

Pratique de l'apprentissage statistique

3. Noyaux de lissage & Plus proches voisins

V. Lefieux



École des Ponts
ParisTech

Plan

Méthode du noyau de lissage : fonction de densité

Méthode du noyau de lissage : régression

Plus proches voisins

Plan

平滑核

Méthode du noyau de lissage : fonction de densité

Méthode du noyau de lissage : régression

Plus proches voisins

Objectif

随机变量

Estimer la **fonction de densité f** d'une variable aléatoire $X \in \mathbb{R}$:

- ▶ à partir d'un échantillon i.i.d (X_1, \dots, X_n) ,
- ▶ sans hypothèse paramétrique.

Un point de départ : l'histogramme I

- ▶ Soit (X_1, \dots, X_n) un échantillon i.i.d de $X \in \mathbb{R}$ qui admet une fonction de densité f de support borné $([a, b])$.

Un point de départ : l'histogramme I

- ▶ Soit (X_1, \dots, X_n) un échantillon i.i.d de $X \in \mathbb{R}$ qui admet une fonction de densité f de support borné $([a, b])$.
- ▶ On découpe le support en k petits intervalles :

$$([\alpha_i, \alpha_{i+1}])_{i \in \{1, \dots, k\}}$$

avec $\alpha_1 = a$ and $\alpha_{k+1} = b$ (en toute rigueur les intervalles sont ouverts à droite).

Un point de départ : l'histogramme |

- ▶ Soit (X_1, \dots, X_n) un échantillon i.i.d de $X \in \mathbb{R}$ qui admet une fonction de densité f de support borné $([a, b])$.
允许有界支持的密度函数 f
- ▶ On découpe le support en k petits intervalles :

$$([\alpha_i, \alpha_{i+1}])_{i \in \{1, \dots, k\}}$$

avec $\alpha_1 = a$ and $\alpha_{k+1} = b$ (en toute rigueur les intervalles sont ouverts à droite). 严格来说，右边开区间

- ▶ L'estimateur **histogramme** est une fonction en escaliers :

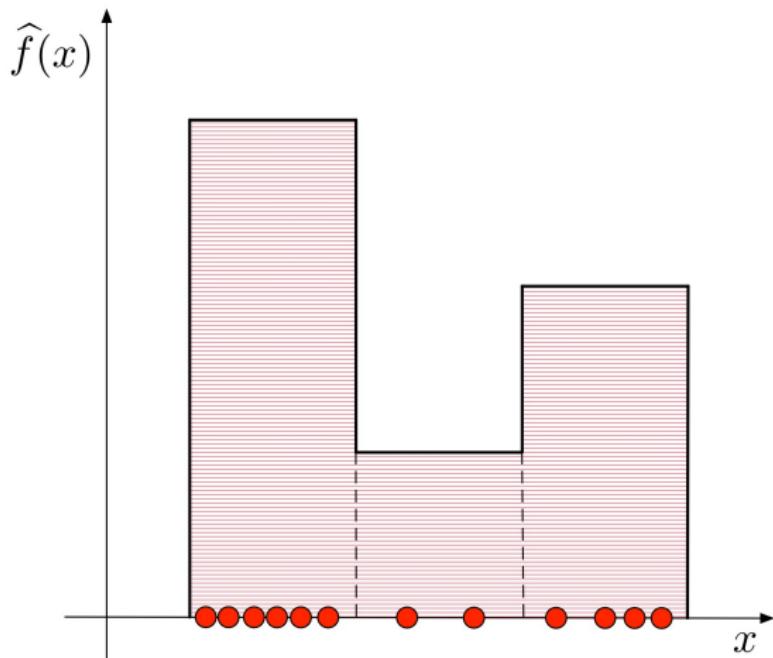
直方图估计器是一个阶跃函数

$$\forall x \in [a, b] : \hat{f}(x) = \sum_{i=1}^k \frac{f_i}{\alpha_{i+1} - \alpha_i} \mathbf{1}_{[\alpha_i, \alpha_{i+1}]}(x)$$

où :

$$f_i = \frac{1}{n} \sum_{j=1}^n \mathbf{1}_{[\alpha_i, \alpha_{i+1}]}(X_j).$$

Un point de départ : l'histogramme II



De l'histogramme à l'estimateur de la fenêtre mobile

- ▶ Soit un point $x \in [\alpha_i, \alpha_{i+1}]$ tout près de α_i .

De l'histogramme à l'estimateur de la fenêtre mobile

- ▶ Soit un point $x \in [\alpha_i, \alpha_{i+1}]$ tout près de α_i .
- ▶ Un point de l'échantillon situé un peu avant α_{i+1} sera pris en compte alors qu'un point de l'échantillon situé un peu avant α_i ne le sera pas.

De l'histogramme à l'estimateur de la fenêtre mobile

从直方图到移动窗口估计器

非常接近 α_i 。

- ▶ Soit un point $x \in [\alpha_i, \alpha_{i+1}]$ tout près de α_i .
- ▶ Un point de l'échantillon situé un peu avant α_{i+1} sera pris en compte alors qu'un point de l'échantillon situé un peu avant α_i ne le sera pas. 将考虑位于 $i+1$ 之前的样本点，而不会考虑位于 i 之前的样本点。
- ▶ Pour pallier ce problème, on peut utiliser l'estimateur de la fenêtre mobile qui est un translaté de l'histogramme de manière à ce que le point x d'estimation se retrouve au centre de la classe. 为了克服这个问题，我们可以使用移动窗口估计器，它是直方图的平移，使得估计点 x 位于类的中心。

Estimateur de la fenêtre mobile I

- ▶ L'estimateur de la fenêtre mobile est un « histogramme mobile ».

Estimateur de la fenêtre mobile I

- ▶ L'estimateur de la fenêtre mobile est un « histogramme mobile ».
- ▶ On considère un intervalle autour de x :

$$[x - h_n, x + h_n]$$

où h_n est la fenêtre (soit la demi-largeur de l'intervalle).

Estimateur de la fenêtre mobile |

移动窗口估计器是“移动直方图”。

- ▶ L'estimateur de la fenêtre mobile est un « histogramme mobile ».

我考虑一个围绕 x 的区间：

- ▶ On considère un intervalle autour de x :

$$[x - h_n, x + h_n]$$

其中 h_n 是窗口 (即间隔的半宽度)。

où h_n est la fenêtre (soit la demi-largeur de l'intervalle).

- ▶ La valeur de l'estimateur au point x est égale à la hauteur du rectangle dans l'histogramme pour l'intervalle considéré :

点 x 处的估计值等于所考虑区间的直方图中矩形的高度

$$\hat{f}(x) = \frac{1}{2n h_n} \sum_{i=1}^n \mathbb{1}_{[x-h_n, x+h_n]}(X_i).$$

Estimateur de la fenêtre mobile II

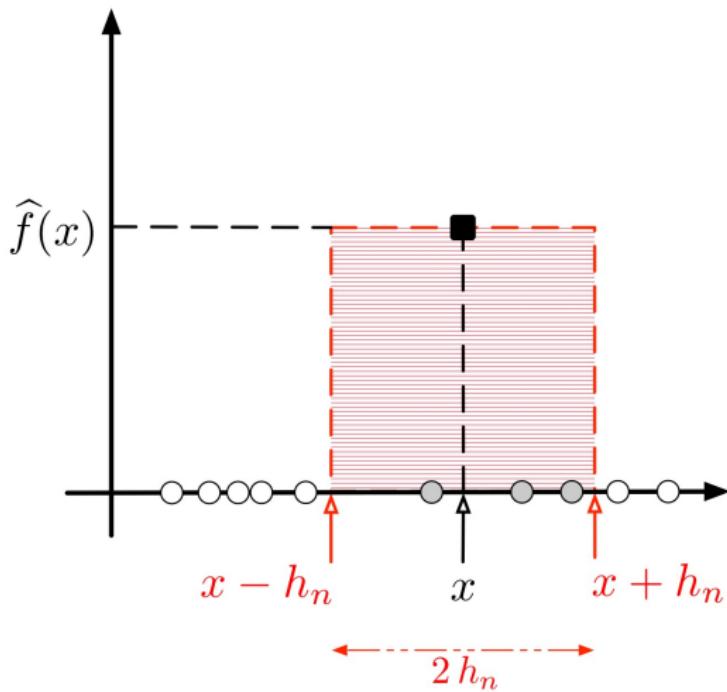
On peut réécrire l'estimateur de la fenêtre mobile sous la forme :

