CH3 Scroing_appli

Youming

2023-09-21

CH3 Evamuation d'un score

1. Context et objectif

$x1, x2...x_n, \in X$ sont des abservation répartie en Group $y_I \in \{1L..G\}$ indique le groupe auquel appartient n_i

2. Eurreur de classement

- Le score s est de bonne qualité si la classe $\hat{y}(x)$ à la quelle il affecte x correspond, en moyenne, à la vraie classe y de x. , autrement dire : S est un bon score si on moyenne le classe estimée de x : $\hat{y}(x)$coinside avec y 若 $\hat{y}$ 与实际 y 符合，则 s 质量好
- Or, si l'on ne sait rien du modèle qui a généré (x,y), l'erreur de classement : $E(x,y)(1_{\{\hat{y}(x)=y\}})$ ne peut pas être déterminée. Il existe cependant plusieurs façons de l'estimer et d'évaluer ainsi la qualité de s. 不过，若不知道生成 (x,y) 的模型，eurreur de classement **autrement dire : on souhaite que l'erreur de classement soit minimale.** 就不能被定论, 好在有以下方法可以评估 s $\varepsilon = E[1\hat{y}(x) \neq y)]$

Deux problèmes :

i) On ne connait pas la loi du couple (x, y) ; on ne peut pas calculer ε

ii) quand on a vrais loi du couple (x, y), ε dépend du classefier à traces $\hat{y}(x)$ et on ne peut pas la rendre arbitrairment proche de 0

Pour(ii) , il n'y a rien à faire, pour (i) on dispose plusieurs de méthode d'estimation

Eurreur théorique

En analyse dicriminante probabiliste, on modélise les donées du groupe g par $f_g(\bullet, \theta_g)$ et le poids du groupe est π_g. L'erreur théorique est alors :

$$\hat{\epsilon}_{th} = \sum_{g=1}^{G} \hat{\pi}_g \int_{\{\infty \in X; \hat{y}(\infty) \neq g\}} f_g(x, \hat{\theta}_g) dx$$

