$\frac{1}{n}\sum_{i=1}^{n} \hat{f}_X^{-i}(X_i)$ est un estimateur sans biais de type "leave-one out" de $\mathbb{E}\left[\int \hat{f}_X(x) f_X(x)\,dx\right]$

Proposition : $\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} \hat{f}_X^{-i}(X_i)\right] = \mathbb{E}\left[\int \hat{f}_X(x) f_X(x)\,dx\right]$

Demonstration :

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} \hat{f}_X^{-i}(X_i)\right] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[\hat{f}_X^{-i}(X_i)\right] \overset{iid}{=} \mathbb{E}\left[\hat{f}_X^{-1}(X_1)\right]$$

$$= \mathbb{E}\left[\frac{1}{n-1}\sum_{\substack{j=1\\ j\neq i}}^{n} \frac{1}{h} k\left(\frac{X_j - X_1}{h}\right)\right] \overset{iid}{=} \frac{1}{n-1}(n-1)\,\mathbb{E}\left[\frac{1}{h} k\left(\frac{X_2 - X_1}{h}\right)\right]$$

$$= \mathbb{E}\left[\frac{1}{h} k\left(\frac{X_2 - X_1}{h}\right)\right] = \int \frac{1}{h} k\left(\frac{x_2 - x_1}{h}\right) f_X(x_1) f_X(x_2)\,dx_1\,dx_2.$$

Or $\mathbb{E}\left[\int \hat{f}_X(x) f_X(x)\,dx\right] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\int \frac{1}{h} k\left(\frac{X_i - x}{h}\right) f_X(x)\,dx\right]$

$$\overset{iid}{=} \mathbb{E}\left[\int \frac{1}{h} k\left(\frac{X_1 - x}{h}\right) f_X(x)\,dx\right]$$
$\longrightarrow$ ou Fubini-Tonelli si $k > 0$.
$\longrightarrow$ ou regarder les $\int$ et vérifier

$$= \mathbb{E}\left[\mathbb{E}\left[\frac{1}{h} k\left(\frac{X_2 - X_1}{h}\right) \Big| X_2\right]\right] = \mathbb{E}\left[\frac{1}{h} k\left(\frac{X_2 - X_1}{h}\right)\right]$$

D'où $\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} \hat{f}_X^{-i}(X_i)\right] = \mathbb{E}\left[\int \hat{f}_X(x) f_X(x)\,dx\right]$.

La méthode "least squares cross validation" consiste donc à prendre $\hat{h}_{LSCV} = \underset{h>0}{\operatorname{argmin}}\ LSCV(h)$.

V. Estimation du mode

On va proposer une méthode d'estimation non-paramétrique du mode θ.

Rappel : $\theta \in \operatorname{argmax}\{t \in \mathbb{R} : f_X(t)\}$

On propose donc de manière assez naturelle l'estimateur du mode suivant

$$\hat{\theta} = \mathrm{argmax}\{t \in \mathbb{R} : \hat{f}_X(t)\}$$

Cet argmax n'est pas forcément unique. On prend alors l'une des valeurs qui maximise $\hat{f}_X$.

En pratique, la méthode la plus basique consiste à estimer $\hat{f}_X$ sur une grille $t_1, \dots, t_M$ de points équi-espacés d'un pas ε petit et de prendre comme estimateur du mode le point de la grille où $\hat{f}_X$ est maximale

$$\hat{\theta} = \mathrm{argmax}\{j \in [\![1;M]\!], \hat{f}_X(t_j)\}.$$

On peut bien entendu utiliser des méthodes plus raffinées :

↳ Newtown - Raphson

↳ méthode du nombre d'or

↳ ...

C - Estimateur à noyau de la fonction de régression

On suppose maintenant que l'on observe pour chaque individu une observation (X_i, Y_i) de variables réelles. On suppose $(X_i, Y_i)_{1 \leq i \leq m} \overset{iid}{\sim} (X, Y)$.

I - Définition

On souhaite étudier la manière dont la variable d'intérêt (ou réponse) Y dépend de la variable explicative (ou auxiliaire) X.