$$\hat{f}(x) = \frac{1}{n h_n} \sum_{i=1}^n K\left(\frac{x - X_i}{h_n}\right)$$

où :

$$K(x) = \frac{1}{2} \mathbb{1}_{[-1,1]}(x).$$

Estimateur de la fenêtre mobile III



De l'estimateur de la fenêtre mobile à l'estimateur à noyau

- ▶ Tous les points de l'échantillon dans l'intervalle centré en x ont la même importance pour le calcul de $\hat{f}(x)$.

De l'estimateur de la fenêtre mobile à l'estimateur à noyau

- ▶ Tous les points de l'échantillon dans l'intervalle centré en x ont la même importance pour le calcul de $\hat{f}(x)$.
- ▶ Une idée naturelle est de pondérer les observations en mettant d'autant plus de poids qu'elles sont proches de x .

De l'estimateur de la fenêtre mobile à l'estimateur à noyau

从移动窗口估计器到核估计器

- ▶ Tous les points de l'échantillon dans l'intervalle centré en x ont la même importance pour le calcul de $\hat{f}(x)$.
以 x 为中心的区间中的所有样本点对于计算 $\hat{f}(x)$ 具有相同的重要性。
- ▶ Une idée naturelle est de pondérer les observations en mettant d'autant plus de poids qu'elles sont proches de x .
一个自然的想法是通过在接近 x 时增加更多的权重来加权观察。
- ▶ On considère pour cela une fonction de poids, un noyau, moins brutal que la loi uniforme qui equipondère les observations.
为此，我们考虑了一个权重函数，一个内核，它没有均衡观察的统一法则那么残酷。

Estimateur à noyau

L'estimateur à noyau de la fonction de densité (KDE : *Kernel Density Estimator*), dans le cas univarié, s'écrit :

$$\hat{f}(x) = \frac{1}{n h_n} \sum_{i=1}^n K\left(\frac{x - X_i}{h_n}\right)$$

où :

- ▶ h_n est la fenêtre,
- ▶ K est un noyau de Parzen-Rosenblatt.

Noyau

- ▶ Un **noyau** de Parzen-Rosenblatt vérifie :
 - ▶ K est borné.
 - ▶ $\int_{-\infty}^{+\infty} K(x) dx = 1$.
 - ▶ $x K(x) \xrightarrow{x \rightarrow +\infty} 0$.

Noyau

- ▶ Un **noyau** de Parzen-Rosenblatt vérifie :
 - ▶ K est borné.
 - ▶ $\int_{-\infty}^{+\infty} K(x) dx = 1$.
 - ▶ $x K(x) \xrightarrow{x \rightarrow +\infty} 0$.
- ▶ On considère généralement des noyaux symétriques :

$$\forall x \in \mathbb{R} : K(-x) = K(x) .$$

Noyau

- ▶ Un **noyau** de Parzen-Rosenblatt vérifie :
 - ▶ K est borné.
 - ▶ $\int_{-\infty}^{+\infty} K(x) dx = 1$.
 - ▶ $x K(x) \xrightarrow{x \rightarrow +\infty} 0$.
- ▶ On considère généralement des noyaux symétriques :

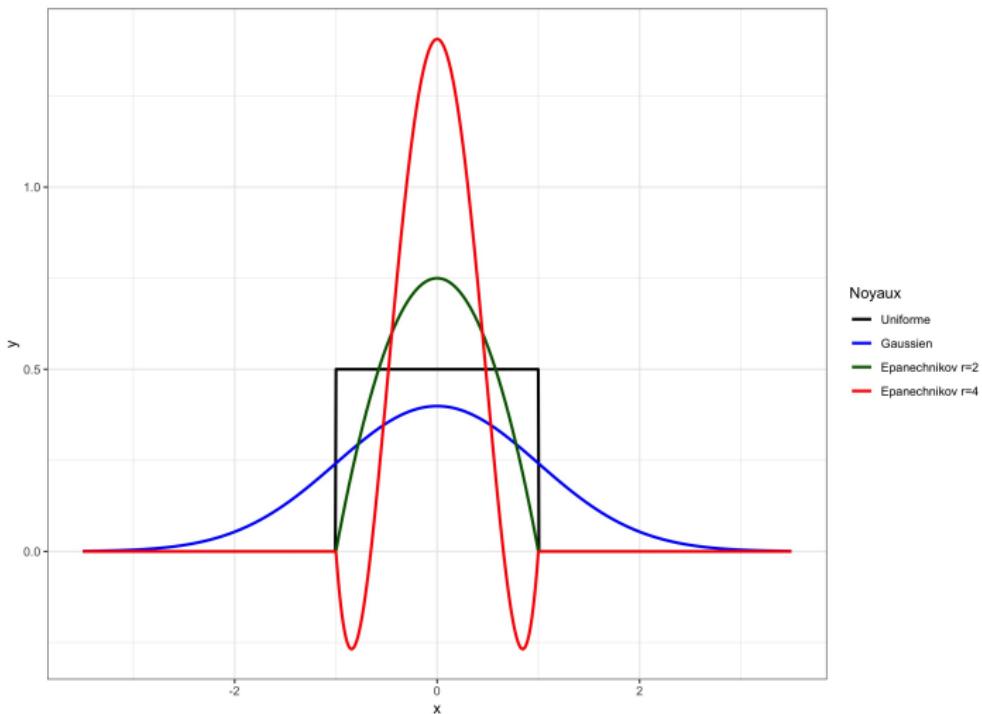
$$\forall x \in \mathbb{R} : K(-x) = K(x) .$$

- ▶ Le noyau n'est pas forcément positif.

Noyaux classiques I

Uniforme	$K(x) = \frac{1}{2} \mathbb{1}_{[-1,1]}(x)$
Gaussien	$K(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x^2\right)$
Epanechnikov (r=2)	$K(x) = \frac{3}{4}(1-x^2)\mathbb{1}_{[-1,1]}(x)$
Epanechnikov (r=4)	$K(x) = \frac{15}{8}\frac{3}{4}(1-\frac{7}{3}x^2)(1-x^2)\mathbb{1}_{[-1,1]}(x)$

Noyaux classiques II



Biais et variance de l'estimateur

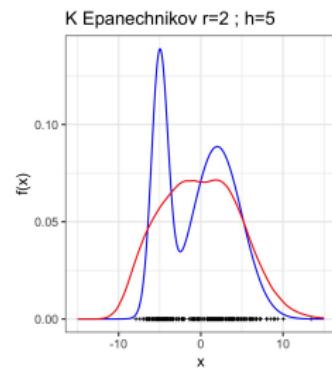
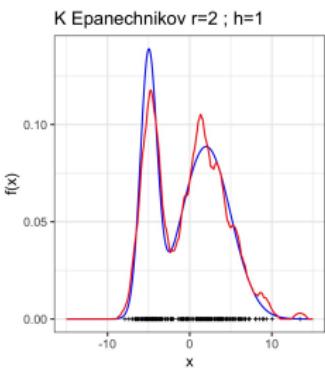
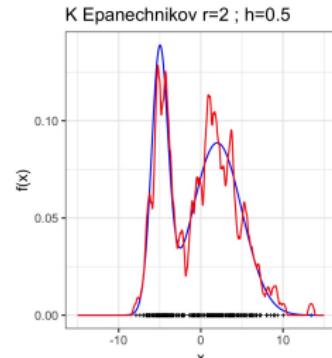
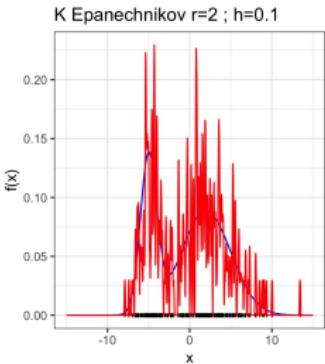
- ▶ Le biais croît avec fenêtre.

