

Apprentissage statistique et grande dimension

TD 3 : classification supervisée

Exercice 1 (Classifieur bayésien)

Soit (X, Y) un couple de loi P avec $Y \in \{0, 1\}$. Soit g^* le classifieur de Bayes, et R^* son risque de classification. On note également $p = \mathbb{P}(Y = 1)$.

1. Montrer que $R^* \leq \min\{p, 1 - p\}$.
2. Montrer que si X et Y sont indépendants, $R^* = \min\{p, 1 - p\}$.
3. Fabriquer un exemple où $R^* = \min\{p, 1 - p\}$ et où X et Y ne sont pas indépendants. Indication : considérer X, Y binaires dépendants.

Réponse de l'exercice 1.

Rappel notation : $\mathcal{X} \subset \mathbb{R}^d$ espace des covariables X ; $i = 1, \dots, n$ indice des observations; X_i vecteur des covariables et Y_i variable réponse du i ème individu; $\mathcal{D}_n = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$. l'espace d'échantillonnage.

0. Voir page 223 de BiauDevroye.pdf déjà vu en cours

Un classifieur binaire est une fonction mesurable $h : \mathcal{X} \times \mathcal{D}_n \mapsto \{0, 1\}$. $h(\cdot, \mathcal{D}_n)$ détermine la règle de classification issue des données \mathcal{D}_n .

La fonction de régression pour un problème de classification est $\eta(x) = E(Y|X = x) = P(Y = 1|X = x)$. Le risque de classification pour un classifieur h conditionnel est $R_{\mathcal{D}_n}(h) = P(h(X) \neq Y|\mathcal{D}_n)$ tandis que son risque de classification (inconditionnel) est $R(h) = E(P(h(X) \neq Y|\mathcal{D}_n)) = P(h(X) \neq Y)$. Dans ce cas binaire pour tout classifieur h le risque est

$$\begin{aligned} R(h) &= E(1_{h(X) \neq Y}) = E(1_{h(X)=0, Y=1} + 1_{h(X)=1, Y=0}) \\ &= E(E(1_{h(X)=0, Y=1}|X)) + E(E(1_{h(X)=1, Y=0}|X))E(1_{h(X)=0}E(1_{Y=1}|X)) + E(1_{h(X)=1}E(1_{Y=0}|X)) \\ &= E(\eta(X)1_{h(X)=0} + (1 - \eta(X))1_{h(X)=1}) \end{aligned}$$

1. Ici on se place en l'absence de données puisqu'on étudie le classifieur g^* de Bayes. Donc son risque vaut

$$R^* = E(P(g^*(X) \neq Y|X)) = E(P(Y = 1|X)1_{g^*(X)=0} + P(Y = 0|X)1_{g^*(X)=1})$$

Le classifieur de Bayes est le classifieur issue de la vraie loi de la variable réponse, c'est à dire

$$g^*(x) = \begin{cases} 1 & \text{si } P(Y = 1|X = x) > P(Y = 0|X = x) \\ 0 & \text{sinon} \end{cases}$$

Or $P(Y = 1|X = x) > P(Y = 0|X = x) \Leftrightarrow 2P(Y = 1|X = x) > 1 \Leftrightarrow P(Y = 1|X = x) > 1/2$. Donc

$$g^*(x) = \begin{cases} 1 & \text{si } P(Y = 1|X = x) > 1/2 \\ 0 & \text{sinon} \end{cases}$$

Réécriture avant majoration - Méthode 1 :

Déterminons son risque

$$\begin{aligned} R(g^*) &= E(P(Y = 1|X)1_{g^*(X)=0} + (1 - P(Y = 1|X))1_{g^*(X)=1}) \\ &= E(P(Y = 1|X)1_{P(Y=1|X) \leq 1/2} + (1 - P(Y = 1|X))1_{P(Y=1|X) > 1/2}) \\ &= E(\min(P(Y = 1|X), 1 - P(Y = 1|X))) \end{aligned}$$

Réécriture avant majoration - Méthode 2 :

Déterminons son risque

$$R(g^*) = E(P(g^*(X) \neq Y|X)) = P(g^*(X) \neq Y) = 1 - P(g^*(X) = Y)$$

puisque g^* est indépendant des données. De plus $\{g^*(X) = Y|X\} = \{g^*(X) = 0 \cap Y = 0|X\} \cup \{g^*(X) = 1 \cap Y = 1|X\}$ Donc

$$\begin{aligned} P(g^*(X) = Y|X) &= P(g^*(X) = 0 \cap Y = 0|X) + P(\{g^*(X) = 1 \cap Y = 1|X\}) \\ &= 1_{g^*(X)=0}P(Y = 0|X) + 1_{g^*(X)=1}P(Y = 1|X) \end{aligned}$$

Or la variable aléatoire $P(g^*(X) = Y|X)$ est une variable avec deux valeurs possibles de probabilités.

— Si $g^*(X) = 0$ alors

$$P(Y = 0|X) > 1/2 > P(Y = 1|X) \Rightarrow P(g^*(X) = Y|X) = \max(P(Y = 0|X), P(Y = 1|X)).$$

— Si $g^*(X) = 1$ alors

$$P(Y = 1|X) > 1/2 > P(Y = 0|X) \Rightarrow P(g^*(X) = Y|X) = \max(P(Y = 0|X), P(Y = 1|X)).$$

Ainsi $P(g^*(X) = Y|X) = \max(P(Y = 0|X), P(Y = 1|X)) \Rightarrow P(g^*(X) = Y) = E(\max(P(Y = 0|X), P(Y = 1|X)))$.
Donc

$$R(g^*) = 1 - E(\max(P(Y = 0|X), P(Y = 1|X))) = E(\min(P(Y = 0|X), P(Y = 1|X)))$$

Majoration :

Comme

$$E(\min(P(Y = 1|X), 1 - P(Y = 1|X))) = \int_{\mathbb{R}} \min(P(Y = 1|X = x), P(Y = 0|X = x))dF_X(x)$$

$$R(g^*) \leq \int_{\mathbb{R}} P(Y = 1|X = x)dF_X(x) = P(Y = 1) \leq \min(P(Y = 1), P(Y = 0))$$

$$\leq \int_{\mathbb{R}} P(Y = 0|X = x)dF_X(x) = P(Y = 0)$$

On obtient la majoration du risque

$$R(g^*) \leq \min(P(Y = 1), 1 - P(Y = 1)) = \min(p, 1 - p).$$

On a aussi

$$R(g^*) = E(\min(P(Y = 1|X), 1 - P(Y = 1|X))) = E(\min(\eta^*(X), 1 - \eta^*(X)))$$

avec $\eta^*(X) = P(Y = 1|X)$.

2. Si indépendance alors $P(Y = .|X) = P(Y = .)$ et donc $g^*(x) = 1_{p > 1/2}$ est déterministe et

$$R(g^*) = P(g^*(X) = 0)P(Y = 1) + P(g^*(X) = 1)P(Y = 0) = 1_{p \leq 1/2}p + 1_{p > 1/2}(1 - p) = \min(p, 1 - p)$$

3. Il nous faut choisir $p \notin \{0, 1/2, 1\}$. En effet

si $p = 0$ alors $Y = 0$ p.s. donc Y est indépendant de X ;

si $p = 1$ alors $Y = 1$ p.s. donc Y est indépendant de X ;

si $p = 1/2$ alors d'après le cours $\eta(x) = 1/2$ et $R^* = 1/2$ donc Y est indépendant de X .

Cherchons un cas dépendance

Fixons une covariable binaire $X \in \{0, 1\}$. Autrement dit on cherche les probabilités élémentaires suivantes :

	conditionnelle		incond.
	$X = 0$	$X = 1$	
$Y = 0$	$P(Y = 0 X = 0) = p_{0 0}$	$P(Y = 0 X = 1) = p_{0 1}$	$1 - p$
$Y = 1$	$P(Y = 1 X = 0) = p_{1 0}$	$P(Y = 1 X = 1) = p_{1 1}$	p
	$p_{0 0} + p_{1 0} = 1$	$p_{0 1} + p_{1 1} = 1$	

La dépendance entraîne $p_{0|0} \neq p_{0|1}$ ou $p_{0|0} \neq p$. Dans ce cadre, le classifieur de Bayes se simplifie

$$g^*(x) = \begin{cases} 1 & \text{si } p_{1|x} > 1/2 \\ 0 & \text{sinon} \end{cases}$$

Deux exemples simples ci-dessous

	$X = 0$	$X = 1$		$X = 0$	$X = 1$
$Y = 0$	$3/8 = p_{0 0}$	$7/8 = p_{0 1}$,	$Y = 0$	$4/8 = p_{0 0}$ $3/8 = p_{0 1}$
$Y = 1$	$5/8 = p_{1 0}$	$1/8 = p_{1 1}$		$Y = 1$	$4/8 = p_{1 0}$ $7/8 = p_{1 1}$
$g^*(X)$	1	0		$g^*(X)$	0 1
$\eta^*(X)$	5/8	1/8		$\eta^*(X)$	4/8 7/8

Dans le premier cas, on a

		général	$P(X = 1)$	
			1/2	1/4
R^*	$3/8P(X = 0) + 1/8P(X = 1)$		1/4	5/16
p	$5/8P(X = 0) + 1/8P(X = 1)$		3/8	1/2
$1 - p$			5/8	1/2
$\min(p, 1 - p)$			3/8	1/2

Dans le second cas, on a

		général	$P(X = 1)$	
			1/2	1/4
R^*	$1/2P(X = 0) + 1/8P(X = 1)$		5/16	13/32
p	$4/8P(X = 0) + 7/8P(X = 1)$		11/16	19/32
$1 - p$			5/16	3/32
$\min(p, 1 - p)$			5/16	3/32

