Data Science for Actuaries (ACT6100)

Arthur Charpentier

Supervisé # 1 (Concepts Fondamentaux - 4)

automne 2020

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

Bayes Classifier

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

$$\widehat{Y} = \underset{y\in\mathcal{Y}}{\text{argmax}}\big\{\mathbb{P}(Y=y|\mathbf{X}=\mathbf{x})\big\} = m^\star(\mathbf{x}).$$

- 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

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

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

Bayes Classifier

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

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

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

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

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

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

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

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

Bayes Classifier

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

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

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

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

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

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

so when $r(\boldsymbol{x}) \geq 1/2$, $m^\star(\boldsymbol{x}) - m(\boldsymbol{x}) = 1 - m(\boldsymbol{x}) \geq 0$
when $r(\boldsymbol{x}) < 1/2$, $m^\star(\boldsymbol{x}) - m(\boldsymbol{x}) = -m(\boldsymbol{x}) \leq 0$

Bayes Classifier

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

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

Oracle Classifier

- let $\mathcal{M}$ denote the set of all classifiers,

$$m_0 = \underset{m \in \mathcal{M}}{\text{argmin}} \big\{ \mathcal{R}(m) \big\}$$

  is called the oracle classifier
- For any $m \in \mathcal{M}$,

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

Analyse discriminante de Fisher

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

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

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

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

Analyse discriminante de Fisher

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

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

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

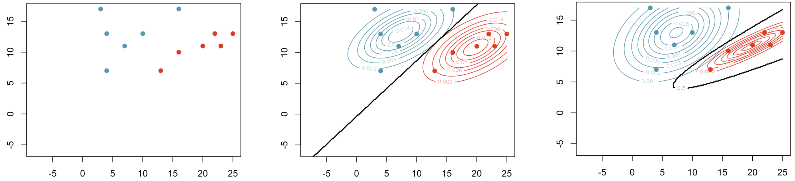
$$\delta_y(\boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_y - \frac{1}{2}\boldsymbol{\mu}_y^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_y + \log \mathbb{P}(Y=y)$$

et la frontière de décision est linéaire en $\boldsymbol{x}$.
Sur la frontière, $\boldsymbol{x}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1 = \boldsymbol{x}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2$ i.e.

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

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

Analyse discriminante de Fisher



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

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

est égal à

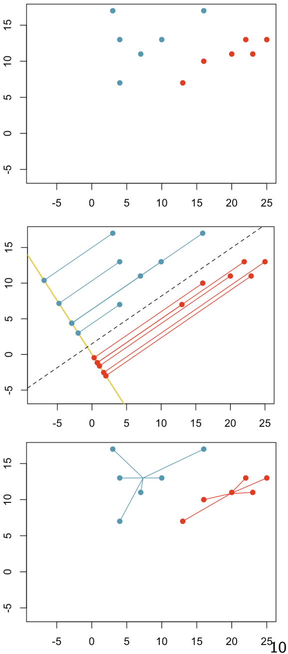
$$\boldsymbol{x}^\top \boldsymbol{\Sigma}^{-1}[\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0] - \frac{1}{2}[\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0]^\top \boldsymbol{\Sigma}^{-1}[\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0] + \log \frac{\mathbb{P}(Y = 1)}{\mathbb{P}(Y = 0)}$$

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

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

ce qui rappelle la régression logistique...

Analyse discriminante de Fisher



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

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

On définie les centroïdes des deux groupes,

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

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

$$\widehat{\boldsymbol{s}}^y = \frac{1}{n_y} \sum_{i:y_i=y} (\boldsymbol{x}_i - \overline{\boldsymbol{z}}^y)(\boldsymbol{x}_i - \overline{\boldsymbol{x}}^y)^\top$$

Analyse discriminante de Fisher

On définit

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

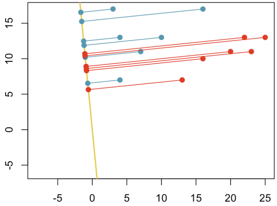
et

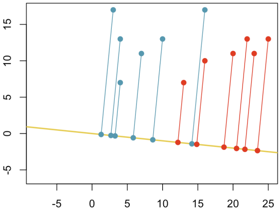
$$\widehat{\boldsymbol{s}}_{\text{between}} = (\overline{\boldsymbol{z}}^0 - \overline{\boldsymbol{z}}^1)(\overline{\boldsymbol{z}}^0 - \overline{\boldsymbol{z}}^1)^\top$$

On cherche alors à maximiser

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

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

$$\frac{df(\vec{\boldsymbol{u}})}{d\vec{\boldsymbol{u}}} = \frac{(2\widehat{\boldsymbol{s}}_{\text{b}}\vec{\boldsymbol{u}})\vec{\boldsymbol{u}}^{\top}\widehat{\boldsymbol{s}}_{\text{w}}\vec{\boldsymbol{u}} - (2\widehat{\boldsymbol{s}}_{\text{w}}\vec{\boldsymbol{u}})\vec{\boldsymbol{u}}^{\top}\widehat{\boldsymbol{s}}_{\text{b}}\vec{\boldsymbol{u}}}{(\vec{\boldsymbol{u}}^{\top}\widehat{\boldsymbol{s}}_{\text{w}}\vec{\boldsymbol{u}})^2} = \vec{\boldsymbol{0}}$$

i.e.