Il s'agit donc d'un modèle de régression de Y sur X
$$Y = r(X) + \varepsilon, \quad r \text{ mesurable et } \mathbb{E}[\varepsilon | X] = 0.$$

<u>Remarque</u> : • $\mathbb{E}[Y|X] = \mathbb{E}[r(X)|X] + \mathbb{E}[\varepsilon|X]$

$\mathbb{E}[Y|X] = r(X)$

• $cov(X, \varepsilon) = \mathbb{E}[\varepsilon X] - \mathbb{E}[\varepsilon]\mathbb{E}[X]$

$= \mathbb{E}[\mathbb{E}[\varepsilon|X]\mathbb{E}[X]] - \mathbb{E}[\mathbb{E}[\varepsilon|X]]\mathbb{E}[X]$

$= 0 \quad$ car $\mathbb{E}[\varepsilon|X] = 0$.

On cherche à estimer r, la fonction de régression. On connaît déjà le modèle de régression linéaire
$$Y = aX + b + \varepsilon.$$
Ce modèle fait une hypothèse forte sur la nature du lien entre X et Y (fonction de régression r) et se ramène à l'estimation des paramètres a et b. On parle de modèle paramétrique.

Nous allons maintenant considérer des modèles plus generaux, appelés modèles non-paramétiques pour lesquels, aucune hypothèse n'est faite sur la nature de r (fonction reliant Y et X)

Nous ferons simplement des hypothèses portant sur sa régularité

<u>Estimateur de Naradaya-Watson (1964).</u>

$$\hat{r}(x) = \begin{cases} \dfrac{\sum_{i=1}^{n} Y_i K\left(\frac{X_i - x}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)} & \text{si } \sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right) \neq 0 \\ \overline{Y} & \text{sinon} \end{cases}$$

Comme nous l'avons dit pour l'estimation de la densité, il est intéressant d'utiliser dans un premier temps un estimateurs non paramétrique afin de réaliser une étude exploratoire des données.
Si le résultat de cette étude met en évidence que la fonction r que l'on cherche à estimer est de nature connue (à quelques paramètres près) on pourra alors proposer des méthodes d'estimations paramétriques qui ont l'avantage de converger plus rapidement et d'être plus faciles à interpréter

Remarques: • L'estimateur $\hat{r}$ est une moyenne pondérée des Y_i :

$$\hat{r}(x) = \sum_{i=1}^{n} Y_i w_i(x) \quad \text{où} \quad w_i(x) = \begin{cases} \dfrac{K\left(\frac{X_i - x}{h}\right)}{\sum_{j=1}^{n} K\left(\frac{X_j - x}{h}\right)} & \text{si } \sum_{j=1}^{n} K\left(\frac{X_j - x}{h}\right) \neq 0 \\ \dfrac{1}{n} & \text{sinon.} \end{cases}$$

•• Cet estimateur peut s'écrire comme quotient de deux estimateurs, dont celui de la densité.

$$\hat{r}(x) = \frac{\hat{g}(x)}{\hat{f}_X(x)} \quad \text{avec} \quad \left| \begin{array}{l} \hat{g}(x) = \dfrac{1}{nh} \displaystyle\sum_{i=1}^{n} Y_i K\left(\frac{X_i - x}{h}\right) \\ \hat{f}_X(x) = \dfrac{1}{nh} \displaystyle\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right) \end{array} \right.$$

Choix de h

• Si h est trop petit, on a un sur-ajustement aux données. Les poids sont nuls sauf au voisinage des observations où ils deviennent proche de 1 (biais faible mais variance forte)
• Si h est trop grand, on a un sous-ajustement aux

données car les poids sont quasiments tous les mêmes (biais fort mais variance faible).

II - Propriétés asymptotiques :

Rappel : Convergence presque-complète

- $Z_n \xrightarrow[n \to +\infty]{p.co} Z \iff \forall \varepsilon > 0, \sum_{n \geq 0} \mathbb{P}(|Z_n - Z| > \varepsilon) < +\infty$

- $Z_n - Z = O_{p.co}(x_n) \iff \exists k > 0, \sum_{n \geq 0} \mathbb{P}(|Z_n - Z| > k x_n) < +\infty$.

- $Z_n - Z = o_{p.co}(x_n) \iff \forall \varepsilon > 0, \sum_{n \geq 0} \mathbb{P}(|Z_n - Z| > \varepsilon x_n) < +\infty$.

- convergence presque-complète $\Rightarrow$ convergence presque-sûr.