Déterminons la conditionnelle $P(Y = .|X = .)$ dans le cas général à l'aide de deux équations par la formule des probabilités totales

$$P(X = 0)p_{1|0} + P(X = 1)p_{1|1} = p$$

$$P(X = 0)p_{0|0} + P(X = 1)p_{0|1} = 1 - p$$

Dans ce cas binaire on a $R^* = E(\min(p_{1|X}, 1 - p_{1|X}))$. On impose

$$\min(p, 1 - p) = \min(p_{1|0}, 1 - p_{1|0})P(X = 0) + \min(p_{1|1}, 1 - p_{1|1})P(X = 1).$$

Prenons $p = P(Y = 1) = 1/4$ on a donc

$$P(X = 0)p_{1|0} + P(X = 1)p_{1|1} = 1/4$$

$$P(X = 0)p_{0|0} + P(X = 1)p_{0|1} = 3/4$$

De plus, on veut une loi conditionnelle tel que $R^* = \min(p, 1 - p) = 1/4$.

$$1/4 = \min(p_{1|0}, 1 - p_{1|0})P(X = 0) + \min(p_{1|1}, 1 - p_{1|1})P(X = 1).$$

On obtient un système de 3 équations à 3 inconnues $p_{1|0}, p_{1|1}, P(X = 0)$:

$$\begin{cases} P(X = 0)p_{1|0} + P(X = 1)p_{1|1} = 1/4 \\ P(X = 0)p_{0|0} + P(X = 1)p_{0|1} = 3/4 \\ \min(p_{1|0}, 1 - p_{1|0})P(X = 0) + \min(p_{1|1}, 1 - p_{1|1})P(X = 1) = 1/4 \end{cases}$$

Supposons $P(X = 1) = 1/2 = P(X = 0)$

$$\begin{cases} p_{1|0} + p_{1|1} = 1/2 \\ p_{0|0} + p_{0|1} = 3/2 \\ \min(p_{1|0}, 1 - p_{1|0}) + \min(p_{1|1}, 1 - p_{1|1}) = 1/2 \end{cases}$$

Si on choisit $p_{1|1} < 1/2 < 1 - p_{1|1}$ alors on obtient

$$\begin{cases} p_{1|0} = 1/2 - p_{1|1} \\ \min(p_{1|0}, 1 - p_{1|0}) = 1/2 - p_{1|1} \end{cases} \Rightarrow p_{1|0} = \min(p_{1|0}, 1 - p_{1|0}) \Rightarrow p_{1|0} < 1/2.$$

Par exemple $p_{1|1} = 1/8$ donc $p_{1|0} = 3/8$ ainsi

		conditionnelle		inconditionnelle
		$X = 0$	$X = 1$	
Proba.	$Y = 0$	$5/8 = p_{0 0}$	$7/8 = p_{0 1}$	$1 - p = \frac{1}{2} \frac{5}{8} + \frac{1}{2} \frac{7}{8} = 3/4$
	$Y = 1$	$3/8 = p_{1 0}$	$1/8 = p_{1 1}$	$p = \frac{1}{2} \frac{3}{8} + \frac{1}{2} \frac{1}{8} = 1/4$
Risque	$g^*(X)$	0	0	$\min(1 - p, p) = 1/4$
	$\eta^*(X)$	3/8	1/8	
	R^*	$3/8 \times 1/2 + 1/8 \times 1/2$		

Exercice 2 (Classifieur bayésien (suite))

On suppose que X admet une densité f par rapport à la mesure de Lebesgue de \mathbb{R}^d et l'on rappelle que

$$R^* = \mathbb{E}[\min\{\eta(X), 1 - \eta(X)\}],$$

où $\eta(x) = \mathbb{P}(Y = 1|X = x)$ est la fonction de régression. Soit f_0 (resp. f_1) la densité de X conditionnellement à $Y = 0$ (resp. $Y = 1$) c'est-à-dire que pour $j = 0, 1$, pour toute fonction borélienne φ , $\mathbb{E}[\varphi(X)|Y = j] = \int_{\mathbb{R}^d} \varphi(x) f_j(x) dx$. Soit $p = \mathbb{P}(Y = 1)$.

0. Montrer que

(a)

$$f(x) = p f_1(x) + (1 - p) f_0(x).$$

(b) En déduire une écriture de $\eta(x)$ en fonction de f_0 , f_1 et p .

1. Écrire R^* en fonction de f_0 , f_1 et p .

2. Dans le cas $p = 1/2$, en déduire que

$$R^* = \frac{1}{2} - \frac{1}{4} \int_{\mathbb{R}^d} |f_0(x) - f_1(x)| dx.$$

On pourra utiliser la formule $\min\{a, b\} = \frac{a+b}{2} - \frac{|a-b|}{2}$.

Réponse de l'exercice 2.

0. (a) **Méthode 1**

Soit $\varphi : \mathbb{R}^d \rightarrow \mathbb{R}$ une fonction continue à support compact. Nous avons

$$\begin{aligned} \mathbb{E}[\varphi(X)] &= \mathbb{E}[\varphi(X)|Y = 1]\mathbb{P}(Y = 1) + \mathbb{E}[\varphi(X)|Y = 0]\mathbb{P}(Y = 0) \\ &= \int_{\mathbb{R}^d} p\varphi(x)f_1(x) + (1 - p)\varphi(x)f_0(x) dx, \end{aligned}$$

d'où le résultat.

Méthode 2

$$\begin{aligned} P(X \leq x, Y = y) &= P(X \leq x|Y = y)P(Y = y) = F_X^{Y=y}(x)p^y(1 - p)^{1-y} \\ \Rightarrow F_X(x) &= \sum_y P(X \leq x, Y = y) = P(X \leq x, Y = 1) + P(X \leq x, Y = 0) = F_X^{Y=1}(x)p + F_X^{Y=0}(x)(1 - p) \end{aligned}$$

$$\Rightarrow f_X(x) = \frac{\partial F_X(x)}{\partial x} = p f_1(x) + (1 - p) f_0(x).$$

(b) Par définition de l'espérance conditionnelle, la fonction η est la seule fonction mesurable telle que, pour toute fonction mesurable $\varphi : \mathbb{R}^d \rightarrow \mathbb{R}$ nous avons

$$\mathbb{E}[\varphi(X)\eta(X)] = \mathbb{E}[\varphi(X)\mathbf{1}_{\{Y=1\}}].$$

Or

$$\mathbb{E}[\varphi(X)\mathbf{1}_{\{Y=1\}}] = \mathbb{E}[\varphi(X)|Y = 1]\mathbb{P}(Y = 1) = p \int_{\mathbb{R}^d} \varphi(x) f_1(x) dx.$$

Comme

$$\mathbb{E}[\varphi(X)\eta(X)] = \int_{\mathbb{R}^d} \varphi(x)\eta(x)f(x)dx,$$

et que le résultat est vrai pour toute fonction borélienne. Cela nous donne

$$\eta(x)f(x) = pf_1(x) \text{ ou encore } \eta(x) = \frac{pf_1(x)}{f(x)},$$

en tout point x où f ne s'annule pas. C'est la formule de Bayes appliquée à notre cas particulier.

1. Par le résultat de la question 0.(ii),

$$R^* = \int_{\mathbb{R}^d} \min\{\eta(x), 1 - \eta(x)\}f(x)dx = \int_{\mathbb{R}^d} \min\left\{\frac{pf_1(x)}{f(x)}, 1 - \frac{pf_1(x)}{f(x)}\right\} f(x)dx$$

Puis par 0.(i), $1 - \frac{pf_1(x)}{f(x)} = \frac{(1-p)f_0(x)}{f(x)}$. D'où

$$R^* = \int_{\mathbb{R}^d} \min\{pf_1(x), (1-p)f_0(x)\} dx.$$

2. Dans le cas $p = 1/2$, nous avons

$$R^* = \frac{1}{2} \int_{\mathbb{R}^d} \min\{f_1(x), f_0(x)\} dx = \frac{1}{2} \int_{\mathbb{R}^d} \frac{f_0(x) + f_1(x)}{2} - \frac{|f_0(x) - f_1(x)|}{2} dx,$$

et le résultat provient du fait que, si $p = 1/2$, $\frac{f_0(x)+f_1(x)}{2} = f(x)$.

Le risque de Bayes est donc lié à la distance entre f_0 et f_1 : plus la distance L^1 entre f_0 et f_1 est élevée et plus le risque de Bayes est petit.

Exercice 3 (Un classifieur universellement consistant)

Soit (X, Y) un couple de loi \mathbb{P} avec $X \in \{1, \dots, p\}$ et $Y \in \{0, 1\}$. On note toujours g^* le classifieur de Bayes et R^* son risque de classification. On se propose d'étudier la règle de classification suivante, étant donné $\mathcal{D}_n = ((X_1, Y_1), \dots, (X_n, Y_n))$,

$$\hat{g}(x) = \mathbf{1}_{\{\hat{\eta}(x) > \frac{1}{2}\}},$$

avec

$$\hat{\eta}(x) = \frac{1}{\text{card}\{i, X_i = x\}} \sum_{i, X_i = x} Y_i$$

et $\hat{\eta}(x) = 0$ si $\text{card}\{i, X_i = x\} = 0$.

1. Déterminer $\hat{\eta}$ et \hat{g} sur les données suivantes \mathcal{D}_7 :

i	1	2	3	4	5	6	7
X_i	3	4	4	3	3	4	5
Y_i	0	1	0	1	0	1	0

2. Montrer que \hat{g} est universellement consistante.

Réponse de l'exercice 3.

