \{0, 1\}$ . On suppose que  $\mathbf{X}|Y = 0 \sim \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$  et  $\mathbf{X}|Y = 1 \sim \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$

$$\mathbb{P}(Y = y|\mathbf{X} = \mathbf{x}) \propto f_y(\mathbf{x}) \cdot \mathbb{P}(Y = y)$$

de telle sorte que  $\log \mathbb{P}(Y = y|\mathbf{X} = \mathbf{x})$  vaut

$$-\frac{1}{2} \log |\boldsymbol{\Sigma}_y| - \frac{1}{2} [\mathbf{x} - \boldsymbol{\mu}_y]^\top \boldsymbol{\Sigma}_y^{-1} [\mathbf{x} - \boldsymbol{\mu}_y] + \log \mathbb{P}(Y = y)$$

# Analyse discriminante de Fisher

Soit  $\delta_y$  la fonction définie (pour  $y \in \{0, 1\}$ ) par

$$\delta_y(\mathbf{x}) = -\frac{1}{2} \log |\boldsymbol{\Sigma}_y| - \frac{1}{2} [\mathbf{x} - \boldsymbol{\mu}_y]^\top \boldsymbol{\Sigma}_y^{-1} [\mathbf{x} - \boldsymbol{\mu}_y] + \log \mathbb{P}(Y = y)$$

La frontière de décision,  $\{\mathbf{x} : \delta_0(\mathbf{x}) = \delta_1(\mathbf{x})\}$  est **quadratique en  $\mathbf{x}$**   
**Fisher (1936)** a rajouté hypothèse  $\boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}_1$ . Alors

$$\delta_y(\mathbf{x}) = \mathbf{x}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_y - \frac{1}{2} \boldsymbol{\mu}_y^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_y + \log \mathbb{P}(Y = y)$$

et la frontière de décision est **linéaire en  $\mathbf{x}$** .

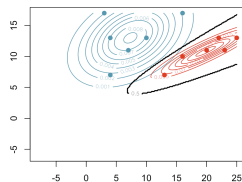
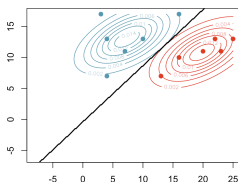
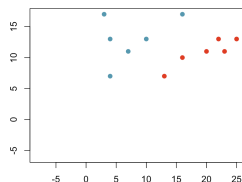
Sur la frontière,  $\mathbf{x}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_1 = \mathbf{x}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_2$  i.e.

$$\text{constant} + \underbrace{\mathbf{x}^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)}_{\vec{u}} = 0$$

qui est un hyperplan de vecteur normal  $\vec{u} = \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)$ .



# Analyse discriminante de Fisher



Si  $\mathbf{X}|Y = 0 \sim \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma})$  et  $\mathbf{X}|Y = 1 \sim \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$  alors

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

est égal à

$$\mathbf{x}^\top \boldsymbol{\Sigma}^{-1}[\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0] - \frac{1}{2}[\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0]^\top \boldsymbol{\Sigma}^{-1}[\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0] + \log \frac{\mathbb{P}(Y = 1)}{\mathbb{P}(Y = 0)}$$

qui est linéaire en  $\mathbf{x}$ , autrement dit

$$\log \frac{\mathbb{P}(Y = 1|\mathbf{X} = \mathbf{x})}{\mathbb{P}(Y = 0|\mathbf{X} = \mathbf{x})} = \mathbf{x}^\top \boldsymbol{\beta}$$

ce qui rappelle la régression logistique...

# Analyse discriminante de Fisher

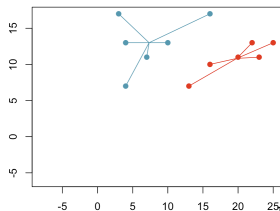
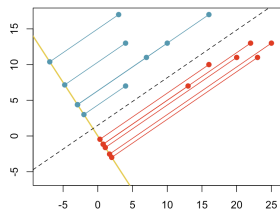
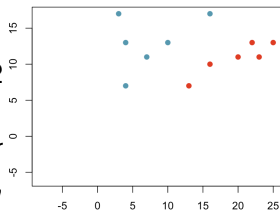
On dispose de données en dimension 2  $\{\mathbf{x}_1, \dots, \mathbf{x}_n\}$  dans deux classes,  $y_i \in \{0, 1\}$ .