1 - Convergence presque complète ponctuelle

Hypothèses : a - r et f sont $k > 0$ fois continuement dérivable au voisinage de x.

b - $f_X(x) > 0$

c - $\lim_{n \to +\infty} h_n = 0$ et $\lim_{n \to +\infty} \frac{n h_n}{\ln(n)} = +\infty$.

d - On suppose K borné, intégrable et à support compact.

e - K est un noyau d'ordre k (Gasser et Muller, 1979) :

$\forall j = 1, .., k-1 : \int t^j K(t) dt = 0$ et $|\int t^k K(t) dt| < +\infty$ et $\neq 0$.

(note : incompatible avec K positif si $k \geq 2$)

f - $|Y| < M < +\infty$ p.s (peut être généralisé).

Théorème 1 : Sous les hypothèses précédentes, on a :

$$\hat{r}(x) - r(x) = O_{p.co}(h^k) + O_{p.co}\left(\sqrt{\frac{\ln(n)}{nh}}\right)$$

2- Convergence presque complète uniforme sur un compact

On appelle S un compact tel que $S \subseteq \mathbb{R}$.

Hypothèses : a') r et f sont $\mathcal{C}^k$ $(k>0)$ sur un voisinage de S

b') $\exists \theta > 0$, $\inf_{x \in S} \{ f_x(x) \} \geq \theta$

g) $\exists \beta > 0, \exists C < +\infty, \forall x \in S, \forall y \in S, \ |k(x) - k(y)| \leq C |x-y|^{\beta}$

Théorème 2 : Sous les hypothèses a') b') et c) à g), on a :
$$\sup_{x \in S} | \hat{r}(x) - r(x) | = O_{p.co}(h^k) + O_{p.co}\left(\sqrt{\frac{\ln(n)}{nh}} \right).$$

3- Convergence presque complète sous hypothèse de type Lipschitz ou de type Holder (sur f_x et r).

Pour $\phi = f_x$ et r

Hypothèses : a'') $\exists \beta > 0, \exists C < +\infty, \exists \beta > 0, \forall y \in]x-\varepsilon, x+\varepsilon[$,
$$|\phi(x) - \phi(y)| \leq C |x-y|^{\beta}$$

a''') $\exists \beta > 0, \exists C < +\infty, \forall x \in S, \forall y \in S$
$$|\phi(x) - \phi(y)| \leq C |x-y|^{\beta}$$

Théorème 3A : Sous les hypothèses a''), b) à d) et f), on a
$$\hat{r}(x) - r(x) = O_{p.co}(h^{\beta}) + O_{p.co}\left(\sqrt{\frac{\ln(n)}{nh}} \right)$$

Théorème 3B : Sous les hypothèses a''') b') c) d) f) et g), on a :
$$\sup_{x \in S} \left(|\hat{r}(x) - r(x)| \right) = O_{p.co}(h^{\beta}) + O_{p.co}\left(\sqrt{\frac{\ln(n)}{nh}} \right)$$

4- Convergence en moyenne quadratique :

Hypothèses : c') $\lim_{n \to +\infty} h_n = 0$ et $\lim_{n \to +\infty} n h_n = +\infty$.

h) $\phi(u) = \mathbb{E}[Y^2 | X = u]$ est continue au point x

d') k borné, intégrable, positif, symétrique et à support compact.

Notation : $g = r \, f_X$

Théorème 4 : Si a) est vérifié avec $k = 2$ et si b), c'), h) et d') sont vérifiées alors on a :

$$\mathbb{E}_x\left[\left(\hat{r}(x) - r(x)\right)^2\right] = B(x)^2 h^4 + V(x)\frac{1}{nh} + o\left(h^4 + \frac{1}{nh}\right)$$

où $B(x) = \dfrac{1}{2 f_X(x)} \displaystyle\int_{\mathbb{R}} t^2 K(t)\,dt \cdot \left(g^{(2)}(x) - r(x) f_X^{(2)}(x)\right)$

et $V(x) = \dfrac{\Phi(x) - r^2(x)}{f_X(x)} \displaystyle\int_{\mathbb{R}} K^2(t)\,dt$

5- Convergence en MISE

- $MISE(\hat{r}) = \mathbb{E}\left[\displaystyle\int_{\mathbb{R}} (\hat{r}(x) - r(x))^2 \omega(x)\,dx\right]$ où

i) ω est une fonction de poids positive, bornée et à support compacte S.

h') $\Phi(u) = \mathbb{E}[Y^2 | X = u]$ est continue sur un voisinage de S.