1. $\hat{\eta}(x, \mathcal{D}_n)$ est en fait la moyenne empirique de la variable réponse Y quand la variable explicative X vaut x . Notons $n_x = \text{card}\{i, X_i = x\} \leq n$ par construction et $\sum_x n_x = n$. Typiquement pour le tableau suivant, on obtient

i	X_i	Y_i
1	3	0
2	4	1
3	4	0
4	3	1
5	3	0
6	4	1
7	5	0

$$\Rightarrow \hat{\eta}(x, \mathcal{D}_7) = \begin{cases} 1/3 & \text{si } x = 3 \\ 2/3 & \text{si } x = 4 \\ 0 & \text{si } x = 5 \end{cases} \Rightarrow g(x, \mathcal{D}_7) = \begin{cases} 0 & \text{si } x = 3 \\ 1 & \text{si } x = 4 \\ 0 & \text{si } x = 5 \end{cases}$$

2. Par la loi des grands nombres, quand $P(X = x) > 0$, i.e. $P(X = 0) \in (0, 1), P(X = 1) \in (0, 1)$ on a

$$\frac{1}{n} \sum_{i=1}^n 1_{X_i=x} Y_i \xrightarrow[n \rightarrow +\infty]{p.s.} E(Y 1_{X=x}), \quad \frac{1}{n} \sum_{i=1}^n 1_{X_i=x} \xrightarrow[n \rightarrow +\infty]{p.s.} E(1_{X=x}) = P(X = x)$$

Or $E(Y 1_{X=x}) = E(1_{X=x})E(Y|X = x) = P(X = x)\eta(x)$. En effet pour A un évènement tel que $P(A) > 0$, la variable conditionnelle $Z|A$ a pour fonction de répartition

$$P(Z \leq x|A) = \frac{P(Z \leq x \cap A)}{P(A)} = F_Z^A(x),$$

et pour espérance

$$\begin{cases} E(Z|A) = \int_{\mathbb{R}} x dF_Z^A(x) = \int_{\mathbb{R}} x \frac{dP(Z \leq x \cap A)}{P(A)} = \int_{\mathbb{R}} x 1_A(x) \frac{dF_Z(x)}{P(A)} & \Rightarrow E(Z|A) = \frac{E(Z 1_A)}{P(A)}. \\ E(Z 1_A) = \int_{\mathbb{R}} x 1_A(x) dF_Z(x) \end{cases}$$

On a bien $E(Z 1_A) = P(A)E(Z|A)$ qui est vrai aussi pour $P(A) = 0$.

Notons $n_X = \text{card}\{i = 1, \dots, n, X_i = X\} \leq n$ le cardinal (aléatoire) du nombre d'observations tel que la variable explicative vaut X .

Etudions l'écart absolu moyen

$$\begin{aligned} E(|\hat{\eta}(X, \mathcal{D}_n) - \eta(X)|) &= E(|\hat{\eta}(X, \mathcal{D}_n) - \eta(X)| \times 1_{n_X \neq 0}) + E(|\hat{\eta}(X, \mathcal{D}_n) - \eta(X)| \times 1_{n_X = 0}) \\ &= E(|\hat{\eta}(X, \mathcal{D}_n) - \eta(X)| \times 1_{n_X \neq 0}) + E(\eta(X) \times 1_{n_X = 0}) \\ &\leq E(|\hat{\eta}(X, \mathcal{D}_n) - \eta(X)| | n_X \neq 0) E(1_{n_X \neq 0}) + E(1_{n_X = 0}) \\ &= E(|\hat{\eta}(X, \mathcal{D}_n) - \eta(X)| | n_X \neq 0) (1 - P(n_X = 0)) + P(n_X = 0) \end{aligned}$$

Or $P(n_X = 0) \xrightarrow[n \rightarrow +\infty]{} 0$ et $E(|\hat{\eta}(X, \mathcal{D}_n) - \eta(X)| | n_X \neq 0) \xrightarrow[n \rightarrow +\infty]{} 0$ donc

$$E(|\hat{\eta}(X, \mathcal{D}_n) - \eta(X)|) \leq E(|\hat{\eta}(X, \mathcal{D}_n) - \eta(X)| | n_X \neq 0) (1 - P(n_X = 0)) + P(n_X = 0) \xrightarrow[n \rightarrow +\infty]{} 0.$$

Exercice 4 (Classification et maximum de vraisemblance)

On suppose que la loi P du couple (X, Y) est donnée de la manière suivante : $Y \sim \mathcal{B}(p)$ pour un paramètre $p \in]0, 1[$ puis $X \in \mathbb{R}^d$ est donnée par sa loi conditionnelle sachant Y :

$$X|Y \sim \mathcal{N}(V_Y, \Sigma)$$

où Σ est une matrice définie positive et V_0, V_1 sont deux vecteurs distincts de \mathbb{R}^d .

1. Donner la loi jointe de (X, Y) ainsi que les lois marginales.
2. Déterminer la fonction de régression

$$\eta(x) = E(Y|X = x).$$

3. En déduire la forme du classifieur de Bayes, g^* , et montrer que son risque de classification s'écrit

$$R^* = pP\left(Z > \delta/2 + \delta^{-1} \log\left(\frac{p}{1-p}\right)\right) + (1-p)P\left(Z < -\delta/2 + \delta^{-1} \log\left(\frac{p}{1-p}\right)\right)$$

où $\delta = \|\Sigma^{-1/2}(V_1 - V_0)\|$ et $Z \sim \mathcal{N}(0, 1)$.

4. En pratique, on dispose d'un n -échantillon (X_i, Y_i) , $1 \leq i \leq n$, de la loi P . On suppose que l'on est dans un cas où l'on connaît p et Σ et on se propose d'estimer V_0 et V_1 par maximum de vraisemblance. Donner la forme des estimateurs \hat{V}_0 et \hat{V}_1 ainsi obtenus.
5. En déduire un estimateur de la fonction de régression, $\hat{\eta}$ et une règle de classification, \hat{g} .
6. Montrer que $R(\hat{g}) \rightarrow R^*$ en probabilité, quand $n \rightarrow \infty$.

7. Montrer que \hat{g} n'est pas universellement consistante. Il suffira de fabriquer une autre loi P' telle que si (X_i, Y_i) et (X, Y) sont iid de loi P' , alors $R(\hat{g})$ ne converge pas vers R^* .

Réponse de l'exercice 4.

1. La loi jointe de (X, Y) est caractérisée par sa densité par rapport à la mesure $\lambda_d \otimes \mu$ où λ_d est la mesure de Lebesgue sur \mathbb{R}^d et μ est la mesure de comptage sur $\{0, 1\}$ ($\mu(S) = \text{card}(S)$ pour tout $S \subseteq \{0, 1\}$). Cette densité est une fonction $f_{X,Y} : \mathbb{R}^d \times \{0, 1\} \rightarrow \mathbb{R}$ telle que, pour toute fonction test mesurable $\varphi : \mathbb{R}^d \times \{0, 1\} \rightarrow \mathbb{R}$

$$\begin{aligned} \mathbb{E}[\varphi(X, Y)] &= \int_{\mathbb{R}^d \times \{0, 1\}} \varphi(x, y) f_{X,Y}(x, y) d(\lambda_d \otimes d\mu)(x, y) \\ &= \sum_{j \in \{0, 1\}} \int_{\mathbb{R}^d} \varphi(x, j) f_{X,Y}(x, j) dx \\ &= \int_{\mathbb{R}^d} \varphi(x, 0) f_{X,Y}(x, 0) dx + \int_{\mathbb{R}^d} \varphi(x, 1) f_{X,Y}(x, 1) dx. \end{aligned}$$

Or

$$\begin{aligned} \mathbb{E}[\varphi(X, Y)] &= \mathbb{E}[\varphi(X, Y)|Y = 0] \times \mathbb{P}(Y = 0) + \mathbb{E}[\varphi(X, Y)|Y = 1] \times \mathbb{P}(Y = 1) \\ &= \int_{\mathbb{R}^d} \varphi(x, 0) f_X^{Y=0}(x) dx \times (1 - p) + \int_{\mathbb{R}^d} \varphi(x, 1) f_X^{Y=0}(x) dx \times p \end{aligned}$$

où

$$f_X^{Y=y}(x) = \frac{1}{\sqrt{2\pi \det(\Sigma)}} \exp\left(-\frac{(x - V_y)^T \Sigma^{-1} (x - V_y)}{2}\right)$$

est la loi densité de X conditionnellement à $Y = y$ (pour $y \in \{0, 1\}$). Ce qui nous donne

$$f_{X,Y}(x, y) = \begin{cases} (1 - p) f_X^{Y=0}(x) & \text{si } y = 0 \\ p f_X^{Y=1}(x) & \text{si } y = 1. \end{cases}$$

Pour obtenir les lois marginales il suffit d'intégrer la densité jointe. Intégrer par rapport à la mesure de comptage revient à sommer sur toutes les valeurs possibles. On obtient

$$f_X(x) = (1 - p) f_X^{Y=0}(x) + p f_X^{Y=1}(x)$$

pour la loi de X . Pour Y , d'après l'énoncé, nous avons,

$$P(Y = 1) = p, P(Y = 0) = 1 - p.$$

2. pour un problème de classification binaire, on a, d'après la formule de Bayes,

$$\eta(x) = E(Y|X = x) = P(Y = 1|X = x) = \frac{p f_X^{Y=1}(x)}{f_X^{Y=1}(x)p + f_X^{Y=0}(x)(1 - p)}$$