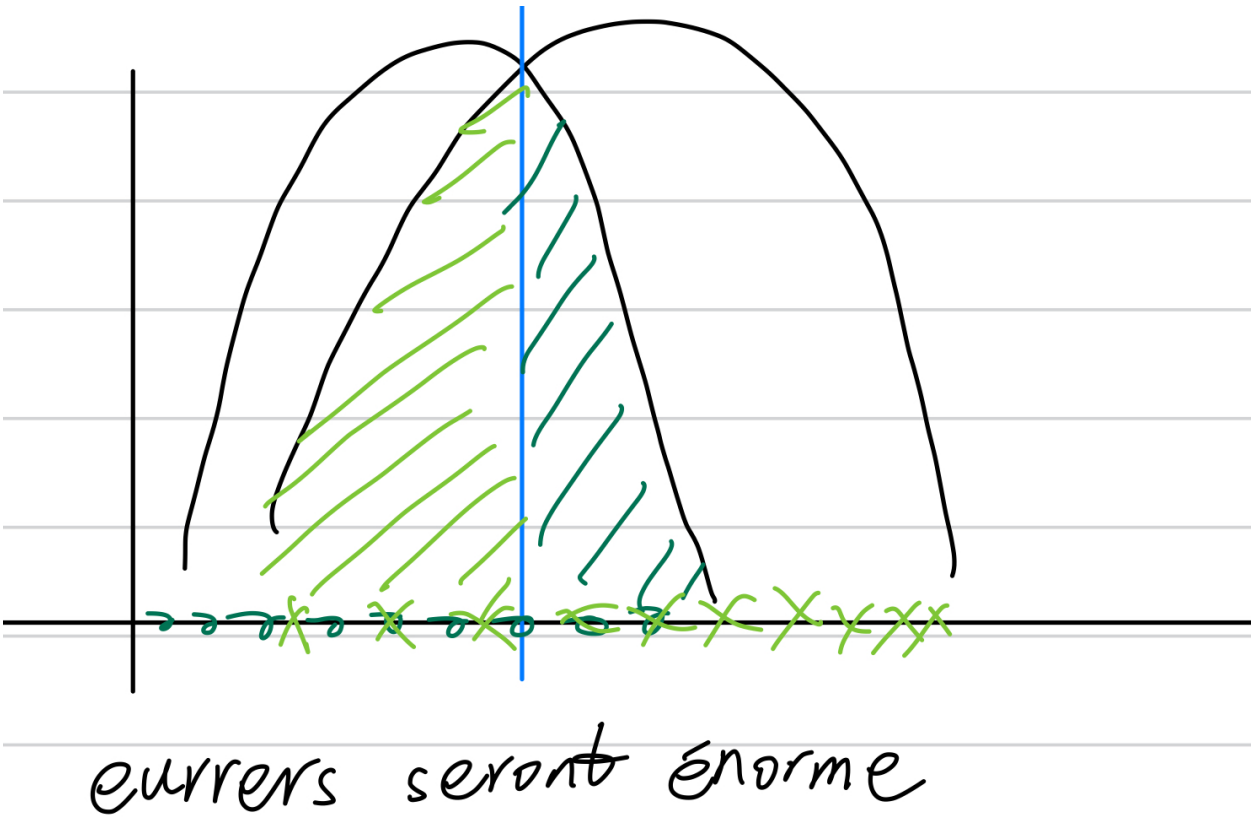


图 1: Erreur sera enorme

Justification : Voir CH2

L'estimateur $\hat{\varepsilon}$ possède deux inconvénients :

概率建模：在概率判别分析中，我们试图使用条件概率密度函数 $f_g(\bullet, \theta_g)$，对属于不同组（通常表示为 G）的数据进行建模。每个组都有一个权重 π_g，代表属于该组的数据比例。

(a) il hérite du biais du modèle postulé

(b) il est fréquent que les finctions f_g ne permettent pas de calculer $\hat{\varepsilon}_{th}$ explicitement [1]

理论误差：理论误差 $\hat{\epsilon}_{th}$ 是理论分类误差的度量。它被定义为每个组 g 的分类误差之和。总和的每个项都对应于密度函数 $f_g(x, \hat{\theta}_g)$ 在数据集上的积分，在数据集上，模型错误地分配了相关的组 g，即 $\hat{y}(\infty) \neq g$。$\hat{\varepsilon}$ 估计器的缺点：

(a) 假设模型的偏差：$\hat{\epsilon}_{th}$ 估计器有一个主要缺点。**它继承了假定模型的偏差**，这意味着如果用于估计 f_g 密度的基本概率模型有偏差或不正确，那么 $\hat{\epsilon}_{th}$ 也会有偏差，无法准确评估真实的分类误差。

(b) 计算困难：**密度函数 f_g 通常不允许明确计算** $\hat{\epsilon}_{th}$。这意味着，在许多情况下，无法通过分析确定理论误差，从而限制了准确估计分类误差的能力。

总之，摘录解释了如何利用各组的条件概率密度来模拟理论误差，但指出由于假设模型的原因，这种估计方法可能存在偏差，而且由于密度函数的复杂性，在实践中可能难以计算。这是概率判别分析中常见的难题。

En probabiliste , les erreur de group G sont modélisés par la fonction de probabilité f_g(*; _g) est on note $T_i g$ le poids de la classe g

---

Excercice 1 TD3 : [2]

On envisager deux modèle : le classifier Gaussien Hétéroscédastique ($\hat{\sigma}_1^2 \neq \hat{\sigma}_2^2$)

Affectation de M. Li (découvert = 1.9)

Poid de classe: $\hat{\pi}_1 = \hat{\pi}_2 = 0.5$

Centre de classe: $\hat{\mu}_1 = 1,5$ et $\hat{\mu}_2 = 2,5$

Variance: $\hat{\sigma}_1^2 = 3$ et $\hat{\sigma}_2^2 = 3,2$

x = décoissant

y = classe

两组分布表达式:

[1] méthode Monte Carlo, 随机抽样统计方法

[2] Rappelle : Loi normale $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

$$(x|y=1) \sim f(x;\hat{\mu}_1,\hat{\sigma}_1^2) = \frac{1}{\sqrt{2\pi}}\frac{1}{\hat{\sigma}_1}\exp\left(-\frac{1}{2}\left(\frac{x-\hat{\mu}_1}{\hat{\sigma}_1}\right)^2\right)$$

$$(x|y=2) \sim f(x;\hat{\mu}_2,\hat{\sigma}_2^2) = \frac{1}{\sqrt{2\pi}}\frac{1}{\hat{\sigma}_2}\exp\left(-\frac{1}{2}\left(\frac{x-\hat{\mu}_2}{\hat{\sigma}_2}\right)^2\right)$$

