Exercice 1 (K-Nearest Neighbours):

Consider the KNN rule with equal weights for classification. One observes n i.i.d. replications $D_n = \{(X_i, Y_i),\ i \leq n\}$ of a random pair (X, Y) withe $X \in \mathbb{R}^d$ and $Y \in \{0, 1\}$. Recall that the KNN classifier given the dataset D_n is

$$g_{k,n}(x) = \begin{cases} +1 & \text{if } \sum_{i=1}^{k} Y_{(i)}(x) \geq 1/2 \\ 0 & \text{otherwise.} \end{cases}$$

Where $Y_{(i)}$ is the label of the i^{th} nearest neighbour $X_{(i)}$ of the point x. Recall also that the regression function is defined as $\eta(x) = \mathbb{P}[Y = 1|X]_{X=x}$. We assume here that X is uniformly distributed over $[0, 1]$ and that $\eta(x) = 1/3$ for all $x \in [0, 1]$.

1. Compute the joint probability $\mathbb{P}[X \in [a, b], Y = 1]$ for all $0 < a < b$.
2. Give the expression for the Bayes classifier g^* For a generic regression function $\eta(x)$ and for this particular problem. Give its error risk $\mathbb{P}[g^*(X) \neq Y]$ (the Bayes error) in general and in this particular case.
3. Consider the training set

   $$(X_1 = 0.4, Y_1 = 0), (X_2 = 0.2, Y_2 = 0), (X_3 = 0.7, Y_3 = 1)$$

   Compute the k-NN classifier (for all x) in the case $k = 1$ and $k = 3$.
4. Compute in each case the expected classification error $R(g_{knn}) = \mathbb{P}[g_{knn}(X) \neq Y]$ for a new observation (X, Y). Compare with the Bayes error
5. Consider now another model : X is again uniformly distributed over $[0, 1]$ but now $\eta(x) = \mathbb{1}_{(1/2,1]}(x)$. Repeat questions 1,2,3,4 with the same dataset as in Q3.

1) $\mathbb{P}(X \in [a,b],\ Y=1) = \mathbb{E}\left[\mathbb{P}(Y=1,\ X \in (a,b) \mid X)\right]$

$\mathbb{E}\left(\mathbb{E}\left(\mathbb{1}(Y=1,\ X \in [a,b] \mid X)\right)\right)$

$= \mathbb{E}\left(\mathbb{P}(Y=1 \mid X)\, \mathbb{1}_{(a,b)}(X)\right)$

$= \int_a^b 1/3 \, dx = \frac{b-a}{3}$.

2) if $\eta^*(x) = \mathbb{P}(Y=1 \mid X=x)$ $g^*(x) = \begin{cases} 1 & \text{if } \eta^*(x) \geq 1/2 \\ 0 & \text{otherwise} \end{cases}$

here : $\forall x,\ \eta^*(x) = 1/3 < 1/2$ so $g^*(x) = 0$.