Théorème 5 : Sous les conditions a') avec $k = 2$ et si de plus b') d') i) et h') sont vérifiées alors :

$$MISE(\hat{r}) = B^2 h^4 + \frac{V}{nh} + o\left(h^4 + \frac{1}{nh}\right)$$

avec $B^2 = \int_{\mathbb{R}} B^2(x)\,\omega(x)\,dx$ et $V = \int_{\mathbb{R}} V(x)\,\omega(x)\,dx$

<u>Remarque :</u> On retrouve l'équilibre biais / variance entrevu pour la densité. Et on obtient de manière similaire des expressions asymptotiquement optimales pour h par rapport aux termes dominants de la MISE et de la MSE :

$$\hat{h}_{AMISE} = \left(\frac{V}{4B^2}\right)^{1/5} n^{-1/5} \quad \text{et} \quad \hat{h}_{AMSE} = \left(\frac{V(x)}{4B^2(x)}\right)^{1/5} n^{-1/5}.$$

6 - Vitesses de convergences

On obtient comme corollaire des résultats précédents :

- Sous les hypothèses du théorème 1 et 2 (respectivement) avec $h \sim C\left(\frac{n}{\ln(n)}\right)^{-1/2k+1}$

$$\hat{r}(x) - r(x) = O_{p.co}\left(\left(\frac{n}{\ln(n)}\right)^{-\frac{k}{2k+1}}\right)$$

$$\sup_{x\in S} |\hat{r}(x) - r(x)| = O_{p.co}\left(\left(\frac{n}{\ln(n)}\right)^{-\frac{k}{2k+1}}\right)$$

- Sous les hypothèses du théorème 3A et 3B avec $h \sim C\left(\frac{n}{\ln(n)}\right)^{-\frac{1}{2\beta+1}}$

$$\hat{r}(x) - r(x) = O_{p.co}\left(\left(\frac{n}{\ln(n)}\right)^{-\frac{\beta}{2\beta+1}}\right)$$

$$\sup_{x\in S} |\hat{r}(x) - r(x)| = O_{p.co}\left(\left(\frac{n}{\ln(n)}\right)^{-\frac{\beta}{2\beta+1}}\right)$$

- Sous les hypothèses du théorème 4 et 5, avec $h \sim C n^{-1/5}$.

$$MSE(\hat{r}(x)) = O(n^{-4/5}) = \frac{5}{4} B^{2/5}(x)\, V^{4/5}(x)\, 4^{1/5}$$

$$MISE(\hat{r}) = \mathcal{O}(n^{-4/5}) = \frac{5}{4} B^{2/5} V^{4/5} 4^{1/5}$$

7 - Choix du paramètre de lissage en pratique

Les choix optimaux que nous venons de voir nous fournissent des valeurs théoriques inutilisable en pratique.

Une méthode populaire : la validation croisée :

Notation : $MISE(\hat{r}) := MISE(h)$.

On choisit $h_{CV} = \underset{h \in \mathcal{H}}{\arg\min}\ CV(h)$

avec $CV(h) = \sum_{i=1}^{n} (Y_i - \hat{r}^{-i}(X_i))^2 \omega(X_i)$

où : • $\mathcal{H}$ est un ensemble de valeur possible pour h

• $$\hat{r}^{-i}(x) = \begin{cases} \dfrac{1}{\sum_{j=1, j\neq i}^{n} k\left(\frac{X_j - x}{h}\right)} \sum_{j=1, j\neq i}^{n} Y_j\, k\left(\frac{X_j - x}{h}\right) & \text{si } \sum_{j=1, j\neq i}^{n} k\left(\frac{X_j - x}{h}\right) \neq 0 \\ \overline{Y^{-i}} = \dfrac{1}{n-1} \sum_{j=1, j \neq i}^{n} Y_j & \text{sinon} \end{cases}$$

$$CV(h) = \sum_{i=1}^{n} \left(r(X_i) + \varepsilon_i - \hat{r}^{-i}(X_i)\right)^2 \omega(X_i)$$

$$= \sum_{i=1}^{n} \varepsilon_i^2 \omega(X_i) + \sum_{i=1}^{n} (r(X_i) - \hat{r}^{-i}(X_i))^2 \omega(X_i) + \underbrace{2\sum_{i=1}^{n} \varepsilon_i (r(X_i) - \hat{r}^{-i}(X_i)) \omega(X_i)}_{= A \text{ et } \mathbb{E}[A] = 0}$$