Biais et variance de l'estimateur

- ▶ Le biais croît avec fenêtre.
- ▶ La variance tend vers 0 si :

$$n h_n \xrightarrow{n \rightarrow +\infty} +\infty .$$

Illustration



Fenêtre optimale : validation croisée I

- ▶ Considérons l'**erreur quadratique intégrée** (**ISE** : *Integrated Squared Error*) :

$$\text{ISE}(\hat{f}) = \int_{-\infty}^{+\infty} (\hat{f}(x) - f(x))^2 dx .$$

Fenêtre optimale : validation croisée I

- ▶ Considérons l'**erreur quadratique intégrée** (**ISE** : *Integrated Squared Error*) :

$$\text{ISE}(\hat{f}) = \int_{-\infty}^{+\infty} (\hat{f}(x) - f(x))^2 dx .$$

- ▶ On peut réécrire :

$$\begin{aligned}\text{ISE}(\hat{f}) &= \int_{-\infty}^{+\infty} (\hat{f}(x) - f(x))^2 dx \\ &= \int_{-\infty}^{+\infty} \hat{f}^2(x) dx - 2 \int_{-\infty}^{+\infty} \hat{f}(x) f(x) dx + \int_{-\infty}^{+\infty} f^2(x) dx .\end{aligned}$$

Fenêtre optimale : validation croisée II

- ▶ $\int_{-\infty}^{+\infty} f^2(x) dx$ ne dépend pas de h_n .

Fenêtre optimale : validation croisée II

- ▶ $\int_{-\infty}^{+\infty} f^2(x) dx$ ne dépend pas de h_n .
- ▶ Pour une fenêtre h_n donnée, $\int_{-\infty}^{+\infty} \widehat{f}^2(x) dx$ peut être calculé.

Fenêtre optimale : validation croisée II

- ▶ $\int_{-\infty}^{+\infty} f^2(x) dx$ ne dépend pas de h_n .
- ▶ Pour une fenêtre h_n donnée, $\int_{-\infty}^{+\infty} \hat{f}^2(x) dx$ peut être calculé.
- ▶ On peut considérer que $\int \hat{f}(x) f(x) dx$ est équivalent à $\mathbb{E}[\hat{f}(X)]$, abusivement car \hat{f} dépend de l'échantillon.

Fenêtre optimale : validation croisée II

- ▶ $\int_{-\infty}^{+\infty} f^2(x) dx$ ne dépend pas de h_n .
- ▶ Pour une fenêtre h_n donnée, $\int_{-\infty}^{+\infty} \hat{f}^2(x) dx$ peut être calculé.
- ▶ On peut considérer que $\int \hat{f}(x) f(x) dx$ est équivalent à $\mathbb{E}[\hat{f}(X)]$, abusivement car \hat{f} dépend de l'échantillon.
- ▶ Afin de corriger cette dépendance, on utilise l'estimateur **leave-one out** pour estimer $\mathbb{E}[\hat{f}(X)]$:

$$\frac{1}{n} \sum_{i=1}^n \hat{f}_{-i}(X_i)$$

où :

$$\hat{f}_{-i}(x) = \frac{1}{(n-1)h} \sum_{j=1, j \neq i}^n K\left(\frac{x - X_j}{h_n}\right).$$

Fenêtre optimale : validation croisée III

La fenêtre obtenue par validation croisée minimise :

$$CV(h_n) = \int \widehat{f}^2(x) dx - \frac{2}{n(n-1)h_n} \sum_{i=1}^n \sum_{j=1, j \neq i}^n K\left(\frac{X_i - X_j}{h_n}\right).$$

Illustration |

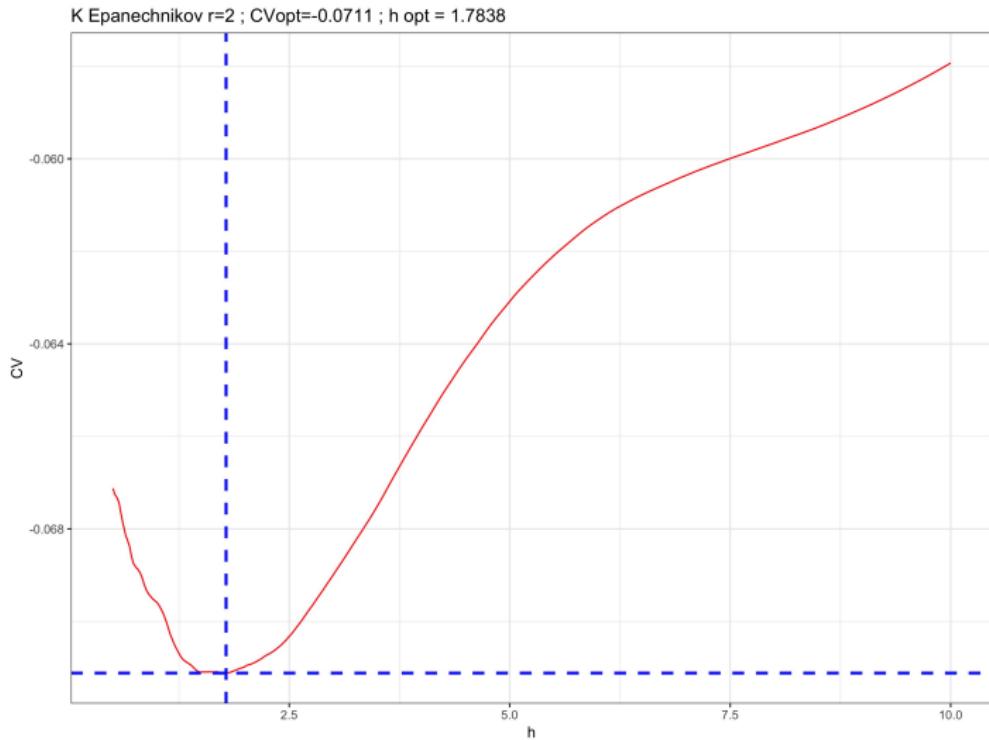
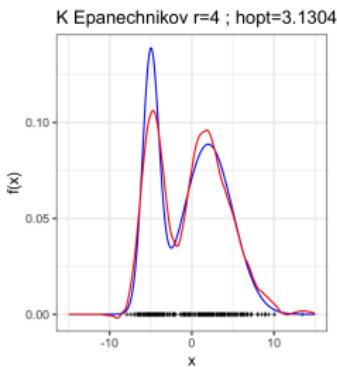
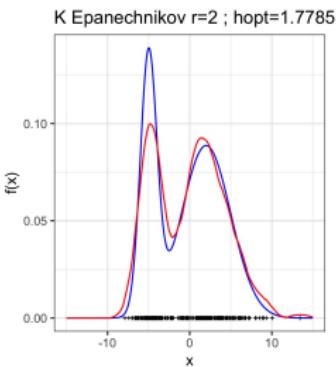
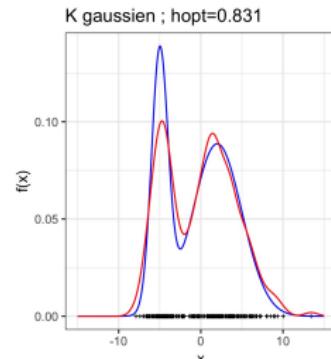
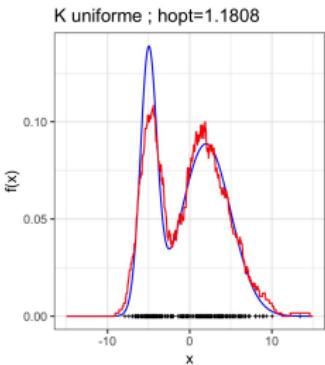