Bayes error $\mathbb{P}(Y \neq g^*(X)) = \mathbb{P}(Y=1) = \frac{1}{3}$

in general $\mathbb{P}(Y \neq g^*(X)) =$

$\mathbb{E}\left(\mathbb{P}(Y \neq g^*(X) \mid X)\right)$

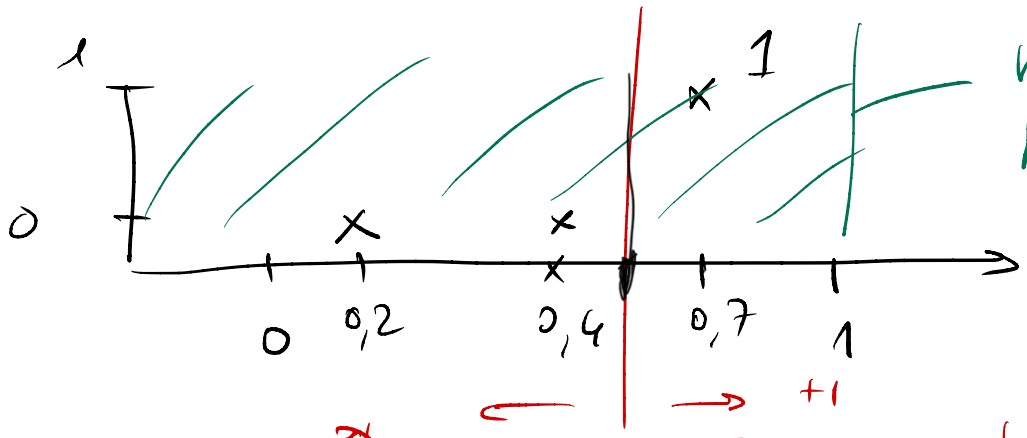
$= \int_{\eta^*(x) < 1/2} \mathbb{P}(Y=1)\, P_X(dx) + \int_{\eta^* > 1/2} \mathbb{P}(Y=0)\, P_X(dx) =$

$$P(Y \neq g^*(X)) = \int_{\eta^* < \frac{1}{2}} \eta^*(x)\, dP_X(x) + \int_{\eta^* \geq \frac{1}{2}} (1-\eta^*)(x)\, dP_X(x)$$

$$= \int \min(\eta^*, 1-\eta^*)(x)\, dP_X(x)$$

$$= \boxed{\mathbb{E}\left(\min\left(\eta^*, 1-\eta^*(X)\right)\right)}$$

3)



$k=3$: all points are classified as 0

0 ← → +1

$x = 0,55$ 1-nn boundary

$k = 1$

$k=3$: $g_{knn}(x) \equiv 0$

$k=1$ $g_{knn}(x) = \mathbb{1}\{x \geq 0,55\}$.

errors $R(g_{1,nn}) = P(Y \neq g_{knn}(X))$

$= P(Y=1, X < 0,55)$

$+ P(Y=0, X \geq 0,55)$

$= \frac{0,55}{3} + \frac{0,45 \times 2}{3}$

$= \frac{1,45}{3} = \boxed{0,48\ldots\ (> \frac{1}{3})}$

$R(g_{3,nn}) = P(Y=1) = \frac{1}{3}$

5] $\eta(x) = \mathbb{P}(Y=1 \mid X=x)$

$= \mathbb{1}_{\{x \geq 1/2\}}$

5-1) si $b < 1/2$: $P(X \in (a,b), Y=1) = 0$

si $a > 1/2$: $\int_a^b 1 - x = b - a$

si $a < 1/2 < b$: $= P(X \in (1/2, b), Y=1) = b - \frac{1}{2}$

5-2 $g^*(x) = \eta(x)$

$P(Y = g^*(x)) = \mathbb{E}[\min(\eta, 1-\eta)] = 0$

5-3 $R(g_{1nn}) = P(Y=1, g_{1nn}(X)=0) = P(X \in [0.5, 0.5]) = 0$

$R(g_{3nn}) = P(Y=-1) = 0.5$

① $X_n \xrightarrow{P} X$: $\forall \varepsilon > 0 \quad \lim_{n\to\infty} P(\|X_n - X\| \geq \varepsilon) \to 0$

conv in probability

② $X_n \xrightarrow{a.s} X$: almost sure convergence

$\bullet\ (\exists \Omega \in \mathcal{A})\ P(\Omega) = 1$

$(\forall \omega \in \Omega)\ \|X_n(\omega) - X(\omega)\| \to 0$ as $n \to \infty$

$\Leftrightarrow \bullet\ P(\lim_{n\to\infty} \|X_n - X\| = 0) = 1$

$X_n \xrightarrow{P} X$: X is constant

$\Leftrightarrow (\forall \varepsilon > 0)\ P(\limsup_n \{\omega, \|X_n(\omega) - X(\omega)\| \geq \varepsilon\}) = 0$