On va projeter les deux nuages de points sur une droite, de direction  $\vec{u}$ , avec  $\|\vec{u}\| = 1$ ,  $\{z_1, \dots, z_n\}$ . La distance de  $z_i$  à 0 est  $\vec{u}^\top \mathbf{x}_i$ . On définit les centroïdes des deux groupes,

$$\bar{\mathbf{x}}^y = \frac{1}{n_y} \sum_{i:y_i=y} \mathbf{x}_i$$

et les variances intra (des  $\mathbf{x}$ )

$$\hat{\mathbf{s}}^y = \frac{1}{n_y} \sum_{i:y_i=y} (\mathbf{x}_i - \bar{\mathbf{x}}^y)(\mathbf{x}_i - \bar{\mathbf{x}}^y)^\top$$



# Analyse discriminante de Fisher

On définit

$$\hat{\mathbf{s}}_{\text{within}} = \frac{n_0 \hat{\mathbf{s}}^0 + n_1 \hat{\mathbf{s}}^1}{n}$$

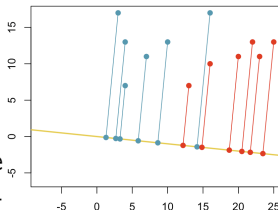
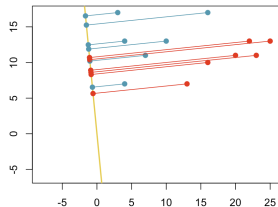
et

$$\hat{\mathbf{s}}_{\text{between}} = (\bar{\mathbf{z}}^0 - \bar{\mathbf{z}}^1)(\bar{\mathbf{z}}^0 - \bar{\mathbf{z}}^1)^\top$$

On cherche alors à maximiser

$$f(\vec{\mathbf{u}}) = \frac{\vec{\mathbf{u}}^\top \hat{\mathbf{s}}_{\text{b}} \vec{\mathbf{u}}}{\vec{\mathbf{u}}^\top \hat{\mathbf{s}}_{\text{w}} \vec{\mathbf{u}}}$$

On cherche alors la direction qui maximise le ratio entre la variance inter et intra de la projection



# Analyse discriminante de Fisher

the first order condition is

$$\frac{df(\vec{u})}{d\vec{u}} = \frac{(2\hat{\mathbf{s}}_b\vec{u})\vec{u}^\top\hat{\mathbf{s}}_w\vec{u} - (2\hat{\mathbf{s}}_w\vec{u})\vec{u}^\top\hat{\mathbf{s}}_b\vec{u}}{(\vec{u}^\top\hat{\mathbf{s}}_w\vec{u})^2} = \vec{0}$$

i.e.

$$(2\hat{\mathbf{s}}_b\vec{u})\vec{u}^\top\hat{\mathbf{s}}_w\vec{u} - (2\hat{\mathbf{s}}_w\vec{u})\vec{u}^\top\hat{\mathbf{s}}_b\vec{u} = \vec{0}$$

can be written

$$\hat{\mathbf{s}}_b\vec{u} - \underbrace{\frac{\vec{u}^\top\hat{\mathbf{s}}_b\vec{u}}{\vec{u}^\top\hat{\mathbf{s}}_w\vec{u}}}_{=\lambda}\hat{\mathbf{s}}_w\vec{u} = \vec{0}$$

If  $\hat{\mathbf{s}}_w$  is full-rank matrix,

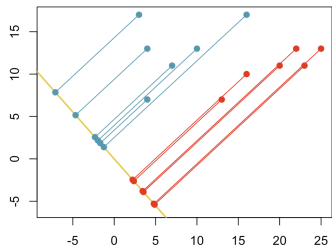
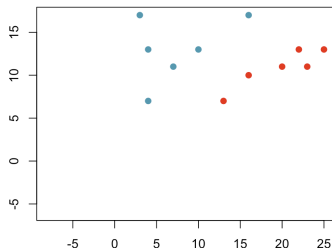
$$\hat{\mathbf{s}}_w^{-1}\hat{\mathbf{s}}_b\vec{u} = \lambda\vec{u}$$

and furthermore, since  $\hat{\mathbf{s}}_b\vec{u} = \alpha(\bar{\mathbf{x}}^0 - \bar{\mathbf{x}}^1)$ , we get that

$$\vec{u} \propto \hat{\mathbf{s}}_w^{-1}(\bar{\mathbf{x}}^0 - \bar{\mathbf{x}}^1)$$

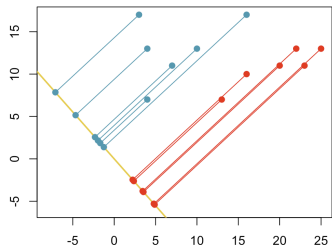
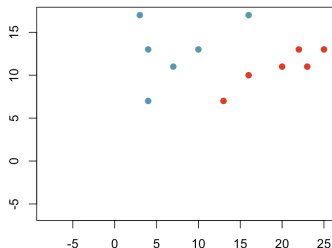
# Analyse discriminante de Fisher

```
1 > b0=data.frame(  
2   x=c(4,4,3,7,10,16),  
3   y=c(7,13,17,11,13,17))  
4 > b1=data.frame(  
5   x=c(13,16,20,23,22,25),  
6   y=c(7,10,11,11,13,13))  
7 > b=rbind(b0,b1)  
8 > plot(b,col=rep(colr2,each=6),  
9       pch=19)  
9 > m0=apply(b0,2,mean)  
10 > m0=t(rep(1,6))%*%as.matrix(b0)/  
11   nrow(b0)  
11 > m1=apply(b1,2,mean)  
12 > centX0=as.matrix(b0)-rep(m0,  
13   each=nrow(b0))  
12 > centX1=as.matrix(b1)-rep(m1,  
13   each=nrow(b1))
```