Illustration II



Noyau optimal

- ▶ Epanechnikov a mené des travaux pour déterminer le noyau optimal et a obtenu une version homothétique du noyau K suivant :

$$K(x) = \frac{3}{4} (1 - x^2) \mathbb{1}_{[-1,1]}(x) .$$

Noyau optimal

- ▶ Epanechnikov a mené des travaux pour déterminer le noyau optimal et a obtenu une version homothétique du noyau K suivant :

$$K(x) = \frac{3}{4} (1 - x^2) \mathbb{1}_{[-1,1]}(x) .$$

- ▶ Le choix du noyau n'est pas fondamental, contrairement à celui de la fenêtre.

Objectif

Estimer la **fonction de densité f** d'une variable aléatoire $X \in \mathbb{R}^p$:

- ▶ à partir d'un échantillon i.i.d (X_1, \dots, X_n) ,
- ▶ sans hypothèse paramétrique.

Estimateur à noyau

- Soit (X_1, \dots, X_n) un échantillon i.i.d de $X \in \mathbb{R}^p$, $p \in \mathbb{N}^*$.

Estimateur à noyau

- ▶ Soit (X_1, \dots, X_n) un échantillon i.i.d de $X \in \mathbb{R}^p$, $p \in \mathbb{N}^*$.
- ▶ L'estimateur à noyau de la fonction de densité (dans le cas multivarié) est défini par :

$$\forall x = (x_1, \dots, x_p)^\top \in \mathbb{R}^p : \hat{f}(x) = \frac{1}{n h_{n1} \dots h_{np}} \sum_{i=1}^n K_p \left(\frac{x_1 - X_{i1}}{h_{n1}}, \dots, \frac{x_p - X_{ip}}{h_{np}} \right)$$

Estimateur à noyau

- ▶ Soit (X_1, \dots, X_n) un échantillon i.i.d de $X \in \mathbb{R}^p$, $p \in \mathbb{N}^*$.
- ▶ L'estimateur à noyau de la fonction de densité (dans le cas multivarié) est défini par :

$$\forall x = (x_1, \dots, x_p)^\top \in \mathbb{R}^p : \hat{f}(x) = \frac{1}{n h_{n1} \dots h_{np}} \sum_{i=1}^n K_p \left(\frac{x_1 - X_{i1}}{h_{n1}}, \dots, \frac{x_p - X_{ip}}{h_{np}} \right)$$

où :

- ▶ $H = (h_{n1}, \dots, h_{np})^\top \in \mathbb{R}^p$ est la fenêtre,

Estimateur à noyau

- ▶ Soit (X_1, \dots, X_n) un échantillon i.i.d de $X \in \mathbb{R}^p$, $p \in \mathbb{N}^*$.
- ▶ L'estimateur à noyau de la fonction de densité (dans le cas multivarié) est défini par :

$$\forall x = (x_1, \dots, x_p)^\top \in \mathbb{R}^p : \hat{f}(x) = \frac{1}{n h_{n1} \dots h_{np}} \sum_{i=1}^n K_p \left(\frac{x_1 - X_{i1}}{h_{n1}}, \dots, \frac{x_p - X_{ip}}{h_{np}} \right)$$

où :

- ▶ $H = (h_{n1}, \dots, h_{np})^\top \in \mathbb{R}^p$ est la fenêtre,
- ▶ K_p un noyau multivarié de Parzen-Rosenblatt :
 - ▶ K_p est bornée sur \mathbb{R}^p ,
 - ▶ $\int_{\mathbb{R}^p} K_p(x) dx = 1$,
 - ▶ $\|x\|^p K_p(x) \xrightarrow{\|x\| \rightarrow +\infty} 0$ où $\|\cdot\|$ désigne la norme euclidienne.

Fenêtre et noyau

- ▶ Usuellement on choisit la même fenêtre
 $H = (h_n, \dots, h_n) \in \mathbb{R}^p$:

$$\widehat{f}(x) = \frac{1}{n(h_n)^p} \sum_{i=1}^n K_p\left(\frac{x - X_i}{h_n}\right)$$

où $X_i = (X_{i1}, \dots, X_{ip})^\top$.

Fenêtre et noyau

- ▶ Usuellement on choisit la même fenêtre
 $H = (h_n, \dots, h_n) \in \mathbb{R}^p$:

$$\widehat{f}(x) = \frac{1}{n(h_n)^p} \sum_{i=1}^n K_p \left(\frac{x - X_i}{h_n} \right)$$

où $X_i = (X_{i1}, \dots, X_{ip})^\top$.

- ▶ Usuellement on choisit un noyau produit :

$$K_p(x_1, \dots, x_p) = \prod_{j=1}^p K(x_j).$$

Biais et variance

- ▶ Le biais croît avec la fenêtre.

Biais et variance

- ▶ Le biais croît avec la fenêtre.
- ▶ La variance tend vers 0 si :

$$n(h_n)^p \xrightarrow{n \rightarrow +\infty} +\infty .$$

Choix de la fenêtre

On sélectionne la fenêtre par validation croisée.

Fléau de la dimension

- ▶ La convergence de la méthode décroît avec la dimension, c'est ce qu'on appelle communément le **fléau de la dimension**.

Fléau de la dimension

- ▶ La convergence de la méthode décroît avec la dimension, c'est ce qu'on appelle communément le **fléau de la dimension**.
- ▶ En cas de dimension élevée, il faut disposer d'un très grand échantillon pour que la méthode soit efficace (ce qui est le cas pour toutes les méthodes de moyennage locale).

Plan

Méthode du noyau de lissage : fonction de densité

Méthode du noyau de lissage : régression

Plus proches voisins

Cas de la régression simple

- Soit $(X_i, Y_i)_{i \in \{1, \dots, n\}}$ un échantillon i.i.d de $(X, Y) \in \mathbb{R} \times \mathbb{R}$.

Cas de la régression simple

- ▶ Soit $(X_i, Y_i)_{i \in \{1, \dots, n\}}$ un échantillon i.i.d de $(X, Y) \in \mathbb{R} \times \mathbb{R}$.
- ▶ Soit le modèle de régression suivant :

$$Y = m(X) + \varepsilon .$$

Cas de la régression simple

- ▶ Soit $(X_i, Y_i)_{i \in \{1, \dots, n\}}$ un échantillon i.i.d de $(X, Y) \in \mathbb{R} \times \mathbb{R}$.
- ▶ Soit le modèle de régression suivant :

$$Y = m(X) + \varepsilon .$$

- ▶ L'objectif est d'estimer la **fonction de lien m** sans hypothèse paramétrique.

Vers l'estimateur de Nadaraya-Watson I

La fonction de lien :

$$\begin{aligned}m(x) &= \mathbb{E}(Y/X=x) \\&= \int_{-\infty}^{+\infty} y f_{Y/X=x}(y) dy \\&= \frac{1}{f_X(x)} \int_{-\infty}^{+\infty} y f_{X,Y}(x,y) dy\end{aligned}$$

minimise le risque quadratique.