3. dans ce cadre le classificateur de Bayes est

$$g^*(x) = 1_{\eta(x) > 1/2} = 1_{f_X^{Y=1}(x)p > f_X^{Y=0}(x)(1-p)}$$

Son risque vaut

$$\begin{aligned} R^* &= E(\min(\eta(X), 1 - \eta(X))) = E\left(\min\left(\frac{p f_X^{Y=1}(X)}{f_X(X)}, 1 - \frac{p f_X^{Y=1}(X)}{f_X(X)}\right)\right) \\ &= E\left(\min\left(\frac{p f_X^{Y=1}(X)}{f_X(X)}, \frac{(1 - p) f_X^{Y=0}(X)}{f_X(X)}\right)\right) = E\left(\frac{\min(p f_X^{Y=1}(X), (1 - p) f_X^{Y=0}(X))}{f_X(X)}\right) \\ &= \int \frac{\min(p f_X^{Y=1}(x), (1 - p) f_X^{Y=0}(x))}{f_X(x)} f_X(x) dx = \int \min(p f_X^{Y=1}(x), (1 - p) f_X^{Y=0}(x)) dx \end{aligned}$$

Calculons le minimum

$$\begin{aligned}
& pf_X^{Y=1}(x) < (1-p)f_X^{Y=0}(x) \\
\Leftrightarrow p \frac{1}{\sqrt{2\pi\det(\Sigma)}} \exp\left(-\frac{(x-V_1)^T \Sigma^{-1}(x-V_1)}{2}\right) < (1-p) \frac{1}{\sqrt{2\pi\det(\Sigma)}} \exp\left(-\frac{(x-V_0)^T \Sigma^{-1}(x-V_0)}{2}\right) \\
& \Leftrightarrow \log \frac{p}{1-p} - \frac{(x-V_1)^T \Sigma^{-1}(x-V_1)}{2} < -\frac{(x-V_0)^T \Sigma^{-1}(x-V_0)}{2} \\
& \Leftrightarrow 2 \log \frac{p}{1-p} < (x-V_1)^T \Sigma^{-1}(x-V_1) - (x-V_0)^T \Sigma^{-1}(x-V_0)
\end{aligned}$$

Or

$$\begin{aligned}
(x-V_0)^T \Sigma^{-1}(x-V_0) &= (x-V_1+V_1-V_0)^T \Sigma^{-1}(x-V_1+V_1-V_0) \\
&= (x-V_1)^T \Sigma^{-1}(x-V_1+V_1-V_0) + (V_1-V_0)^T \Sigma^{-1}(x-V_1+V_1-V_0) \\
&= (x-V_1)^T \Sigma^{-1}(x-V_1) + 2(x-V_1)^T \Sigma^{-1}(V_1-V_0) + \underbrace{(V_1-V_0)^T \Sigma^{-1}(V_1-V_0)}_{\delta^2}
\end{aligned}$$

Ainsi

$$pf_X^{Y=1}(x) < (1-p)f_X^{Y=0}(x) \Leftrightarrow 2 \log \frac{p}{1-p} < -2(x-V_1)^T \Sigma^{-1}(V_1-V_0) - \delta^2$$

Autrement dit

$$\begin{aligned}
R^* &= \int \mathbf{1}_{2 \log \frac{p}{1-p} + \delta^2 < -2(x-V_1)^T \Sigma^{-1}(V_1-V_0)} pf_X^{Y=1}(x) dx + \int \mathbf{1}_{2 \log \frac{p}{1-p} + \delta^2 > -2(x-V_1)^T \Sigma^{-1}(V_1-V_0)} (1-p) f_X^{Y=0}(x) dx \\
&= pE\left(\mathbf{1}_{2 \log \frac{p}{1-p} + \delta^2 < -2(X-V_1)^T \Sigma^{-1}(V_1-V_0)} | Y=1\right) + (1-p)E\left(\mathbf{1}_{2 \log \frac{p}{1-p} + \delta^2 > -2(X-V_1)^T \Sigma^{-1}(V_1-V_0)} | Y=0\right) \\
&= p \underbrace{P\left(2 \log \frac{p}{1-p} + \delta^2 < -2(X-V_1)^T \Sigma^{-1}(V_1-V_0) | Y=1\right)}_{p_{1,X}} \\
&\quad + (1-p) \underbrace{P\left(2 \log \frac{p}{1-p} + \delta^2 > -2(X-V_1)^T \Sigma^{-1}(V_1-V_0) | Y=0\right)}_{p_{0,X}}
\end{aligned}$$

Comme $X|Y=1 \sim \mathcal{N}(V_1, \Sigma)$ on a

$$(V_1-V_0)^T \Sigma^{-1}(X-V_1) | Y=1 \sim \mathcal{N}((V_1-V_1)^T \Sigma^{-1}(V_1-V_0), (V_1-V_0)^T \Sigma^{-1} \Sigma \Sigma^{-1}(V_1-V_0)) = \mathcal{N}(0, \delta^2).$$

Donc $(V_1-V_0)^T \Sigma^{-1}(X-V_1) | Y=1$ a même loi que δZ avec $Z \sim \mathcal{N}(0, 1)$ et

$$\begin{aligned}
p_{1,X} &= P\left(2 \log\left(\frac{p}{1-p}\right) + \delta^2 < -2\delta Z\right) = P\left(\log\left(\frac{p}{1-p}\right)/\delta + \delta/2 < -Z\right) \\
&= P\left(\log\left(\frac{p}{1-p}\right)/\delta + \delta/2 < Z\right)
\end{aligned}$$

Comme $(V_1-V_0)^T \Sigma^{-1}(X-V_1) = (V_1-V_0)^T \Sigma^{-1}(X-V_0) - \delta^2$ on a

$$(V_1-V_0)^T \Sigma^{-1}(X-V_1) | Y=0 = (V_1-V_0)^T \Sigma^{-1}(X-V_0) - \delta^2 | Y=0 \sim \mathcal{N}(-\delta^2, \delta^2).$$

Donc $(V_1-V_0)^T \Sigma^{-1}(X-V_1) | Y=0$ a même loi que $-\delta^2 + \delta Z$ avec $Z \sim \mathcal{N}(0, 1)$ et

$$\begin{aligned}
p_{0,X} &= P\left(2 \log \frac{p}{1-p} + \delta^2 > 2\delta^2 - 2\delta Z\right) = P\left(2 \log \frac{p}{1-p} - \delta^2 > -2\delta Z\right) \\
&= P\left(\log \frac{p}{1-p} / \delta - \delta/2 > -Z\right) = P\left(\log \frac{p}{1-p} / \delta - \delta/2 > Z\right)
\end{aligned}$$

Ainsi

$$R^* = pP\left(Z > \delta/2 + \delta^{-1} \log\left(\frac{p}{1-p}\right)\right) + (1-p)P\left(Z < -\delta/2 + \delta^{-1} \log\left(\frac{p}{1-p}\right)\right)$$

4. Nous pouvons réécrire la densité du couple (X, Y) de façon suivante

$$\begin{aligned} f_{X,Y}(x, y) &= (f_X^{Y=1}(x)p)^y (f_X^{Y=0}(x)(1-p))^{1-y} \\ &= \left(\frac{p}{\sqrt{2\pi\det(\Sigma)}} \exp\left(-\frac{(x-V_1)^T \Sigma^{-1}(x-V_1)}{2}\right) \right)^y \left(\frac{1-p}{\sqrt{2\pi\det(\Sigma)}} \exp\left(-\frac{(x-V_0)^T \Sigma^{-1}(x-V_0)}{2}\right) \right)^{1-y} \\ &= \frac{p^y(1-p)^{1-y}}{\sqrt{2\pi\det(\Sigma)}} \exp\left(-y\frac{(x-V_1)^T \Sigma^{-1}(x-V_1)}{2} - (1-y)\frac{(x-V_0)^T \Sigma^{-1}(x-V_0)}{2}\right) \end{aligned}$$

La log-vraisemblance s'écrit donc

$$\begin{aligned} \mathcal{L}(V_0, V_1) &= -n \log(\sqrt{2\pi\det(\Sigma)}) + \sum_{i=1}^n y_i \log(p) + \sum_{i=1}^n (1-y_i) \log(1-p) \\ &\quad - \sum_{i=1}^n y_i \frac{(x_i-V_1)^T \Sigma^{-1}(x_i-V_1)}{2} - \sum_{i=1}^n (1-y_i) \frac{(x_i-V_0)^T \Sigma^{-1}(x_i-V_0)}{2} \\ \frac{\partial \mathcal{L}(V_0, V_1)}{\partial V_0} &= - \sum_{i=1}^n (1-y_i) \Sigma^{-1}(x_i-V_0) = -\Sigma^{-1} \sum_{i=1}^n (1-y_i)(x_i-V_0) \end{aligned}$$

Annulons la dérivée en utilisant Σ^{-1} est inversible (puisque c'est l'inverse de Σ)

$$\frac{\partial \mathcal{L}(V_0, V_1)}{\partial V_0} = 0 \Leftrightarrow \sum_{i=1}^n (1-y_i)(x_i-V_0) = 0 \Leftrightarrow V_0 = \frac{\sum_{i=1}^n (1-y_i)x_i}{n - \sum_{i=1}^n y_i}$$

De même

$$\frac{\partial \mathcal{L}(V_0, V_1)}{\partial V_1} = -\Sigma^{-1} \sum_{i=1}^n y_i(x_i-V_1) \Rightarrow V_1 = \frac{\sum_{i=1}^n y_i x_i}{\sum_{i=1}^n y_i}$$