Théorème : Sous les hypothèses du théorème 4B et si $\mathcal{H}$ contient que des largeurs de fenêtre vérifiant c') alors

$$\frac{1}{MISE(h_{CV})} \times \inf_{h \in \mathcal{H}} (MISE(h)) \xrightarrow[n \to +\infty]{p.s} 1$$

On choisit $h_{CV} = \underset{h \in H}{\operatorname{argmin}} \; CV(h)$

$$\text{avec } CV(h) = \sum_{i=1}^{n} \left(Y_i - \hat{r}^{-i}(X_i)\right)^2 w(X_i)$$

où · H est un ensemble de valeurs possibles pour h

· $$\hat{r}^{-i}(x) = \begin{cases} \dfrac{\sum_{\substack{j=1,n \\ j \neq i}} Y_j \, K\left(\frac{X_j - x}{h}\right)}{\sum_{\substack{j=1,n \\ j \neq i}} K\left(\frac{X_j - x}{h}\right)} & \text{si } \sum_{j \neq i} K(\,) \neq 0 \\ \overline{Y}^{-i} & \end{cases}$$

début cours.

$$CV(h) = \sum_{i=1}^{n} \left(r(X_i) + \varepsilon_i - \hat{r}^{-i}(X_i)\right)^2 w(X_i)$$

$$= \sum_{i=1}^{n} \varepsilon_i^2 \, w(X_i) + \sum_{i=1}^{n} \left(r(X_i) - \hat{r}^{-i}(X_i)\right)^2 w(X_i)$$

$$+ \underbrace{2 \sum_{i=1}^{n} \varepsilon_i \left(r(X_i) - \hat{r}^{-i}(X_i)\right) w(X_i)}_{A} \quad E[A] = 0 \text{ car } E[$$

Théorème : sous les hypothèses du Théorème 4b et si H contient que des largeurs de fenêtre vérifiant $c')$ alors :

$$\frac{\inf_{h \in H} MISE(h)}{MISE(h_{cv})} \xrightarrow[n \to +\infty]{P.s} 1$$

Il existe d'autres méthodes pour choisir h (voir poly ...

Choix de K (cf densité, si on se restreint aux noyaux de cet

Modèle de regression étudié

$$Y = r(X) + \varepsilon \quad \text{avec} \quad \mathbb{E}[\varepsilon \mid X] = 0$$

où $r(X)$ est l'espérance de la loi conditionnelle de Y sachant X.

On peut aussi s'interesser à l'estimation d'autres paramètres de cette loi : mode, médiane, quantiles qui sont plus robustes à la présence de valeurs extrêmes/abbé

K_0 noyau >0, d'intégral 1 (densité)
$H(u) = \int_{-\infty}^{u} K_0(t)dt$ (fdr)

Mode conditionnel

densité conditionnelle de Y sachant X

$$\hat{f}_{Y|X}(y,x) = \begin{cases} \dfrac{\sum_{i=1}^{n} \frac{1}{g} K_0\left(\frac{Y_i - y}{g}\right) K\left(\frac{X_i - x}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)} & \text{si } \sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right) \neq 0 \\ 0 & \text{sinon} \end{cases}$$

$$\hat{\Theta}_{Y|X}(x) = \underset{y}{\operatorname{argmax}}\ \hat{f}_{Y|X}(y,x)$$

Mediane conditionnelle

Fdr conditionnelle

$$\hat{F}_{Y|X}(y,x) = \begin{cases} \dfrac{\sum_{i=1}^{n} H\left(\frac{Y_i - y}{g}\right) K\left(\frac{X_i - x}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)} & \text{si } \sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right) \neq 0 \\ 0 & \text{sinon} \end{cases}$$

$\hat{q}_{Y|X}(\alpha, x)$ solution en y de $\hat{F}_{Y|X}(y,x) = \alpha$

(par la mediane $\alpha = 0,5$)

III Applications à la discrimination (classification supervisée)

Pour un même échantillon on observe

Y classe $\in \{0, \dots, k-1\}$

X donnée numérique

Echantillon $(X_1, Y_1), \dots, (X_n, Y_n) \overset{iid}{\sim} (X, Y)$

Objectif : Construire à partir de l'échantillon un classifieur qui a une nouvelle valeur x_0 de X associe une classe (à ce nouvel individu)

Classifieur de Bayes

$$\hat{y}_0 = \underset{y}{\operatorname{argmax}}\ \mathbb{P}(Y = y \mid X = x_0)$$

C'est le meilleur classifieur possible mais il n'est pas utilisable car $y \to \mathbb{P}(Y = y \mid X = x)$ sont inconnue.