$$(2\widehat{\boldsymbol{s}}_{\text{b}}\vec{\boldsymbol{u}})\vec{\boldsymbol{u}}^{\top}\widehat{\boldsymbol{s}}_{\text{w}}\vec{\boldsymbol{u}} - (2\widehat{\boldsymbol{s}}_{\text{w}}\vec{\boldsymbol{u}})\vec{\boldsymbol{u}}^{\top}\widehat{\boldsymbol{s}}_{\text{b}}\vec{\boldsymbol{u}} = \vec{\boldsymbol{0}}$$

can be written

$$\widehat{\boldsymbol{s}}_{\text{b}}\vec{\boldsymbol{u}} - \underbrace{\frac{\vec{\boldsymbol{u}}^{\top}\widehat{\boldsymbol{s}}_{\text{b}}\vec{\boldsymbol{u}}}{\vec{\boldsymbol{u}}^{\top}\widehat{\boldsymbol{s}}_{\text{w}}\vec{\boldsymbol{u}}}}_{=\lambda}\widehat{\boldsymbol{s}}_{\text{w}}\vec{\boldsymbol{u}} = \vec{\boldsymbol{0}}$$

If $\widehat{\boldsymbol{s}}_{\text{w}}$ is full-rank matrix,

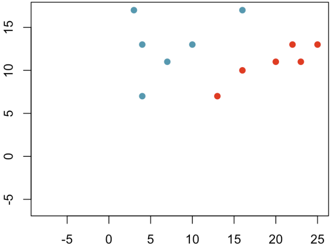
$$\widehat{\boldsymbol{s}}_{\text{w}}^{-1}\widehat{\boldsymbol{s}}_{\text{b}}\vec{\boldsymbol{u}} = \lambda\vec{\boldsymbol{u}}$$

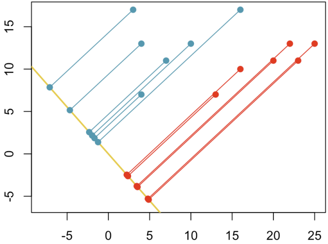
and furthermore, since $\widehat{\boldsymbol{s}}_{\text{b}}\vec{\boldsymbol{u}} = \alpha(\overline{\boldsymbol{x}}^0 - \overline{\boldsymbol{x}}^1)$, we get that

$$\vec{\boldsymbol{u}} \propto \widehat{\boldsymbol{s}}_{\text{w}}^{-1}(\overline{\boldsymbol{x}}^0 - \overline{\boldsymbol{x}}^1)$$

Analyse discriminante de Fisher

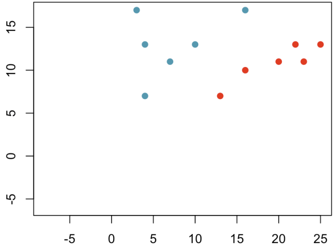
```
> b0=data.frame(
  x=c(4,4,3,7,10,16),
  y=c(7,13,17,11,13,17))
> b1=data.frame(
  x=c(13,16,20,23,22,25),
  y=c(7,10,11,11,13,13))
> b=rbind(b0,b1)
> plot(b,col=rep(colr2,each=6),
    pch=19)
> m0=apply(b0,2,mean)
> m0=t(rep(1,6))%*%as.matrix(b0)/
    nrow(b0)
> m1=apply(b1,2,mean)
> centX0=as.matrix(b0)-rep(m0,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

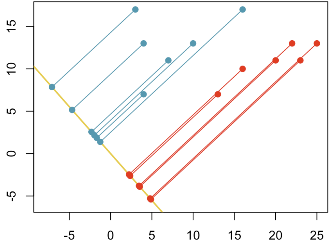




Analyse discriminante de Fisher

```
> S0 = t(centX0)%*%centX0
> S1 = t(centX1)%*%centX1
> Sw = S0+S1
> Sw
           x        y
x 226.16667 83.83333
y  83.83333 96.83333
> Sb = t(m0-m1)%*%(m0-m1)
> Sb
           x          y
x 156.25000 -27.083333
y -27.08333   4.694444
> u  = solve(Sw)%*%t(m0-m1)
> u
          [,1]
x -0.09359965
y  0.10340899
```





Analyse discriminante de Fisher

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

```
> library(MASS)
> fit_lda = lda(PRONO ~. , data=myocarde)

Prior probabilities of groups:
        0         1
0.4084507 0.5915493

Coefficients of linear discriminants:
 FRCAR INCAR  INSYS  PRDIA PAPUL  PVENT  REPUL
-0.013 1.089 -0.019 -0.026  0.02 -0.038 -0.001
```

Analyse discriminante de Fisher

Le classifieur de Fisher est

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

où $\boldsymbol{u}_n = \widehat{\boldsymbol{s}}_w^{-1}(\overline{\boldsymbol{x}}^0 - \overline{\boldsymbol{x}}^1)$ et

$$c = \frac{1}{2}(\overline{\boldsymbol{x}}^0 - \overline{\boldsymbol{x}}^1)^\top \widehat{\boldsymbol{s}}_w^{-1}(\overline{\boldsymbol{x}}^0 - \overline{\boldsymbol{x}}^1) - \log \frac{1 - \widehat{\pi}_1}{\widehat{\pi}_1}$$

Note: inférence en maximisant la vraisemblance globale

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

Régression logistique

Un autre classifieur classique est celui de la régression logistique

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

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

Note: inférence en maximisant la vraisemblance conditionnelle

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

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