Vers l'estimateur de Nadaraya-Watson II

- ▶ A partir des deux estimateurs de densité suivants (noyau produit pour X et Y , noyau K symétrique, même fenêtre h_n) :

Vers l'estimateur de Nadaraya-Watson II

- ▶ A partir des deux estimateurs de densité suivants (noyau produit pour X et Y , noyau K symétrique, même fenêtre h_n) :

$$\hat{f}_X(x) = \frac{1}{n h_n} \sum_{i=1}^n K\left(\frac{x - X_i}{h_n}\right),$$

$$\hat{f}_{X,Y}(x, y) = \frac{1}{n(h_n)^2} \sum_{i=1}^n K\left(\frac{x - X_i}{h_n}\right) K\left(\frac{y - Y_i}{h_n}\right),$$

Vers l'estimateur de Nadaraya-Watson II

- ▶ A partir des deux estimateurs de densité suivants (noyau produit pour X et Y , noyau K symétrique, même fenêtre h_n) :

$$\hat{f}_X(x) = \frac{1}{n h_n} \sum_{i=1}^n K\left(\frac{x - X_i}{h_n}\right),$$

$$\hat{f}_{X,Y}(x, y) = \frac{1}{n(h_n)^2} \sum_{i=1}^n K\left(\frac{x - X_i}{h_n}\right) K\left(\frac{y - Y_i}{h_n}\right),$$

on peut considérer l'estimateur suivant :

$$\hat{m}_{nw}(x) = \frac{1}{\hat{f}_X(x)} \int_{-\infty}^{+\infty} y \hat{f}_{X,Y}(x, y) dy.$$

Vers l'estimateur de Nadaraya-Watson II

- ▶ A partir des deux estimateurs de densité suivants (noyau produit pour X et Y , noyau K symétrique, même fenêtre h_n) :

$$\hat{f}_X(x) = \frac{1}{n h_n} \sum_{i=1}^n K\left(\frac{x - X_i}{h_n}\right),$$

$$\hat{f}_{X,Y}(x, y) = \frac{1}{n(h_n)^2} \sum_{i=1}^n K\left(\frac{x - X_i}{h_n}\right) K\left(\frac{y - Y_i}{h_n}\right),$$

on peut considérer l'estimateur suivant :

$$\hat{m}_{nw}(x) = \frac{1}{\hat{f}_X(x)} \int_{-\infty}^{+\infty} y \hat{f}_{X,Y}(x, y) dy.$$

- ▶ On obtient après calculs :

$$\hat{m}_{nw}(x) = \frac{\sum_{i=1}^n Y_i K\left(\frac{x - X_i}{h_n}\right)}{\sum_{i=1}^n K\left(\frac{x - X_i}{h_n}\right)}.$$

Cas de la régression multiple

- Soit $(X_i, Y_i)_{i \in \{1, \dots, n\}}$ un échantillon i.i.d de $(X, Y) \in \mathbb{R}^p \times \mathbb{R}$.

Cas de la régression multiple

- ▶ Soit $(X_i, Y_i)_{i \in \{1, \dots, n\}}$ un échantillon i.i.d de $(X, Y) \in \mathbb{R}^p \times \mathbb{R}$.
- ▶ Soit le modèle de régression suivant :

$$Y = m(X) + \varepsilon .$$

Cas de la régression multiple

- ▶ Soit $(X_i, Y_i)_{i \in \{1, \dots, n\}}$ un échantillon i.i.d de $(X, Y) \in \mathbb{R}^p \times \mathbb{R}$.
- ▶ Soit le modèle de régression suivant :

$$Y = m(X) + \varepsilon .$$

- ▶ L'objectif est d'estimer la **fonction de lien m** sans hypothèse paramétrique.

Estimateur de Nadaraya-Watson

- ▶ L'estimateur de Nadaraya-Watson vaut (avec la même fenêtre pour toutes les dimensions) :

$$\hat{m}_{nw}(x) = \frac{\sum_{i=1}^n Y_i K_p\left(\frac{x-X_i}{h_n}\right)}{\sum_{i=1}^n K_p\left(\frac{x-X_i}{h_n}\right)}.$$

Estimateur de Nadaraya-Watson

- ▶ L'estimateur de Nadaraya-Watson vaut (avec la même fenêtre pour toutes les dimensions) :

$$\hat{m}_{nw}(x) = \frac{\sum_{i=1}^n Y_i K_p\left(\frac{x-X_i}{h_n}\right)}{\sum_{i=1}^n K_p\left(\frac{x-X_i}{h_n}\right)}.$$

- ▶ Usuellement on choisit un noyau produit.

Estimateur de Nadaraya-Watson

- ▶ L'estimateur de Nadaraya-Watson vaut (avec la même fenêtre pour toutes les dimensions) :

$$\hat{m}_{nw}(x) = \frac{\sum_{i=1}^n Y_i K_p\left(\frac{x-X_i}{h_n}\right)}{\sum_{i=1}^n K_p\left(\frac{x-X_i}{h_n}\right)}.$$

- ▶ Usuellement on choisit un noyau produit.
- ▶ On peut l'écrire sous la forme :

$$\hat{m}_{nw}(x) = \sum_{i=1}^n \omega_i Y_i$$

où :

$$\omega_i = \frac{K_p\left(\frac{x-X_i}{h_n}\right)}{\sum_{j=1}^n K_p\left(\frac{x-X_j}{h_n}\right)}.$$

Biais et variance de l'estimateur

- ▶ Le biais croît avec la fenêtre.

Biais et variance de l'estimateur

- ▶ Le biais croît avec la fenêtre.
- ▶ La variance tend vers 0 si :