Les deux estimateurs sont donc

$$\hat{V}_0 = \frac{\sum_{i=1}^n (1-Y_i)X_i}{n - \sum_{i=1}^n Y_i} = \frac{\sum_{i, Y_i=0} X_i}{n - \sum_{i=1}^n Y_i}, \quad \hat{V}_1 = \frac{\sum_{i=1}^n Y_i X_i}{\sum_{i=1}^n Y_i} = \frac{\sum_{i, Y_i=1} X_i}{\sum_{i, Y_i=1} Y_i}.$$

5. en utilisant la question 1, on a la fonction de régression

$$\hat{\eta}(x) = \frac{p f_X^{Y=1}(x; \hat{V}_1)}{f_X^{Y=1}(x; \hat{V}_1)p + f_X^{Y=0}(x; \hat{V}_0)(1-p)} = \frac{p \exp\left(-\frac{(x-\hat{V}_1)^T \Sigma^{-1}(x-\hat{V}_1)}{2}\right)}{p \exp\left(-\frac{(x-\hat{V}_1)^T \Sigma^{-1}(x-\hat{V}_1)}{2}\right) + (1-p) \exp\left(-\frac{(x-\hat{V}_0)^T \Sigma^{-1}(x-\hat{V}_0)}{2}\right)}$$

et le classifieur

$$\hat{g}(x) = 1_{\hat{\eta}(x) > 1/2} = 1_{p \exp\left(-\frac{(x-\hat{V}_1)^T \Sigma^{-1}(x-\hat{V}_1)}{2}\right) > (1-p) \exp\left(-\frac{(x-\hat{V}_0)^T \Sigma^{-1}(x-\hat{V}_0)}{2}\right)}$$

6. par la loi des grands nombres, on a

$$\hat{V}_1 = \frac{1/n \sum_{i=1}^n Y_i X_i}{1/n \sum_{i=1}^n Y_i} \xrightarrow[n \rightarrow +\infty]{p.s.} \frac{E(YX)}{E(Y)} = \frac{E(YX|Y=1)P(Y=1) + E(YX|Y=0)P(Y=0)}{P(Y=1)} = E(X|Y=1) + 0$$

De même $\hat{V}_0 \xrightarrow[n \rightarrow +\infty]{p.s.} E(X|Y=0)$. Etant une fonction continue, la fonction de régression converge aussi

$\hat{\eta}(X, \mathcal{D}_n) \xrightarrow[n \rightarrow +\infty]{p.s.} \eta(X)$. De plus

$$0 \leq |\hat{\eta}(X, \mathcal{D}_n) - \eta(X)| \leq |\hat{\eta}(X, \mathcal{D}_n)| + |\eta(X)| \leq 2$$

Par le théorème de convergence dominée, $E(|\hat{\eta}(X, \mathcal{D}_n) - \eta(X)|) \xrightarrow[n \rightarrow +\infty]{} 0$. En utilisant l'exercice 2, on a le résultat souhaité.

7. Considérons $X \sim \mathcal{N}(V_0, \Sigma)$ indépendamment de $Y \sim \mathcal{B}(p)$ avec $0 < p < 1/2$ par exemple $1/3$. Par indépendance on a $\eta(x) = E(Y) = p$ et $g^*(x) = \mathbf{1}_{\eta(x) > 1/2} = 0$ donc

$$R^* = P(g^*(X) \neq Y) = P(Y \neq 0) = P(Y = 1) = p.$$

Pour le classifieur, on a par indépendance

$$R(\hat{g}) = P(\hat{g}(X) \neq Y) = P(\hat{g}(X) = 0, Y = 1) + P(\hat{g}(X) = 1, Y = 0) = pP(\hat{g}(X) = 0) + (1-p)P(\hat{g}(X) = 1).$$

Or

$$P(\hat{g}(X) = 0) = P\left(p \exp\left(-\frac{(X - \hat{V}_1)^T \Sigma^{-1} (X - \hat{V}_1)}{2}\right) \leq (1-p) \exp\left(-\frac{(X - \hat{V}_0)^T \Sigma^{-1} (X - \hat{V}_0)}{2}\right)\right)$$

La convergence presque sûre des estimateurs garantit que

$$P(\hat{g}(X) = 0) \xrightarrow[n \rightarrow +\infty]{p.s.} P\left(Z - \delta/2 \leq \delta^{-1} \log\left(\frac{1-p}{p}\right)\right) =: p_{\delta,p}$$

Pour $P(\hat{g}(X) = 1)$, on trouve

$$P(\hat{g}(X) = 1) = 1 - P(\hat{g}(X) = 0) = 1 - p_{\delta,p}.$$

car $\{\hat{g}(X) = 0\}^c = \{\hat{g}(X) = 1\}$. Donc

$$R(\hat{g}) \xrightarrow[n \rightarrow +\infty]{p.s.} p(1 - p_{\delta,p}) + (1-p)p_{\delta,p}.$$

S'il était universellement consistant alors les deux limites devraient être égales

$$p(1 - p_{\delta,p}) + (1-p)p_{\delta,p} = p$$

C'est impossible avec $p < 1/2$.

Exercice 5 (Risque de classification pondéré)

Soit $\mathcal{D}_n = \{(X_i, Y_i), i = 1, \dots, n\}$ un échantillon de copies du couple de variables aléatoires (X, Y) avec $X \in \mathbb{R}^d$ et $Y \in \{0, 1\}$. Soient $\omega_0, \omega_1 > 0$ fixé, le risque de classification pondéré R_ω d'un classifieur h est la quantité

$$R_\omega(h) = \omega_0 \mathbb{P}(Y = 0, h(X) = 1 | \mathcal{D}_n) + \omega_1 \mathbb{P}(Y = 1, h(X) = 0 | \mathcal{D}_n).$$

1. Soit $\eta(X) = \mathbb{E}[Y | X = x]$ la fonction de régression. Montrer que

$$\mathbb{E}[R_\omega(h)] \geq \mathbb{E}[\min\{\omega_0(1 - \eta(X)), \omega_1 \eta(X)\}].$$

2. Soit

$$h_\omega^*(x) = \mathbf{1}_{\{\eta(x) > \frac{\omega_0}{\omega_0 + \omega_1}\}},$$

calculer $R_\omega(h_\omega^*)$ et commenter.

3. Montrer que, pour tout classifieur h

$$\mathbb{E}[R_\omega(h)] - R_\omega(h_\omega^*) = \mathbb{E}[|\omega_1 \eta(X) - \omega_0(1 - \eta(X))| \mathbf{1}_{\{h(X) \neq h_\omega^*(X)\}}].$$

4. Soit $\hat{\eta}$ un estimateur de la fonction η et

$$h_\omega(x) = \mathbf{1}_{\{\hat{\eta}(x) > \frac{\omega_0}{\omega_0 + \omega_1}\}}.$$

(a) Montrer que $h_\omega(x) \neq h_\omega^*(x)$ implique $|\hat{\eta}(x) - \eta(x)| \geq \left|\eta(x) - \frac{\omega_0}{\omega_0 + \omega_1}\right|$.

(b) En déduire que

$$\mathbb{E}[R_\omega(h_\omega)] - R_\omega(h_\omega^*) \leq (\omega_0 + \omega_1) \mathbb{E}[|\hat{\eta}(x) - \eta(x)|].$$

Réponse de l'exercice 5.

1.

$$\mathbb{E}[R_\omega(h)] = \omega_0 \mathbb{P}(Y = 0, h(X) = 1) + \omega_1 \mathbb{P}(Y = 1, h(X) = 0)$$

Comme $h(X)$ est (X, \mathcal{D}_n) -mesurable, et Y indépendant de \mathcal{D}_n , nous avons (en remarquant que $\mathbf{1}_{\{h(X)=1\}} = h(X)$),

$$\begin{aligned} \mathbb{P}(Y = 0, h(X) = 1) &= \mathbb{E}[\mathbb{P}(Y = 0, h(X) = 1 | X, \mathcal{D}_n)] = \mathbb{E}[\mathbf{1}_{\{h(X)=1\}} \mathbb{P}(Y = 0 | X, \mathcal{D}_n)] \\ &= \mathbb{E}[h(X) \mathbb{P}(Y = 0 | X)] = \mathbb{E}[h(X)(1 - \mathbb{P}(Y = 1 | X))] \\ &= \mathbb{E}[h(X)(1 - \eta(X))]. \end{aligned}$$

De même,

$$\mathbb{P}(Y = 1, h(X) = 0) = \mathbb{E}[(1 - h(X))\eta(X)].$$

Nous avons donc

$$\begin{aligned} \mathbb{E}[R_\omega(h)] &= \mathbb{E}[\omega_0 h(X)(1 - \eta(X)) + \omega_1 (1 - h(X))\eta(X)] \\ &\geq \mathbb{E}[h(X) \min\{\omega_0(1 - \eta(X)), \omega_1 \eta(X)\} + (1 - h(X)) \min\{\omega_0(1 - \eta(X)), \omega_1 \eta(X)\}] \\ &= \mathbb{E}[\min\{\omega_0(1 - \eta(X)), \omega_1 \eta(X)\}]. \end{aligned}$$

Attention ici au fait que R_ω n'est en général pas égal au risque de classification vu en cours (sauf dans le cas $\omega_0 = \omega_1 = 1$). Aucun résultat du cours ne permet donc d'affirmer que le classifieur $h^*(x) = \mathbf{1}_{\{\eta(x) > 1/2\}}$ (classifieur de Bayes) minimise R_ω .