③ $X_n \xrightarrow{D} X$: $\mathbb{E}(\varphi(X_n)) \to \mathbb{E}(\varphi(X))$ $\forall \varphi\ C_b$

$\mathbb{E}(e^{itX_n}) \to \mathbb{E}(e^{itX})$

Exercice 2 (Key lemma explaining K-nn consistency):
Let P_X be the law of X on $\mathbb{R}^d$. Let x in the support of P_X, that is for all $\epsilon > 0$, for all ball $B(x,\epsilon)$ of radius ϵ centered at x, $P_X(B(x,\epsilon)) > 0$. Let k_n be a sequence of integers such that $k_n/n \to 0$. Let $(X_i, i \in \mathbb{N})$ be an iid sequence distributed as X.

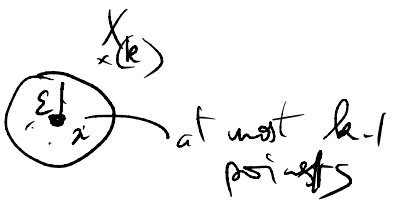
- Show that, almost surely , $\|X_{(k_n,n)} - x\| \to 0$ where $X_{(k_n)}$ is the k_n'th nearest neighbour of x among $(X_1, \dots, X_n)$.

Reminder : $X_n \xrightarrow{as} X$ iff $\mathrm{ess\,sup} P\left(\limsup \{\omega : |X_n(\omega) - X(\omega)| > \varepsilon\} \right) = 0$

where for a sequence of sets $(A_n)_{n\geq 1}$

$\limsup A_n = \inf_{k\geq 1} \sup_{n \geq k} A_n$

$= \bigcap_{k\geq 1} \bigcup_{n\geq k} A_n$



at most $k-1$ points

here : let $\varepsilon > 0$ and $A_n(\varepsilon) = \{ \|X_{(k_n,n)} - x\| > \varepsilon \}$ (k-th neighbour)

$A_n(\varepsilon) \subset \sum_{i=1}^{n} \mathbb{1}\{X_i \in B(x,\varepsilon)\} \leq k_n - 1 < k_n$

$\subset \left\{ \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{X_i \in B(x,\varepsilon)\} \leq \frac{k_n}{n} \right\}$

or LGN $\Rightarrow$ avec proba 1 (avec Ω_1 s.t. $\mathbb{P}(\Omega_1) = 1$)

$\frac{1}{n} \sum \mathbb{1}\{X_i \in B(x,\varepsilon)\} \longrightarrow p_\varepsilon = P(X \in B(x,\varepsilon)) > 0$

Now, $\forall \omega \in \Omega_1$, $\exists n_0$ s.t. $\forall n \geq n_0$, $\exists n_0$ s.t. $\forall n \geq n_0$

$\frac{1}{n} \sum \mathbb{1}\{X_i(\omega) \in B(x,\varepsilon)\} \geq \frac{p_\varepsilon}{2} > \frac{k_n}{n}$

$\exists n_1$ s.t. $n \geq n_1 \Rightarrow \frac{k_n}{n} \leq \frac{p_\varepsilon}{2}$

$n_2 = \max(n_0, n_1) \Rightarrow n \geq n_2$

$\Rightarrow \omega \notin \bigcup_{n \geq n_0} A_n$ ($\frac{k_{n_0}}{n_0} < \frac{p_\varepsilon}{2}$, n_0 chosen such that)

$\Rightarrow \omega \notin \limsup A_n$.

Whence $\limsup A_n \subset \Omega_1^c$ of probability 0.

$P(\limsup A_n) = 0$

Exercice 3 (Nadaraya-Watson Regression):
Consider a regression problem for a random pair $(X, Y) \in \mathbb{R}^d \times \mathbb{R}$ with target Y and covariate X. We assume that $\mathbb{E}[Y^2] < \infty$. The goal is to approach $m(x) = \mathbb{E}[Y|X = x]$. We assume that the pair (X, Y) has a density $f(x, y)$ with respect to the Lebesgue measure on $\mathbb{R}^{d+1}$.