$$n(h_n)^p \xrightarrow{n \rightarrow +\infty} +\infty .$$

Illustration |

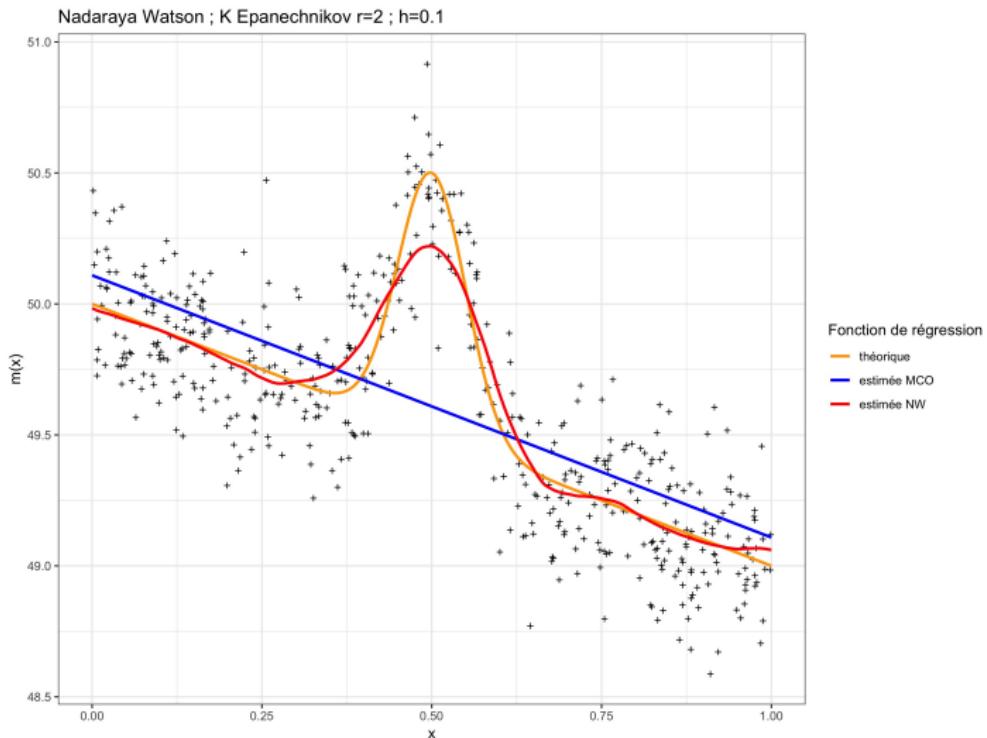
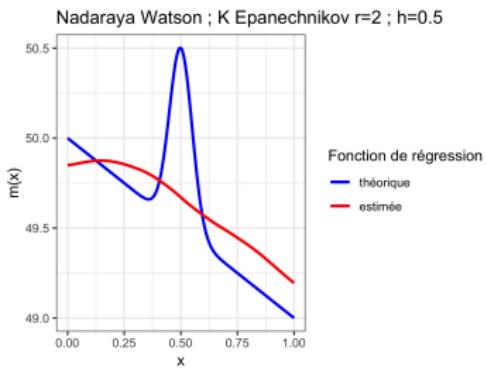
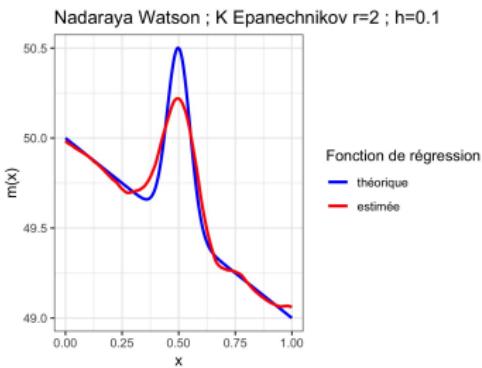
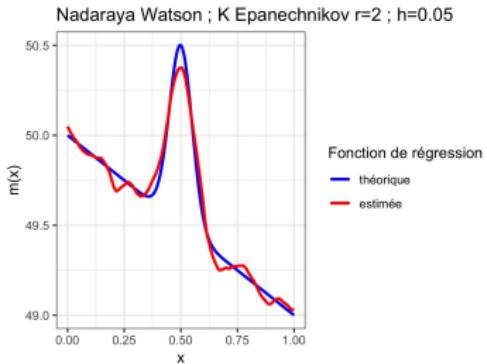
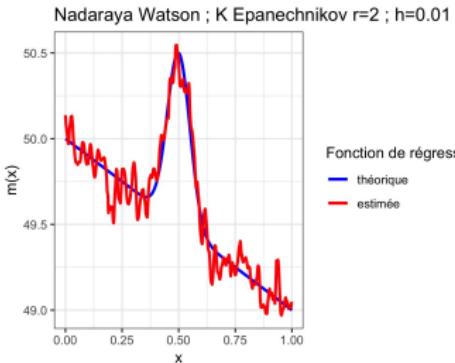


Illustration II



Validation croisée

- ▶ On peut choisir la fenêtre optimale par validation croisée.

Validation croisée

- ▶ On peut choisir la fenêtre optimale par validation croisée.
- ▶ On peut estimer le MISE :

$$\text{MISE}(m) = \mathbb{E} \left[(\hat{m}_{nw}(X) - m(X))^2 \right]$$

par :

$$\frac{1}{n} \sum_{i=1}^n [\hat{m}_{nw}(X_i) - m(X_i)]^2 .$$

Validation croisée

- ▶ On peut choisir la fenêtre optimale par validation croisée.
- ▶ On peut estimer le MISE :

$$\text{MISE}(m) = \mathbb{E} \left[(\hat{m}_{nw}(X) - m(X))^2 \right]$$

par :

$$\frac{1}{n} \sum_{i=1}^n [\hat{m}_{nw}(X_i) - m(X_i)]^2 .$$

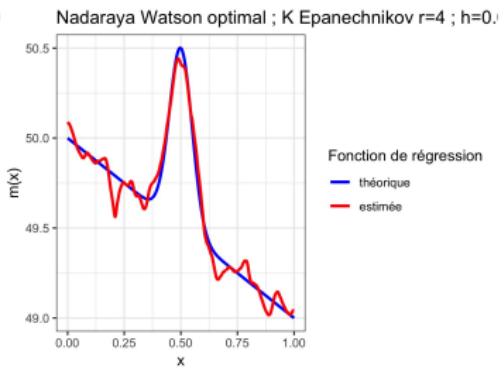
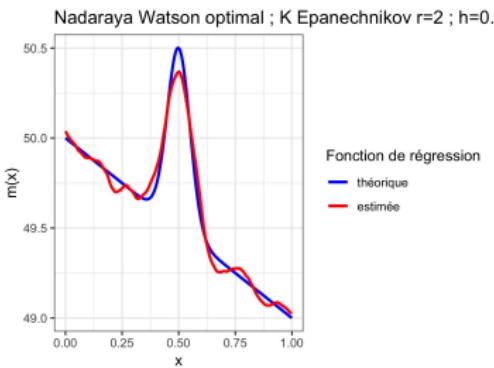
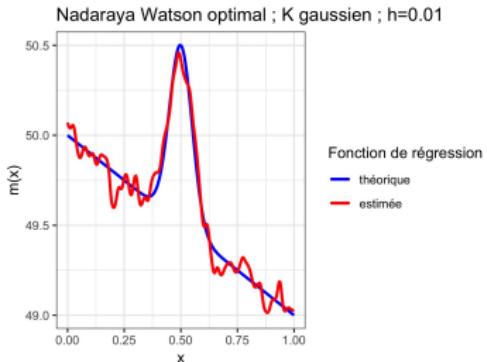
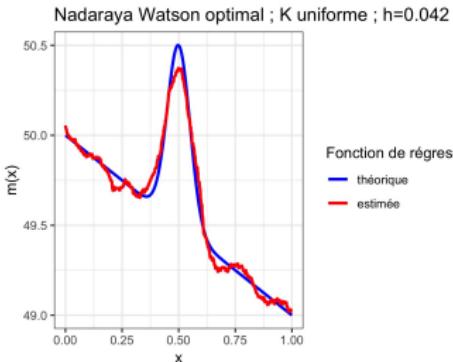
- ▶ On obtient la fenêtre par validation croisée en minimisant :

$$CV(h_n) = \frac{1}{n} \sum_{i=1}^n [\hat{m}_{nw,-i}(X_i) - Y_i]^2$$

où :

$$\hat{m}_{nw,-i}(x) = \frac{\sum_{j=1, j \neq i}^n Y_j K\left(\frac{x-X_j}{h}\right)}{\sum_{j=1, j \neq i}^n K\left(\frac{x-X_j}{h}\right)} .$$

Illustration



Plan

Méthode du noyau de lissage : fonction de densité

Méthode du noyau de lissage : régression

Plus proches voisins

Données considérées

- On dispose d'un échantillon de (X, Y) :

$$\mathcal{D}_n = (X_i, Y_i)_{i \in \{1, \dots, n\}} .$$

On note :

$$d_n = (x_i, y_i)_{i \in \{1, \dots, n\}} .$$

Données considérées

- ▶ On dispose d'un échantillon de (X, Y) :

$$\mathcal{D}_n = (X_i, Y_i)_{i \in \{1, \dots, n\}} .$$

On note :

$$d_n = (x_i, y_i)_{i \in \{1, \dots, n\}} .$$

- ▶ On considère dans la suite que :
 - ▶ $X \in \mathbb{R}^p$.
Toutes les covariables sont considérés quantitatives.
 - ▶ $Y \in \{1, \dots, K\}$ dans le cas de la classification supervisée.
 - ▶ $Y \in \mathbb{R}$ dans le cas de la régression.

Plus proches voisins

- ▶ On calcule les distances entre $x \in \mathbb{R}^p$ et l'échantillon :

$$\forall i \in \{1, \dots, n\} : D_i = d(x, X_i)$$

où d est une distance (ex : euclidienne, Manhattan).

Plus proches voisins

- ▶ On calcule les distances entre $x \in \mathbb{R}^p$ et l'échantillon :

$$\forall i \in \{1, \dots, n\} : D_i = d(x, X_i)$$

où d est une distance (ex : euclidienne, Manhattan).