2. Comme h_ω^* ne dépend pas de \mathcal{D}_n ,

$$R_\omega(h_\omega^*) = \omega_0 \mathbb{P}(Y = 0, h_\omega^*(X) = 1) + \omega_1 \mathbb{P}(Y = 1, h_\omega^*(X) = 0).$$

Nous avons,

$$\mathbb{P}(Y = 0, h_\omega^*(X) = 1) = \mathbb{E}[\mathbf{1}_{\{h_\omega^*(X)=1\}} \mathbb{P}(Y = 0 | X)] = \mathbb{E}\left[\mathbf{1}_{\{\eta(X) > \frac{\omega_0}{\omega_0 + \omega_1}\}} (1 - \eta(X))\right],$$

et de même,

$$\mathbb{P}(Y = 1, h_\omega^*(X) = 0) = \mathbb{E}\left[\mathbf{1}_{\{\eta(X) \leq \frac{\omega_0}{\omega_0 + \omega_1}\}} \eta(X)\right].$$

Or $\eta(X) > \frac{\omega_0}{\omega_0 + \omega_1} \Leftrightarrow \omega_1 \eta(X) > \omega_0(1 - \eta(X))$ d'où

$$\begin{aligned} R_\omega(h_\omega^*) &= \mathbb{E}[\omega_0(1 - \eta(X)) \mathbf{1}_{\{\omega_1 \eta(X) > \omega_0(1 - \eta(X))\}} + \omega_1 \eta(X) \mathbf{1}_{\{\omega_1 \eta(X) \leq \omega_0(1 - \eta(X))\}}] \\ &= \mathbb{E}[\min\{\omega_0(1 - \eta(X)), \omega_1 \eta(X)\}]. \end{aligned}$$

D'après 1., tout classifieur h vérifie $\mathbb{E}[R_\omega(h)] \geq R_\omega(h_\omega^*)$, h_ω^* est donc le classifieur minimisant le risque pondéré. C'est le classifieur de Bayes pour ce risque-là.

3. Soit h un classifieur, en conditionnant par X à l'intérieur de l'espérance et en remarquant que $\mathbf{1}_{\{h(X)=1\}} = h(X)$ et $\mathbf{1}_{\{h(X)=0\}} = 1 - h(X)$ (de même en remplaçant h par h_ω^*), nous obtenons,

$$\begin{aligned} \mathbb{E}[R_\omega(h)] - R_\omega(h_\omega^*) &= \mathbb{E}[\omega_0 \mathbf{1}_{\{Y=0\}} (\mathbf{1}_{\{h(X)=1\}} - \mathbf{1}_{\{h_\omega^*(X)=1\}}) + \omega_1 \mathbf{1}_{\{Y=1\}} (\mathbf{1}_{\{h(X)=0\}} - \mathbf{1}_{\{h_\omega^*(X)=0\}})] \\ &= \mathbb{E}[\omega_0(1 - \eta(X)) (h(X) - h_\omega^*(X)) + \omega_1 \eta(X) ((1 - h(X)) - (1 - h_\omega^*(X)))] \\ &= \mathbb{E}[(\omega_0(1 - \eta(X)) - \omega_1 \eta(X)) (h(X) - h_\omega^*(X))] \\ &= \mathbb{E}[(\omega_0(1 - \eta(X)) - \omega_1 \eta(X)) (-\mathbf{1}_{\{h(X) \neq h_\omega^*(X), h_\omega^*(X)=1\}}) \\ &\quad + \mathbb{E}[(\omega_0(1 - \eta(X)) - \omega_1 \eta(X)) \mathbf{1}_{\{h(X) \neq h_\omega^*(X), h_\omega^*(X)=0\}}]. \end{aligned}$$

Or $h_\omega^*(X) = 1 \Leftrightarrow \omega_0(1 - \eta(X)) - \omega_1 \eta(X) < 0$ d'où le résultat.

4. (a) Soit x tel que $h_\omega(x) \neq h_\omega^*(x)$, deux cas sont possibles.

Soit $h_\omega(x) = 1$ et $h_\omega^*(x) = 0$ dans ce cas-là, $\hat{\eta}(x) > \frac{\omega_0}{\omega_0 + \omega_1}$ et $\eta(x) \leq \frac{\omega_0}{\omega_0 + \omega_1}$. Nous avons donc

$$|\hat{\eta}(x) - \eta(x)| = \hat{\eta}(x) - \eta(x) > \frac{\omega_0}{\omega_0 + \omega_1} - \eta(x) = \left| \eta(x) - \frac{\omega_0}{\omega_0 + \omega_1} \right|.$$

Soit $h_\omega(x) = 0$ et $h_\omega^*(x) = 1$ dans ce cas-là, $\hat{\eta}(x) \leq \frac{\omega_0}{\omega_0 + \omega_1}$ et $\eta(x) > \frac{\omega_0}{\omega_0 + \omega_1}$. Nous avons donc

$$|\hat{\eta}(x) - \eta(x)| = \eta(x) - \hat{\eta}(x) \geq \eta(x) - \frac{\omega_0}{\omega_0 + \omega_1} = \left| \eta(x) - \frac{\omega_0}{\omega_0 + \omega_1} \right|.$$

(b) Par 3.

$$\begin{aligned} \mathbb{E}[R_\omega(h_\omega)] - R_\omega(h_\omega^*) &= \mathbb{E}[\omega_0(1 - \eta(X)) - \omega_1\eta(X) | \mathbf{1}_{\{h_\omega(X) \neq h_\omega^*(X)\}}] \\ &= \mathbb{E}[\omega_0 - (\omega_0 + \omega_1)\eta(X) | \mathbf{1}_{\{h_\omega(X) \neq h_\omega^*(X)\}}] \\ &\leq (\omega_0 + \omega_1) \mathbb{E}[|\hat{\eta}(X) - \eta(X)| \mathbf{1}_{\{h_\omega(X) \neq h_\omega^*(X)\}}] \leq (\omega_0 + \omega_1) \mathbb{E}[|\hat{\eta}(X) - \eta(X)|], \end{aligned}$$

où l'avant-dernière inégalité provient de (a).

Exercice 6 (Vitesse de convergence sous hypothèse de régularité de η)

On suppose que les couples i.i.d. $(X, Y), (X_1, Y_1), \dots, (X_n, Y_n)$ sont à valeurs dans $[0, 1]^d \times \{0, 1\}$ pour $d \geq 3$. L'hypothèse sur la loi de probabilité P est que $\eta(\cdot)$ est L -Lipschitz, pour une certaine constante $L > 0$. On note μ la première marginale de P .

Soit \hat{g}_k le classifieur des k -plus proches voisins : $\hat{g}_k(x) = \mathbf{1}_{\hat{\eta}(x) \geq 1/2}$ où $\hat{\eta}(x) = \frac{1}{k} \sum_{m=1}^k Y_{i_m(x)}$ et $m \mapsto i_m(x)$ est une permutation $\sigma(X_1, \dots, X_n)$ -mesurable telle que $\|X_{i_1(x)} - x\| \leq \dots \leq \|X_{i_n(x)} - x\|$.

1. Pour $0 < \varepsilon < 1$, montrer qu'il existe une partition de $[0, 1]^d$ formée de moins de $(\lfloor \frac{1}{\varepsilon} + 1 \rfloor)^d$ éléments de diamètre maximal au plus $\sqrt{d}\varepsilon$.

On note $A_1, \dots, A_{m_\varepsilon}$ les éléments de cette partition.

2. On rappelle que $X_{i_1(x)}$ est le plus proche voisin de x parmi X_1, \dots, X_n . Montrer que

$$\forall x \in A_j, \quad P\left(\|X_{i_1(x)} - x\| > \varepsilon\sqrt{d}\right) \leq [1 - \mu(A_j)]^n.$$

3. En déduire que $(X$ est maintenant aléatoire)

$$P\left(\|X_{i_1(X)} - X\| > \varepsilon\sqrt{d}\right) \leq m_\varepsilon \sup_{\alpha > 0} \alpha \exp(-\alpha n) \leq \left(\frac{2}{\varepsilon}\right)^d \frac{1}{n}$$

4. En déduire que pour $c = \frac{4d^2}{d-2}$ on obtient

$$E(\|X_{i_1(X)} - X\|^2) \leq 2d \int_0^1 \min\left\{1, \left(\frac{2}{\varepsilon}\right)^d \frac{1}{n}\right\} \varepsilon d\varepsilon \leq \frac{c}{n^{2/d}}$$

5. Vérifier que

$$\begin{aligned} &E[(\hat{\eta}(x) - \eta(x))^2] \\ &= E\left[\left(\frac{1}{k} \sum_{m=1}^k Y_{i_m(x)} - \eta(X_{i_m(x)})\right)^2\right] + E\left[\left(\frac{1}{k} \sum_{m=1}^k \eta(x) - \eta(X_{i_m(x)})\right)^2\right] \\ &= A(x) + B(x) \end{aligned}$$

6. Montrer que pour $m \neq m'$ (attention $i_m(x)$ et $i_{m'}(x)$ sont aléatoires) :

$$E[(Y_{i_m(x)} - \eta(X_{i_m(x)}))(Y_{i_{m'}(x)} - \eta(X_{i_{m'}(x)})) | X_1, \dots, X_n] = 0$$

7. En déduire que $A(x) \leq \frac{1}{k}$

8. Soient S_1, \dots, S_k des sous ensembles disjoints de $\{X_1, \dots, X_n\}$ tous de taille $\lfloor n/k \rfloor$ et $\hat{X}_j(x)$ le plus proche voisin de x dans S_j . Montrer que

$$B(x) \leq \frac{L^2}{k} \sum_{j=1}^k E\|\hat{X}_j(x) - x\|^2$$

9. En déduire que $B(x) \leq cL^2 \left(\frac{k}{n}\right)^{2/d}$ et en déduire la vitesse de convergence des k -plus proches voisins.
 10. Qu'obtient-on pour $d = 1$ et $d = 2$?