概率表达式:

$$P(Li \in Class1|x(Li)=1,9) = \frac{\hat{\Pi}_1 f_1(1,9;\hat{\mu};\hat{\sigma}_1^2)}{\sum_{j=1}^{2}\hat{\Pi}_j f(1,9;\hat{\mu}_j,\hat{\sigma}_j^2)} = \frac{\text{组 1}}{\text{两组}}$$

Sur R on a

```
library(stats)
pi1 = 0.5; pi2 = 0.5
mu1 = 1.5; mu2 = 2.5
sicar1 = 3; sicar2 = 3.2
pf1 = pi1 * dnorm(1.9, mean = mu1, sd = sqrt(sicar1))
pf2 = pi2 * dnorm(1.9, mean = mu2, sd = sqrt(sicar2))
t1 = pf1/(pf1+pf2)
print(t1)
```

```
## [1] 0.5154582
```

M.Li a 52% de class apportement à la classe 1, on affect à la class 1

Estimation de l'erreur de classement :

On détermine la fontiere de classement γ en réselment <12>

$$t1(r) = t2(r)$$

$$\hat{\pi}_1 \times f_1(r;\hat{\mu}_1,\hat{\sigma}_1^2) = \hat{\pi}_2 \times f_2(r;\hat{\mu}_2,\hat{\sigma}_2^2)$$

$$\hat{\pi}_1\frac{1}{\sqrt{2\pi}}\frac{1}{\hat{\sigma}_1}\exp\left(-\frac{1}{2}\left(\frac{\lambda-\hat{\mu}_1}{\sigma_1^2}\right)\right) = \hat{\pi}_2\frac{1}{\sqrt{2\pi}}\frac{1}{\hat{\sigma}_2}\exp\left(-\frac{1}{2}\left(\frac{\lambda-\hat{\mu}_2}{\sigma_2^2}\right)\right) \qquad (*)$$

On résoudre à la mains et on résoudre avec R :

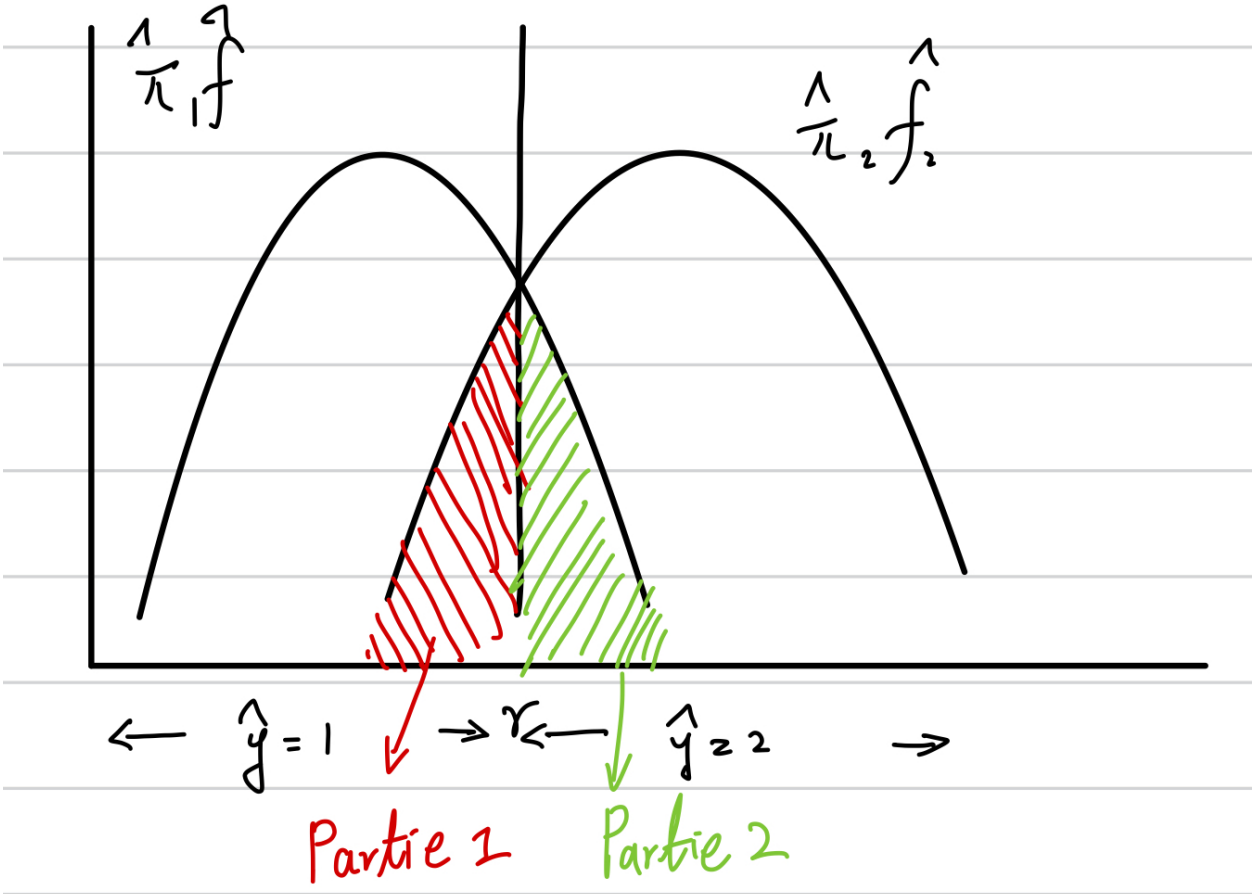
```
y = function(x){
  pi1*dnorm(x,mu1,sqrt(sicar1))-pi2*dnorm(x, mu2, sqrt(sicar2))
}
out = uniroot (y, lower=1.5, upper = 2.5)
```

```
gamma = out$root
print(gamma)
```

```
## [1] 2.091595
```