- ▶ On réordonne alors ces distances :

$$D_{(1)} \leq \dots \leq D_{(n)} .$$

Plus proches voisins

- ▶ On calcule les distances entre $x \in \mathbb{R}^p$ et l'échantillon :

$$\forall i \in \{1, \dots, n\} : D_i = d(x, X_i)$$

où d est une distance (ex : euclidienne, Manhattan).

- ▶ On réordonne alors ces distances :

$$D_{(1)} \leq \dots \leq D_{(n)} .$$

- ▶ Les **k -plus proches voisins** de x (**k -pp**) sont les k points associés à $D_{(1)} \leq \dots \leq D_{(k)}$:

$$X_{(1)}(x), \dots, X_{(k)}(x) .$$

Plus proches voisins

- ▶ On calcule les distances entre $x \in \mathbb{R}^p$ et l'échantillon :

$$\forall i \in \{1, \dots, n\} : D_i = d(x, X_i)$$

où d est une distance (ex : euclidienne, Manhattan).

- ▶ On réordonne alors ces distances :

$$D_{(1)} \leq \dots \leq D_{(n)} .$$

- ▶ Les **k -plus proches voisins** de x (**k -pp**) sont les k points associés à $D_{(1)} \leq \dots \leq D_{(k)}$:

$$X_{(1)}(x), \dots, X_{(k)}(x) .$$

- ▶ En cas d'égalité, il est possible d'utiliser un tirage au sort.

Enjeux

A partir de ces k -plus proches voisins, on peut :

- ▶ Estimer une **loi de probabilité**.
- ▶ Effectuer une **classification supervisée**.
- ▶ Effectuer une **régression**.

Classification supervisée

- ▶ La règle consiste à déterminer la modalité la plus représentée parmi les k -plus proches voisins.

Classification supervisée

- ▶ La règle consiste à déterminer la modalité la plus représentée parmi les k -plus proches voisins.
- ▶ Il s'agit d'un **vote majoritaire**.