# Analyse discriminante de Fisher

```
1 > S0 = t(centX0)%*%centX0
2 > S1 = t(centX1)%*%centX1
3 > Sw = S0+S1
4 > Sw
5           x           y
6 x 226.16667  83.83333
7 y  83.83333  96.83333
8 > Sb = t(m0-m1)%*%(m0-m1)
9 > Sb
10          x           y
11 x 156.25000 -27.083333
12 y -27.08333   4.694444
13 > u = solve(Sw)%*%t(m0-m1)
14 > u
15           [,1]
16 x -0.09359965
17 y  0.10340899
```



# Analyse discriminante de Fisher

```
1 > myocarde=read.table("http://freakonometrics.free.fr/
  saporta.csv", header=TRUE, sep=";")
2 > levels(myocarde$PRONO)=c("0", "1")
3 > m0 = apply(myocarde[myocarde$PRONO=="0", 1:7], 2, mean)
4 > m1 = apply(myocarde[myocarde$PRONO=="1", 1:7], 2, mean)
5 > Sigma = var(myocarde[, 1:7])
6 > omega = solve(Sigma)%*%(m1-m0)
7 > omega
8   FRCAR  INCAR   INSYS  PRDIA  PAPUL   PVENT   REPUL
9 -0.013  1.089 -0.019 -0.026   0.02 -0.038 -0.001

1 > library(MASS)
2 > fit_lda = lda(PRONO ~. , data=myocarde)
3
4 Prior probabilities of groups:
5           0           1
6 0.4084507 0.5915493
7
8 Coefficients of linear discriminants:
9   FRCAR  INCAR   INSYS  PRDIA  PAPUL   PVENT   REPUL
10 -0.013  1.089 -0.019 -0.026   0.02 -0.038 -0.001
```

# Analyse discriminante de Fisher

Le classifieur de Fisher est

$$\hat{m}(\mathbf{x}) = \begin{cases} 1 & \text{si } \mathbf{u}_n^\top \mathbf{x} < c \\ 0 & \text{si } \mathbf{u}_n^\top \mathbf{x} \geq c \end{cases}$$

où  $\mathbf{u}_n = \hat{\mathbf{s}}_w^{-1}(\bar{\mathbf{x}}^0 - \bar{\mathbf{x}}^1)$  et

$$c = \frac{1}{2}(\bar{\mathbf{x}}^0 - \bar{\mathbf{x}}^1)^\top \hat{\mathbf{s}}_w^{-1}(\bar{\mathbf{x}}^0 - \bar{\mathbf{x}}^1) - \log \frac{1 - \hat{\pi}_1}{\hat{\pi}_1}$$

**Note:** inférence en maximisant la vraisemblance globale

$$\prod_{i=1}^n f(\mathbf{x}_i, y_i) = \underbrace{\prod_{i=1}^n f(\mathbf{x}_i | y_i)}_{\text{Gaussien}} \underbrace{\prod_{i=1}^n \mathbb{P}(y_i)}_{\text{Bernoulli}}$$



# Régression logistique

Un autre classifieur classique est celui de la régression logistique

$$\hat{m}(\mathbf{x}) = \begin{cases} 1 & \text{si } \hat{r}_n(\mathbf{x}) > 1/2 \\ 0 & \text{si } \hat{r}_n(\mathbf{x}) \leq 1/2 \end{cases} \quad \text{où } \hat{r}_n(\mathbf{x}) = \frac{\exp[\hat{\beta}_0 + \mathbf{x}^\top \hat{\beta}]}{1 + \exp[\hat{\beta}_0 + \mathbf{x}^\top \hat{\beta}]}$$

$$\hat{m}(\mathbf{x}) = \begin{cases} 1 & \text{si } \hat{\beta}_n \mathbf{x} > \gamma \\ 0 & \text{si } \hat{\beta}_n \mathbf{x} \leq \gamma \end{cases}$$

**Note:** inférence en maximisant la vraisemblance conditionnelle

$$\prod_{i=1}^n f(\mathbf{x}_i, y_i) = \underbrace{\prod_{i=1}^n \mathbb{P}(y_i | \mathbf{x}_i)}_{\text{Logistique}} \underbrace{\prod_{i=1}^n f(\mathbf{x}_i)}_{\text{ignoré}}$$

**Note:** autres classifieurs linéaires: LASSO et SVM (à suivre...)