–> r = 2.09

$$\hat{\varepsilon}_{th} = \underbrace{\hat{\Pi}_1 \int_r^+ \infty f(x;\hat{\mu}_1;\hat{\sigma}_1^2)dx}_{Partie_1} + \underbrace{\hat{\Pi}_2 \int_r^+ \infty f(x;\hat{\mu}_2;\hat{\sigma}_2^2)dx}_{Partie_2}$$



$$\begin{aligned}\hat{\varepsilon}_{th} &= \hat{\Pi}_1 P(N|\hat{\mu}_1;\hat{\sigma}_1^2 > \gamma) + \hat{\Pi}_2 P(N|\hat{\mu}_2;\hat{\sigma}_2^2 > \gamma) \\ &= \hat{\Pi}_1\Big(1 - f(\frac{\gamma - \hat{\mu}_1}{\hat{\sigma}_1})\Big) + \hat{\Pi}_2\Big(f(\frac{\gamma - \hat{\mu}_2}{\hat{\sigma}_2})\Big)\end{aligned}$$

Sur R,

```
Varpsilon = pi1 * (1-pnorm(2.03,mu1,sqrt(sicar1))) + pi2*pnorm(2.09,mu2,sqrt(sicar2))
print(Varpsilon)
```

```
## [1] 0.3945809
```

$\hat{\varepsilon} = 0.3880232$

ii) Classifier Gamma homosédastique : Même démarche avec

$$\hat{\sigma}_1^2 = \hat{\sigma}_2^2 = \hat{\Pi}_1 \times 3.0 + \hat{\Pi}_2 \times 3.2 = 3.1$$

Affectation de M.Li :

$$t_1(Li) = 0.51$$

On affecté M.Li à classe 1 avec 51% de classe erreur de classement $\hat{\varepsilon}_{th} = 0.3882125$ et $\gamma = 2$

Eurreur apparente (ou de resubstitution)

L'erreur apparente esst calculée par restitution : les observations de D sont prises à la fois pour données d'apprentissage et données de test.

- Construir estimée un classifieur sur les donné de D
- Pour chaque $i \in \{1, ..., n\}$ déterminer la classe estimée de $x_i : \hat{y}(x_i)$
- Calculer l'eurreur apparente :

$$\hat{\varepsilon} = \#\{i; \hat{y}(x_i) \neq y_i\}/n$$

[3]

- $\hat{\varepsilon}_{ap}$ **sous estime l'erreur de classement (biais d'optimisme)**. Le biais de $\hat{\varepsilon}_{ap}$ est d'autant plus grand que n est petit ou que le modèle est complexe (beaucoup de paramètre).