Illustration |

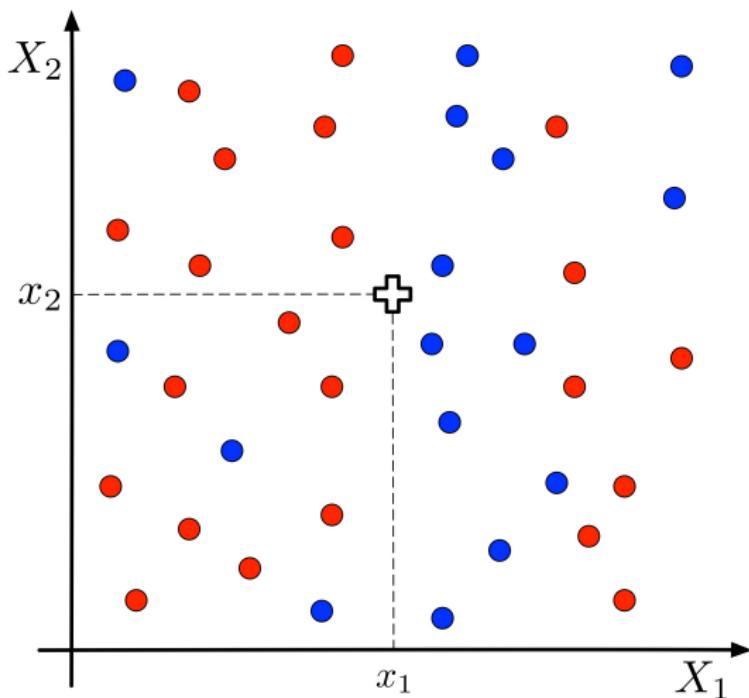


Illustration II

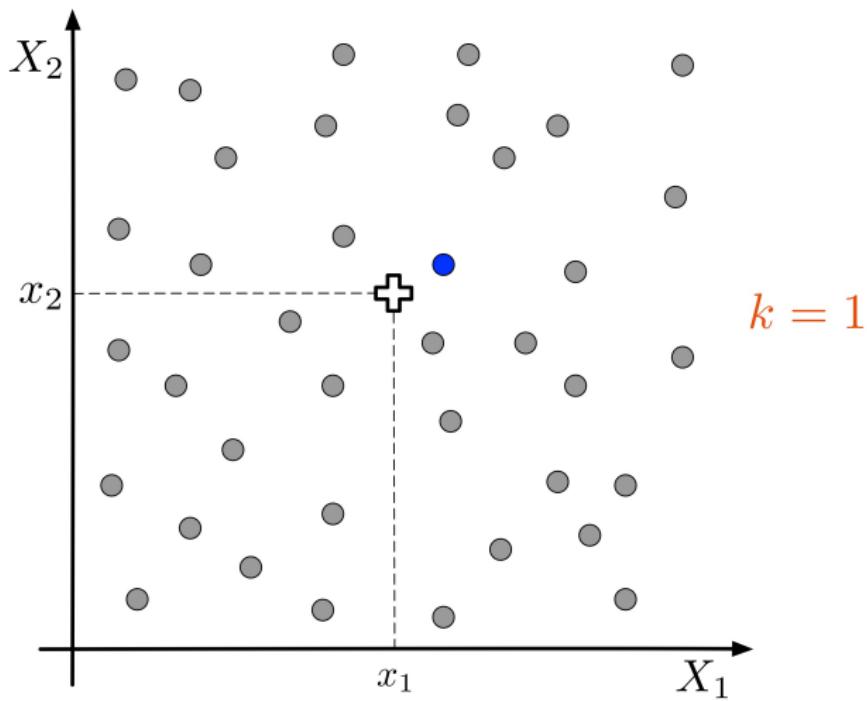


Illustration III

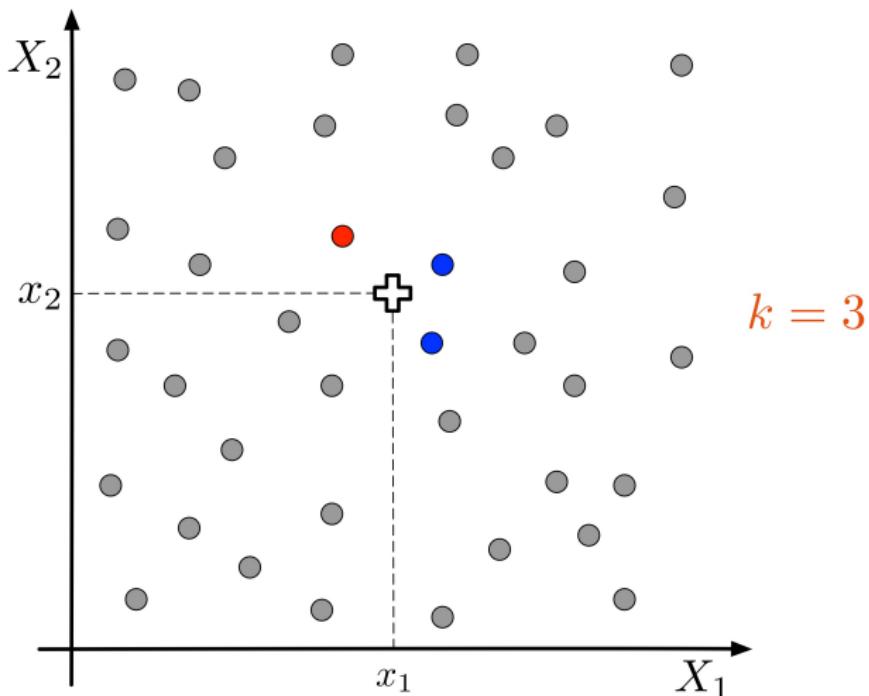


Illustration IV

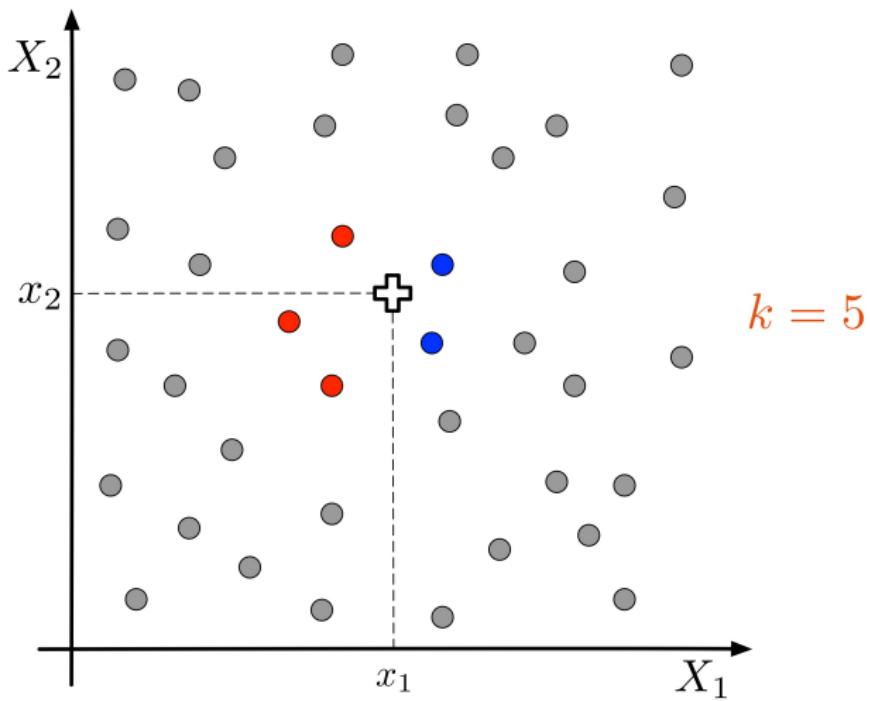
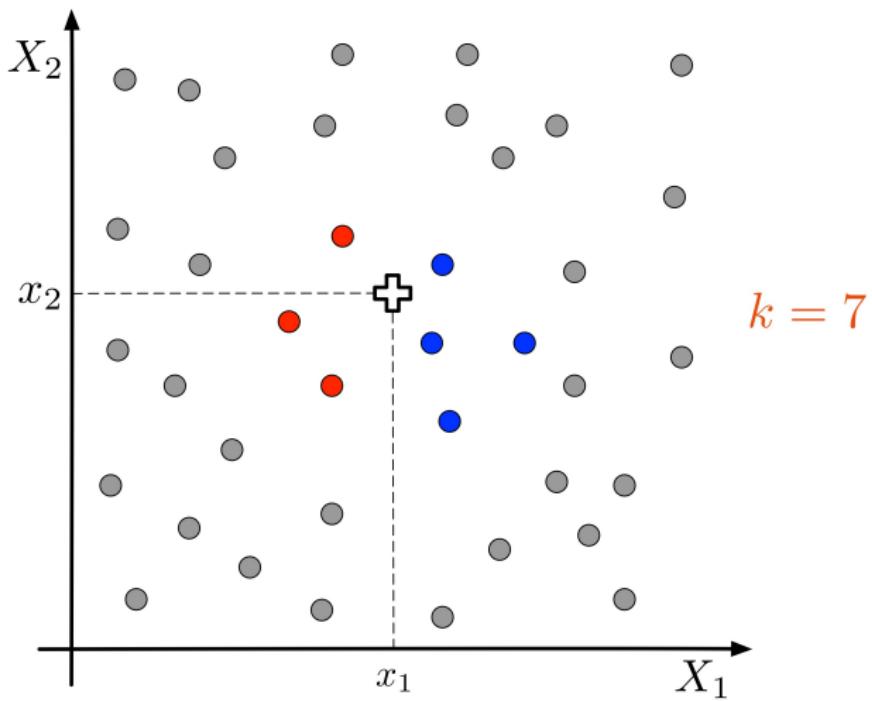


Illustration V



Régression

- ▶ Soit le modèle de régression suivant :

$$Y = m(X) + \varepsilon .$$

Régression

- Soit le modèle de régression suivant :

$$Y = m(X) + \varepsilon .$$

- La régression des k -plus proches voisins est :

$$\widehat{m}_{kpp}(x) = \frac{1}{k} \sum_{i=1}^n \mathbb{1}_{d(x, X_i) \leq D_{(k)}} Y_i .$$

Choix de k

- ▶ Plus k est important, plus on « lisse ».

Choix de k

- ▶ Plus k est important, plus on « lisse ».
- ▶ Le biais croît avec k .

Choix de k

- ▶ Plus k est important, plus on « lisse ».
- ▶ Le **biais croît avec k .**
- ▶ La **variance décroît avec k .**

Choix de k

- ▶ Plus k est important, plus on « lisse ».
- ▶ Le biais croît avec k .
- ▶ La variance décroît avec k .
- ▶ On peut déterminer une valeur de k optimale par validation croisée.

Remarques

- ▶ Cette méthode est très simple à mettre en œuvre.

Remarques

- ▶ Cette méthode est très simple à mettre en œuvre.
- ▶ Elle ne fournit pas de bonnes prévisions en général.

Le coin R

On peut utiliser plusieurs packages, parmi lesquels :

Le coin R

On peut utiliser plusieurs packages, parmi lesquels :

- ▶ Le package `gbm` :
 - ▶ La fonction de base est `gbm`.
 - ▶ On utilise l'option :
 - ▶ `method="ada"` pour AdaBoost,
 - ▶ `method="LogitBoost"` pour LogitBoost,
 - ▶ `method="gaussian"` pour L_2 -Boosting.
 - ▶ L'option `n.trees` permet de choisir le nombre d'arbres.
 - ▶ L'option `interaction.depth` permet de choisir la complexité des arbres.
 - ▶ L'option `shrinkage` correspond à λ .
 - ▶ La fonction `gbm.perf` permet de choisir le nombre d'itérations optimal.

Le coin R

On peut utiliser plusieurs packages, parmi lesquels :

- ▶ Le package **gbm** :
 - ▶ La fonction de base est **gbm**.
 - ▶ On utilise l'option :
 - ▶ **method="ada"** pour AdaBoost,
 - ▶ **method="LogitBoost"** pour LogitBoost,
 - ▶ **method="gaussian"** pour L_2 -Boosting.
 - ▶ L'option **n.trees** permet de choisir le nombre d'arbres.
 - ▶ L'option **interaction.depth** permet de choisir la complexité des arbres.
 - ▶ L'option **shrinkage** correspond à λ .
 - ▶ La fonction **gbm.perf** permet de choisir le nombre d'itérations optimal.
- ▶ Le package **caret** :
 - ▶ La fonction de base est **train**.
 - ▶ On utilise une des options correspondant aux SVM, par exemple :
 - ▶ **method="ada"** pour AdaBoost,
 - ▶ **method="LogitBoost"** pour LogitBoost.

Références

- Epanechnikov, V. A. 1969, «Nonparametric estimation of a multidimensional probability density», *Theory of Probability and Applications*, vol. 14, n° 1, p. 156–161.
- Fan, J. et I. Gijbels. 1996, *Local polynomial modelling and its applications*, Monographs on Statistics & Applied Probability (66), Chapman & Hall.
- Härdle, W. 1992, *Applied nonparametric regression*, Econometric Society Monographs, vol. 19, Cambridge University Press.
- Härdle, W., M. Müller, S. Sperlich et A. Werwatz. 2004, *Nonparametric and semiparametric models*, Springer.
- Silverman, B. W. 1986, *Density estimation for statistics and data analysis*, Monographs on Statistics & Applied Probability (26), Chapman & Hall.