Réponse de l'exercice 6.

1. Quelques exemples en dimension 2 pour 3 ϵ différents
 Définissons la partition de la manière suivante

$$A_j = \prod_{l=1}^d [j_l \epsilon, (j_l + 1)\epsilon] \cap [0, 1]^d, \quad j_l \in \mathbb{N}.$$

Le nombre d'éléments de cette partition est $\lfloor 1 + 1/\epsilon \rfloor^d$. On pose k_ϵ le nombre de découpage par dimension, c'est à dire $k_\epsilon = \lfloor 1 + 1/\epsilon \rfloor$. Donc on a

$$\forall l = 1, \dots, d, 0 \leq j_l < k_\epsilon \text{ et } j_l \in \mathbb{N} \Leftrightarrow \forall l = 1, \dots, d, j_l = \{0, \dots, k_\epsilon - 1\} \Leftrightarrow \forall l = 1, \dots, d, j_l = \{0, \dots, \lfloor 1/\epsilon \rfloor\}.$$

Typiquement pour $\epsilon = 3/4$,

$$A_1 = [0, \epsilon]^2, A_2 = [\epsilon, 2\epsilon] \times [0, \epsilon] \cap [0, 1]^2, A_3 = [0, \epsilon] \times [\epsilon, 2\epsilon] \cap [0, 1]^2, A_4 = [\epsilon, 2\epsilon] \times [\epsilon, 2\epsilon] \cap [0, 1]^2.$$

Posons $m_\epsilon = \lfloor \frac{1}{\epsilon} + 1 \rfloor^d$. On a pour le volume

$$\forall j = 1, \dots, m_\epsilon, \text{Vol}(A_j) \leq \text{Vol}([0, \epsilon]^d) = \epsilon^d,$$

et pour la longueur de la diagonale

$$\forall j = 1, \dots, m_\epsilon, \text{ldiag}(A_j) \leq \text{ldiag}([0, \epsilon]^d) = \sqrt{\epsilon^2 + \dots + \epsilon^2} = \epsilon\sqrt{d}.$$

2. Le plus proche voisin vérifie

$$X_{i_1} = \arg \min_{i=1, \dots, n} \|X_i - x\| \Leftrightarrow \forall i = 1, \dots, n, \|X_{i_1} - x\| \leq \|X_i - x\|$$

Ainsi la probabilité peut se majorer en utilisant l'indépendance

$$\forall x \in A_j, P(\|X_{i_1} - x\| > \epsilon\sqrt{d}) = P(\forall i = 1, \dots, n, \|X_i - x\| > \epsilon\sqrt{d}) = \prod_{i=1, \dots, n} P(\|X_i - x\| > \epsilon\sqrt{d}).$$

L'identique distribution entraine

$$\forall x \in A_j, P(\|X_{i_1} - x\| > \epsilon\sqrt{d}) \leq P(\|X - x\| > \epsilon\sqrt{d})^n$$

Comme $\text{ldiag}(A_j) \leq \epsilon\sqrt{d}$, on a

$$\forall x \in A_j, P(\|X - x\| > \epsilon\sqrt{d})^n \leq P(X \notin A_j)^n = (1 - P(X \in A_j))^n = (1 - \mu(A_j))^n$$

Donc on obtient

$$\forall x \in A_j, P(\|X_{i_1} - x\| > \epsilon\sqrt{d}) \leq (1 - \mu(A_j))^n.$$

3. Passons par l'espérance conditionnelle

$$P(\|X_{i_1} - X\| > \epsilon\sqrt{d}) = E\left(P(\|X_{i_1} - X\| > \epsilon\sqrt{d} | X)\right) = \sum_{j=1}^{m_\epsilon} E\left(P(\|X_{i_1} - X\| > \epsilon\sqrt{d} | X) 1_{X \in A_j}\right)$$

Utilisons la question précédente pour tout x ,

$$P(\|X_{i_1} - X\| > \epsilon\sqrt{d} | X = x) \leq (1 - \mu(A_j))^n$$

Donc

$$P(\|X_{i_1} - X\| > \epsilon\sqrt{d}) \leq \sum_{j=1}^{m_\epsilon} E\left((1 - \mu(A_j))^n 1_{X \in A_j}\right) = \sum_{j=1}^{m_\epsilon} (1 - \mu(A_j))^n P(X \in A_j) = \sum_{j=1}^{m_\epsilon} (1 - \mu(A_j))^n \mu(A_j).$$

Or

$$(1 - \mu(A_j))^n \mu(A_j) = \mu(A_j) e^{n \ln(1 - \mu(A_j))} \leq \mu(A_j) e^{-n\mu(A_j)}$$

car $\ln(1 - x) \leq -x$. Par conséquent

$$P(\|X_{i_1} - X\| > \epsilon\sqrt{d}) \leq \sum_{j=1}^{m_\epsilon} \mu(A_j) e^{-n\mu(A_j)}.$$

Etudions la fonction

$$r(p) = pe^{-np}, r'(p) = e^{-np} - npe^{-np} = e^{-np}(1 - np)$$

Donc r est décroissante pour $p > 1/n$ et croissante pour $p < 1/n$. donc

$$\sup_p r(p) = r(1/n) = \frac{1}{n} e^{-1}.$$

Ainsi

$$P(\|X_{i_1} - X\| > \epsilon\sqrt{d}) \leq m_\epsilon \sup_{\mu(A_j)} r(\mu(A_j)) = m_\epsilon \frac{1}{n} e^{-1} = \left[\frac{1}{\epsilon} + 1\right]^d \frac{1}{n} e^{-1}.$$

Enfin en utilisant $e^{-1} < 1$ et

$$0 < \epsilon < 1 \Rightarrow 1 < 1/\epsilon < +\infty \Rightarrow 1 + 1/\epsilon < 1/\epsilon + 1/\epsilon \Rightarrow [1 + 1/\epsilon] < 2/\epsilon$$

on trouve

$$P(\|X_{i_1} - X\| > \epsilon\sqrt{d}) \leq (2/\epsilon)^d / n$$

4. Comme $N = \|X_{i_1} - X\|$ est une variable positive, son espérance est l'intégrale de sa fonction de survie

$$E(N^2) = \int_0^\infty P(N^2 > t) dt = \int_0^\infty P(\|X_{i_1} - X\|^2 > t) dt$$

En posant $t = u^2 d$, $dt = 2ud du$ et

Le résultat de la question 3. nous donne

$$E(N^2) = \int_0^{+\infty} P(\|X_{i_1} - X\| > u\sqrt{d}) 2ud du \leq 2d \int_0^{+\infty} \min\left\{1, \left(\frac{2}{u}\right)^d \frac{1}{n}\right\} udu.$$

Or

$$\left(\frac{2}{u}\right)^d \frac{1}{n} \geq 1 \Leftrightarrow \left(\frac{2}{u}\right)^d \geq n \Leftrightarrow \frac{2}{u} \geq n^{1/d} \Leftrightarrow u \leq 2n^{-1/d}.$$

Cela nous donne :

$$\min\left\{1, \left(\frac{2}{u}\right)^d \frac{1}{n}\right\} = \begin{cases} 1 & \text{si } u \leq 2n^{-1/d}, \\ \left(\frac{2}{u}\right)^d \frac{1}{n} & \text{sinon.} \end{cases}$$

D'où, comme $d > 2$,

$$E(N^2) \leq 2d \int_0^{2n^{-1/d}} udu + 2d \frac{2^d}{n} \int_{2n^{-1/d}}^{+\infty} u^{1-d} du = 2d \times 2n^{-2/d} + \frac{2^{d+1}d}{n} \times \frac{2^{2-d} n^{\frac{d-2}{d}}}{d-2},$$

ce qui donne bien le résultat voulu car $n^{(d-2)/d}/n = n^{(d-2)/d-1} = n^{-d/2}$.

5.

$$\begin{aligned} & E[(\hat{\eta}(x) - \eta(x))^2] \\ &= E\left[\left(\frac{1}{k} \sum_{m=1}^k Y_{i_m(x)} - \eta(x)\right)^2\right] \\ &= E\left[\left(\frac{1}{k} \sum_{m=1}^k Y_{i_m(x)} - \eta(X_{i_m(x)})\right)^2\right] + E\left[\left(\frac{1}{k} \sum_{m=1}^k \eta(x) - \eta(X_{i_m(x)})\right)^2\right] \\ &\quad + 2E\left[\left(\frac{1}{k} \sum_{m=1}^k Y_{i_m(x)} - \eta(X_{i_m(x)})\right) \left(\frac{1}{k} \sum_{m=1}^k \eta(x) - \eta(X_{i_m(x)})\right)\right]. \end{aligned}$$