"Erreur apparentée"（或称重复估计误差）这个概念通常用于衡量模型在训练数据集上的性能。它的主要特点是，**它评估模型对于已经见过的数据的拟合程度，而不是对未见过的数据的泛化性能**。在训练数据集上计算的误差通常会**低估模型的真实性能**，因为模型可能会过度拟合训练数据，无法很好地泛化到新数据，特别是在数据有限或模型复杂时。

因此，"erreur apparentée" 或重复估计误差并不是一种很好的模型性能评估方法，因为它不能提供模型在未见过的数据上的表现。为了更全面地评估模型的性能，通常还需要使用交叉验证或将数据集划分为训练集和测试集，以便评估模型在未见过的数据上的泛化能力。这可以帮助确定模型是否过度拟合了训练数据，以及模型在真实世界中的实际效果如何。

Boostrap

L'erreur estimée par bootstrap vise à réduire le biais de l'erreur apparente $\hat{\varepsilon}_{ap}$ calculée sur D. Pour $j = 1, ..., N$:

[3]这里的井号 # 表示数学符号中的计数符号，用于表示一个集合中元素的数量或计数。在这个上下文中，$\#i; \hat{y}(x_i) \neq y_i$ 表示了满足条件 $\hat{y}(x_i) \neq y_i$ 的数据点的数量。它计算了分类器在测试数据中错误分类的数据点数目。

- objectif : corrigé le biais de $\hat{\varepsilon}_ap$
- étapes : pour i affant de 1 à N (chemin)
  - tirer de D un échantillon D_j de taille n compoetant de possibles répétitions
  - construire le classifieur sur les données de D_j puis:(使用 D_j 上的数据来构建一个分类器，然后：)
    - classer les données de D_i et déterminer me taix d'erreur $\hat{\varepsilon}_{D_j}$
    - classer les données de D et déterminer le taux d'erreur $\hat{\varepsilon}_D$
- calculer $\beta_j = \hat{\varepsilon}_{D_j} - \hat{\varepsilon}_D$

L'erreur estimée par bootstrap est : $\hat{\varepsilon}_{boot} = \hat{\varepsilon}_{ap} + \sum_{j=1}^{N} \beta_j/N$

Le terme $\sum_{j=1}^{N} \beta_j/N$ est destiné à corriger le biais d'optimisme associé à l'erreur apparente $\hat{\varepsilon}_{ap}$

Leave one out

Le leave One Out ou validation croissée sonsidère tout à tout chacun des points de D pour donnée de test et les autres points pour données d'apprntissage. 留一法是一种交叉验证的技巧，它考虑将每一个数据点作为测试数据，其余数据点作为训练数据。这意味着对于每个数据点 x_i，它都会被单独用作测试数据，其余数据点用于训练。

Pour $i = 1, ..., n$:

- construire la classifieur sur D privé de (x_i, y_i) 构建一个分类器，使用数据集 D 但不包括数据点 (x_i, y_i)。
- estimer la classe de $x_i : y_i$ 估计数据点 x_i 的类别，表示为 $\hat{y}_i$。
- comparer $\hat{y}_i$ à la vraie classe de $x_i : y_i$ 比较 $\hat{y}_i$ 与数据点 x_i 的真实类别 y_i。

L'erreur de calidation croisée (Crosse Validation) est $\hat{\varepsilon}_{CV} = \sum_{i=1}^{n} 1_{\{\hat{y}_i \neq y_i\}/n}$

交叉验证错误率 $\hat{\varepsilon}_{CV}$ 是通过计算被分类器错误分类的数据点数量，然后除以数据点总数 n，得到的错误率。具体公式为：$\hat{\varepsilon}_{CV} = \sum_{i=1}^{n} 1_{\hat{y}_i \neq y_i/n}$。

留一法交叉验证的主要思想是通过反复单独测试每个数据点，以获得对分类器性能的更准确估计。它是一种有效的方法，特别适用于小型数据集或在评估分类器性能时需要减小估计偏差的情况。

v-flod Cross Validation

受留一法启发，On établit d'abord une partition de $\{1, ..., n\}$ en K groupes de tailles similaires : $I_1, ..., I_K$. Puis, pour $k = \{1, ..., K\}$:

- on construit un classifieur sur $\{(x_i, y_i); i \notin I_k\}$
- on calcue l'erreur de classement $\hat{\varepsilon}_k$ sur $\{(x_i, y_i); i \in I_k\}$

L'erreur v-fold Cross Validation est : $\hat{\varepsilon}_{v-flod\ CV} = \sum_{k=1}^{K} \hat{\varepsilon}_k / K$.