· Approche paramétrique Regression / Modèle logistique

pour $k = 2$ seulement

$$\mathbb{P}(Y = 1 \mid X = x_0) = \frac{e^{\beta + \langle \alpha, x_0 \rangle}}{1 + e^{\beta + \langle \alpha, x_0 \rangle}}$$

· Approche non paramétrique proposée

On remarque que $\forall y$ $\mathbb{P}(Y = y \mid X = x) = \mathbb{E}[\mathbb{1}_{Y=y} \mid X = x]$

$\mathbb{P}(Y = y \mid X) = \mathbb{E}[\mathbb{1}_{Y=y} \mid X]$

d'où $\forall y$ $X \to \mathbb{P}(Y = y \mid X)$ est la fonction de regression d modèle $\mathbb{1}_{Y=y} = r_y(X) + \varepsilon_y$ avec $\mathbb{E}[\varepsilon_y \mid X] = 0$

avec $r_y(X) = \mathbb{P}(Y = y \mid X)$

On propose donc d'estimer $\forall y$ la proba $\mathbb{P}(Y=y \mid X=x)$ par $\widehat{\mathbb{P}(Y=y \mid X=x)} = \hat{r}_y(x) = \begin{cases} \dfrac{\sum \mathbb{1}_{Y_i=y}\, K\left(\frac{X_i-x}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{X_i-x}{h}\right)} \\ \overline{\mathbb{1}_{Y=y}} \quad \text{sinon} \end{cases}$

Classifieur de Bayes "Estimé"

$$\hat{y}_{x_0} = \underset{y}{\operatorname{argmax}}\ \widehat{\mathbb{P}(Y=y \mid X=x_0)} = \underset{y}{\operatorname{argmax}}\ \hat{r}_y(x_0)$$

Exo : $\sum_{y \in \{0, \dots, k-1\}} \hat{r}_y(x) = 1$

Cas particulier $k=2$

$$\hat{y}_0 = 1 \iff \hat{r}_1(x_0) > 0,5$$

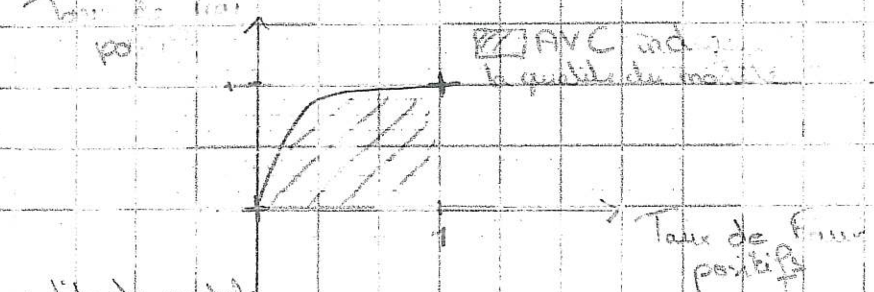
$$(\text{si } \hat{r}_0(x_0) = \hat{r}_1(x_0) = 1/2 \text{ alors } \hat{y}_0 = 0)$$

Qualité de la méthode

- Taux d'erreur global
- Taux d'erreur relatifs (Matrice de Confusion)

classe \ classe estimée	0	1	2
0	ok		
1		ok	
2			ok

- Pour $k=2$, courbe ROC



classe \ classe estimée	0	1	
0	n_{11}	n_{12}	$n_{1\cdot}$
1	n_{21}	n_{22}	
	$n_{\cdot 1}$	$n_{\cdot 2}$	

qualité du modèle = aire sous la courbe

plus elle est grande meilleur est la classification

Taux de Faux positifs : $\frac{N_{12}}{N_{1.}}$

Taux de Vrai positifs : $\frac{N_{22}}{N_{2.}}$

$$\hat{y}_0 = 1 \Leftrightarrow \hat{r}_y(x_0) > s \,, \quad s \in [0,1]$$

$s=1$, $N_{12} = 0$, $N_{22} = 0$

$s=0$, tous de la classe 1, $N_{12} = 1$, $N_{22} = 1$

+ AUC est grande, meilleure est la classification

CV- classif : $\frac{1}{n} \sum_{i=1}^{n} 1_{\{Y_i \neq \hat{y}_i^{-i}\}}$

$$\hat{y}_i^{-i} = \underset{y}{\operatorname{argmax}} \; \hat{r}_y^{-i}(x_i)$$