Or

$$\begin{aligned}
& E \left[\left(\frac{1}{k} \sum_{m=1}^k Y_{i_m(x)} - \eta(X_{i_m(x)}) \right) \left(\frac{1}{k} \sum_{m=1}^k \eta(x) - \eta(X_{i_m(x)}) \right) \right] \\
&= E \left[\mathbb{E} \left[\left(\frac{1}{k} \sum_{m=1}^k Y_{i_m(x)} - \eta(X_{i_m(x)}) \right) \left(\frac{1}{k} \sum_{m=1}^k \eta(x) - \eta(X_{i_m(x)}) \right) \middle| X_1, \dots, X_n \right] \right] \\
&= E \left[\mathbb{E} \left[\left(\frac{1}{k} \sum_{m=1}^k Y_{i_m(x)} - \eta(X_{i_m(x)}) \right) \middle| X_1, \dots, X_n \right] \left(\frac{1}{k} \sum_{m=1}^k \eta(x) - \eta(X_{i_m(x)}) \right) \right] \\
&= E \left[\left(\frac{1}{k} \sum_{m=1}^k \mathbb{E}[Y_{i_m(x)} | X_1, \dots, X_n] - \eta(X_{i_m(x)}) \right) \left(\frac{1}{k} \sum_{m=1}^k \eta(x) - \eta(X_{i_m(x)}) \right) \right] \\
&= E \left[\left(\frac{1}{k} \sum_{m=1}^k \sum_{i=1}^n \mathbf{1}_{\{i_m(x)=i\}} (\mathbb{E}[Y_{i_m(x)} | X_1, \dots, X_n] - \eta(X_{i_m(x)})) \right) \left(\frac{1}{k} \sum_{m=1}^k \eta(x) - \eta(X_{i_m(x)}) \right) \right] \\
&= E \left[\left(\frac{1}{k} \sum_{m=1}^k \sum_{i=1}^n \mathbf{1}_{\{i_m(x)=i\}} (\mathbb{E}[Y_i | X_1, \dots, X_n] - \eta(X_i)) \right) \left(\frac{1}{k} \sum_{m=1}^k \eta(x) - \eta(X_{i_m(x)}) \right) \right] \\
&= E \left[\left(\frac{1}{k} \sum_{m=1}^k \sum_{i=1}^n \mathbf{1}_{\{i_m(x)=i\}} \underbrace{(\mathbb{E}[Y_i | X_i] - \eta(X_i))}_{=0} \right) \left(\frac{1}{k} \sum_{m=1}^k \eta(x) - \eta(X_{i_m(x)}) \right) \right] \\
&= 0,
\end{aligned}$$

d'où le résultat.

6. L'idée est similaire

$$\begin{aligned}
& E [(Y_{i_m(x)} - \eta(X_{i_m(x)}))(Y_{i_{m'}(x)} - \eta(X_{i_{m'}(x)})) | X_1, \dots, X_n] \\
&= \mathbb{E} \left[\sum_{i=1}^n \sum_{i' \neq i} \mathbf{1}_{\{i_m(x)=i, i_{m'}(x)=i'\}} (Y_{i_m(x)} - \eta(X_{i_m(x)}))(Y_{i_{m'}(x)} - \eta(X_{i_{m'}(x)})) | X_1, \dots, X_n \right] \\
&= \mathbb{E} \left[\sum_{i=1}^n \sum_{i' \neq i} \mathbf{1}_{\{i_m(x)=i, i_{m'}(x)=i'\}} (Y_i - \eta(X_i))(Y_{i'} - \eta(X_{i'})) | X_1, \dots, X_n \right] \\
&= \sum_{i=1}^n \sum_{i' \neq i} \mathbf{1}_{\{i_m(x)=i, i_{m'}(x)=i'\}} \mathbb{E} [(Y_i - \eta(X_i))(Y_{i'} - \eta(X_{i'})) | X_1, \dots, X_n] \\
&\quad \text{comme } i_m(x) \text{ est } \sigma(X_1, \dots, X_n) \text{ - mesurable} \\
&= \sum_{i=1}^n \sum_{i' \neq i} \mathbf{1}_{\{i_m(x)=i, i_{m'}(x)=i'\}} \mathbb{E} [Y_i - \eta(X_i) | X_i] \mathbb{E} [(Y_{i'} - \eta(X_{i'})) | X_{i'}] \\
&\quad \text{car } X_1, \dots, X_n \text{ sont indépendants} \\
&= 0.
\end{aligned}$$

7.

$$\begin{aligned}
A(x) &= \mathbb{E} \left[\left(\frac{1}{k} \sum_{m=1}^k Y_{i_m(x)} - \eta(X_{i_m(x)}) \right)^2 \right] \\
&= \frac{1}{k^2} \sum_{m=1}^k \mathbb{E} [(Y_{i_m(x)} - \eta(X_{i_m(x)}))^2] + \frac{1}{k^2} \sum_{m \neq m'} \mathbb{E} [(Y_{i_m(x)} - \eta(X_{i_m(x)})) (Y_{i_{m'}(x)} - \eta(X_{i_{m'}(x)}))]
\end{aligned}$$

En conditionnant par rapport à X_1, \dots, X_n et en utilisant 6., on vérifie que le second terme est nul. Nous

avons donc

$$\begin{aligned}
A(x) &= \frac{1}{k^2} \sum_{m=1}^k \mathbb{E} \left[(Y_{i_m(x)} - \eta(X_{i_m(x)}))^2 \right] = \frac{1}{k^2} \sum_{m=1}^k \mathbb{E} \left[\sum_{i=1}^n \mathbf{1}_{\{i_m(x)=i\}} (Y_{i_m(x)} - \eta(X_{i_m(x)}))^2 \right] \\
&= \frac{1}{k^2} \sum_{m=1}^k \mathbb{E} \left[\sum_{i=1}^n \mathbf{1}_{\{i_m(x)=i\}} (Y_i - \eta(X_i))^2 \right] \\
&= \frac{1}{k^2} \sum_{m=1}^k \mathbb{E} \left[\sum_{i=1}^n \mathbf{1}_{\{i_m(x)=i\}} \mathbb{E} \left[(Y_i - \eta(X_i))^2 \mid X_1, \dots, X_n \right] \right] \\
&= \frac{1}{k^2} \sum_{m=1}^k \mathbb{E} \left[\sum_{i=1}^n \mathbf{1}_{\{i_m(x)=i\}} \mathbb{E} \left[(Y_i - \mathbb{E}[Y_i \mid X_i])^2 \mid X_1, \dots, X_n \right] \right] \\
&= \frac{1}{k^2} \sum_{m=1}^k \mathbb{E} \left[\sum_{i=1}^n \mathbf{1}_{\{i_m(x)=i\}} \mathbb{E} \left[(Y_i - \mathbb{E}[Y_i \mid X_i])^2 \mid X_i \right] \right] \\
&\quad \text{car } X_1, \dots, X_n \text{ sont indépendants} \\
&= \frac{1}{k^2} \sum_{m=1}^k \mathbb{E} \left[\sum_{i=1}^n \mathbf{1}_{\{i_m(x)=i\}} \text{Var}(Y_i \mid X_i) \right].
\end{aligned}$$

Comme, conditionnellement à X_i , Y_i suit une loi biomiale de paramètre $\eta(X_i)$, on a

$$\text{Var}(Y_i \mid X_i) = \eta(X_i)(1 - \eta(X_i)) \leq 1,$$

d'où le résultat.

8. Par Cauchy-Schwarz

$$B(x) \leq E \left[\left(\frac{1}{k} \sum_{m=1}^k \eta(x) - \eta(X_{i_m(x)}) \right)^2 \right] \leq E \left[\frac{1}{k^2} \left(\sum_{m=1}^k 1^2 \right) \times \left(\sum_{m=1}^k (\eta(x) - \eta(X_{i_m(x)}))^2 \right) \right]$$

Comme η est L -höldérienne,

$$(\eta(x) - \eta(X_{i_m(x)}))^2 \leq L^2 \|x - X_{i_m(x)}\|^2.$$

D'où

$$B(x) \leq \frac{L^2}{k} \sum_{m=1}^k \mathbb{E}[\|x - X_{i_m(x)}\|^2] \leq \frac{L^2}{k} \sum_{m=1}^k \mathbb{E}[\|x - \hat{X}_m(x)\|^2]$$

en remarquant que $X_{i_1(x)}, \dots, X_{i_m(x)}$ sont les k -plus proches voisins de x et donc $\sum_{m=1}^k \|x - X_{i_m(x)}\|^2 \leq \sum_{m=1}^k \|x - X_{j_m}\|^2$ pour tout choix possible de j_1, \dots, j_m distincts dans $\{1, \dots, n\}$.

9. En utilisant la question 4., on remarque que, comme $\hat{X}_j(x)$ est le plus proche voisin de x parmi les $\lfloor n/k \rfloor$ éléments de S_j on a

$$\mathbb{E}[\|x - \hat{X}_j(x)\|^2] \leq \frac{c}{\lfloor n/k \rfloor^{2/d}} \leq \frac{c}{(n/k)^{2/d}},$$

d'où le résultat.

En combinant les résultats des questions 5., 7. et 8. , nous avons

$$\mathbb{E}[(\hat{\eta}(x) - \eta(x))^2] = A(x) + B(x) \leq \frac{1}{k} + cL^2 \left(\frac{k}{n} \right)^{2/d}.$$

Le minimum de la fonction $g(x) = x + cL^2 x^{-2/d} n^{-2/d}$ est atteint pour x de l'ordre de constante $\times n^{-\frac{2/d}{2/d+1}} = \text{constante} \times n^{-\frac{2}{d+2}}$ et donc $k = 1/x$ de l'ordre de constante $\times n^{\frac{2}{d+2}}$. Choisissons $k = \lfloor n^{2/(d+2)} \rfloor$, nous avons

$$\mathbb{E}[(\hat{\eta}(x) - \eta(x))^2] \leq \text{constante} \times n^{-\frac{2}{d+2}},$$

c'est la vitesse de convergence de l'estimateur des k -plus proches voisins.

10. $d = 1$, vitesse en $n^{-2/3}$, $d = 2$, vitesse en $n^{-1/2}$. On remarque que la vitesse de convergence se dégrade lorsque la dimension de l'espace augmente.