Lorsque $K = n$, le v-flod Crosse validation est un Leave One Out

v-fold 交叉验证是一种常用的模型评估方法，它允许将数据集分成多个子集，以多次评估模型性能，从而提供更准确的性能估计。这对于在小样本情况下或需要减小估计偏差的情况下非常有用。

Les courbes ROC (receier opérating characteristic)

On suppose le donné répartis en deux classes

- Classe_1 = sans risque
- Classe_2 = à risque

Un indicidu $x \in X$ est affect à l'une de classe par comparaison de non score à un seuil c :

- $\widehat{y}(x) = 2 \quad \text{ssi} \quad s(x) > c$
- $\widehat{y}(x) = 1 \quad \text{ssi} \quad s(x) \leq c$

Les observation se répartient en quatre catégories :

Etat	Explication	Equation
Faux Positifs	真 1 算 2	$FP = \#\{y_i = 1, \widehat{y}(x_i) = 2\}$
Vrais Positifs	真 2 算 2	$VP = \#\{y_i = 2, \widehat{y}(x_i) = 2\}$
Faux Négatif	真 2 算 1	$FN = \#\{y_i = 2, \widehat{y}(x_i) = 1\}$
Vrais Négatif	真 1 算 1	$VN = \#\{y_i = 1, \widehat{y}(x_i) = 1\}$

Nombre de négatif	$N = FP + VN$
Nombre de possitif	$P = VP + FN$
Taux de faux possitif	$TFP = FP/N$
Taux de vrais possitif	$TVP = VP/P$

remarque L'erreur apparente est :

$$\hat{\varepsilon}_{esp} = TFP \times \frac{N}{n} + |1 - TVP| \times \frac{P}{n}$$

Justification

$$\begin{aligned} TFP \times \frac{N}{n} + |1 - TVP| \times \frac{P}{n} &= \frac{FP}{N} \times \frac{N}{n} + (1 - \frac{VP}{P}) \times \frac{P}{n} \\ &= \frac{FP}{n} + \frac{P - VP}{P} \times \frac{P}{n} \\ &= \frac{FR}{n} + \frac{FN}{n} \\ &= \frac{FP + FN}{n} \\ &= \frac{\#\{\text{mal classe}\}}{n} \\ &= \hat{\varepsilon}_{ap} \end{aligned}$$

Le statistique TFP et TVP depandent du seuil c auquel on compare les valeurs du score S calculés en l'échantillon de test

à chaque caleur de c xxxxxxx par du couple (TFP, TVP)