1. Write $m(x)$ as an integral involving the f and the marginal density f_X of X.
2. Given kernels K_x, K_y of order 1 for density estimation of X and Y, (thus $\int uK(u)\mathrm{d}u = 0$ and $\int K(u)\mathrm{d}u = 1$) define the product kernel density estimate as

$$\widehat{f}(x, y) = \frac{1}{nh^2} \sum_{i=1}^{n} K_x(\frac{X_i - x}{h}) K_y(\frac{Y_i - y}{k}).$$

1- $m(x) = \mathbb{E}(Y|X=x) = \int_R y \, f_{Y|X}(y|x)\, dy$

$= \int_R y \, \frac{f(x,y)}{f(x)}\, dy$

2-

(a) Recall the expression for the Kernel density estimate $\widehat{f}_X$ of f_X based on K_x and a dataset $X_{1:n}$.

(b) Show that the Nadaraya-Watson estimate $\widehat{m}(x)$ based on K_x is the plug-in estimate of $\mathbb{E}[Y|X = x]$ based on the expression found in Question 1) up to replacing f_X, f with the kernel density estimates $\widehat{f}_X$ and $\widehat{f}$.

a) $\hat{f}_X(x) = \frac{1}{nh} \sum_{i=1}^{n} K_x\left(\frac{X_i - x}{h}\right)$

b) recall $\hat{m}(x) = \frac{\sum_{i=1}^{n} Y_i K_x\left(\frac{X_i - x}{h}\right)}{\sum_{i=1}^{n} K_x\left(\frac{X_i - x}{h}\right)}$ with Nadaraya-Watson

$= \frac{\frac{1}{nh} \sum_{i=1}^{n} Y_i K_x\left(\frac{X_i - x}{h}\right)}{\frac{1}{nh} \sum_{i=1}^{n} K_x\left(\frac{X_i - x}{h}\right)}$.

let $\tilde{N}(x) = \frac{1}{nh} \sum_{i=1}^{n} Y_i K_x\left(\frac{X_i - x}{h}\right)$ be the numerator.

On the other hand the plug-in estimate is:

$$\hat{m}_{PI}(x) = \int_{\mathbb{R}} y \, \frac{\hat{f}(x,y)}{\hat{f}(x)} \, dy$$

$$= \frac{\hat{N}(x)}{\frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right) Y_i}, \text{ where:}$$

$$\hat{N}(x) = \int_{\mathbb{R}} y \, \frac{1}{nh^2}\sum_{i=1}^{n} K_x\left(\frac{X_i - x}{h}\right) K_y\left(\frac{Y_i - y}{h}\right) dy$$

$$= \frac{1}{nh^2}\sum_{i=1}^{n} K_x\left(\frac{X_i - x}{h}\right) \int_{\mathbb{R}} y \, K_y\left(\frac{Y_i - y}{h}\right) dy$$

change variables: $u = \frac{Y_i - y}{h}$

$$= \frac{1}{nh^2}\sum_{i=1}^{n} K_x\left(\frac{X_i - x}{h}\right) \int_{\mathbb{R}} (Y_i - hu) K_y(u) \, du \cdot h \qquad \left[dy = h\,du\right]$$

$$= \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right) Y_i$$

$$= \tilde{N}(x).$$

whence $\hat{m}_{PI}(x) = \hat{m}_{NW}(x)$ □

$$\left(\frac{\hat{N}(x)}{\hat{f}(x)}\right) \qquad \left(\frac{\tilde{N}(x)}{\hat{f}(x)}\right)$$

$$\hat{m}(x) = \frac{\sum Y_i \, K\left(\frac{x - X_i}{h}\right)}{\sum K\left(\frac{x - X_i}{h}\right)}$$