Les courb ROC est l'emsemble des points {(TFPc; TVPc);c s(X)}

On appelle avec (area under curve) l'aire sous la courbe Roc

Exercice en considère les données de table 3 EX3 td 3

Tracer la courbe ROC du classification

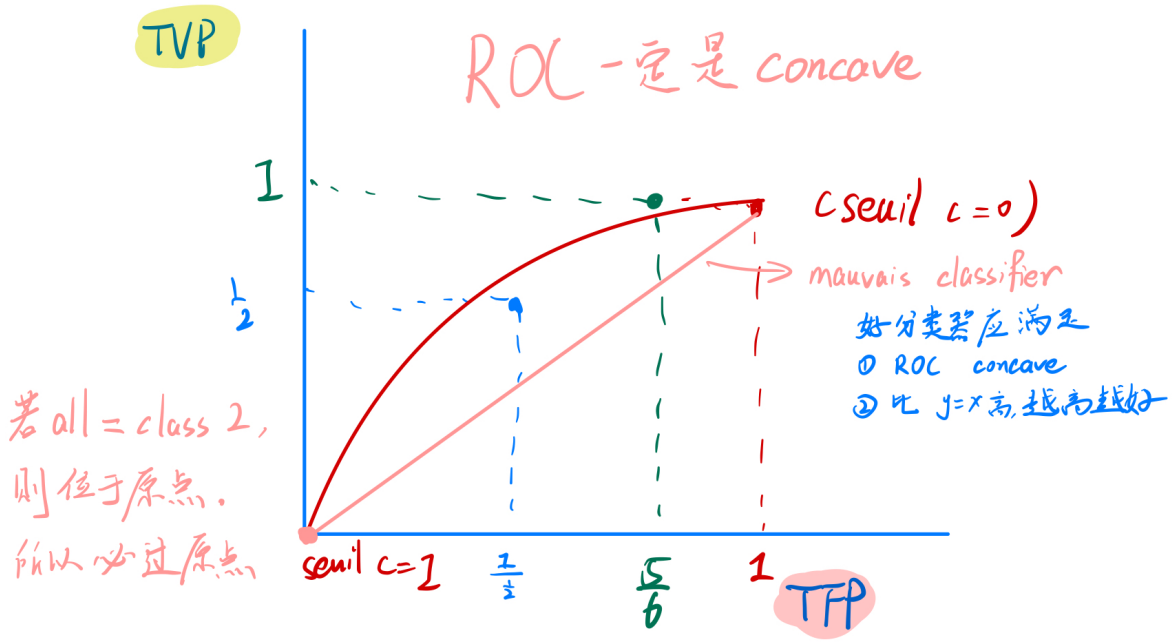
Période	Taux de change en début de période €/$	Rentabilité du marché américain	Rentabilité du marché français
1	0.67	10%	12%
2	0.74	-2%	13%
3	0.8	12%	-4%
4	0.71	22%	5%
5	0.76	5%	14%
6	0.82		

Question 1

Période	Taux de change en début de période €/$	Évolution du change €/$	Évolution du change $/€	Rentabilité du marché américain pour un investisseur français	Rentabilité du marché français pour un investisseur américain
1	0.67				
2	0.74	10.45%	-9.46%	-0.41%	23.70%
3	0.8	8.11%	-7.50%	-9.35%	22.16%
4	0.71	-11.25%	12.68%	26.20%	-14.80%
5	0.76	7.04%	-6.58%	13.97%	12.39%
6	0.82	7.89%	-7.32%	-2.68%	23.00%

	$\hat{1}$	$\hat{2}$		
1	2	4	6	TFP = 2/3
2	2	2	4	TVP = 1/2

Seuille c = 1/2



Choix de modèle

On peut comparer les modèles paramètrique ou semi-paramétrique par des critère d'information Tel que AIC , BIC (voir CH2)

Cependantn un critère de choix de modèle réalise un objectif (AIC estime la déviance d'un modèle, BIC estime la probabilité d'un modèle coditionnellement aux données, etc)

L'objectif du critère de choix de modèle peut être éloigné de celui de la classification. **Autrement dit, un modèle peut être bon du moint de vie du critère qu'il optimise mais ne pas conduire à un classifieur satisfaissant.**

étape : établir une partition de {1,….,n} en K classe de

```
#x = read.table(file = "http://alexandrelourme.free.fr/M2IREF/SCORING/BANKNOTE" , sep=',', dec='.'
```

On considère trois modèle :

- AD Gaussienne honorosédastique
- AD Gaussienne hétérosédastique
- Regression logistique

Determiner sous classe de modèle

- l'erreur LOO
- L'erreur apparente
- La class de nouveau billet : Length = 214,90 Right = 129,96 Left = 130,12

$$\widehat{\varepsilon}_{boost} = (\frac{1}{N}\sum_{i=1}^{N}\beta_i) + \widehat{\varepsilon}_{op}$$

> V flod cross valuation Pricipe : le même que L_oe mesure à chaque estimation stocastique plusieur port de donnés d'appentissage

etape : établir une partition de {1,...n} en t class de taille I_1,...,I_K peu d'allant de 1 à K

cotinuer un classifier sur D1{(); I_k} Estimer le taux d'errer sur {(x,z); i I_l} : $\widehat{\varepsilon}$

Erreur V-flod cross valuation est

$$\widehat{\varepsilon}_{V-flod} = \sum_{k=1}^{k}\widehat{\varepsilon}_k/k$$