# Chapitre 4: Méthode non linéaire en apprentissage supervisé: Classifieurs plus proches voisins K-NN

2 Ingénierie des Données et Systèmes Décisionnels

**Ecole Nationale d'Electronique et des Télécommunications de Sfax**

05 Octobre 2022

# Plan

1. Introduction
2. Lien entre estimation de la fonction de régression et classification
3. Classifieurs plus proches voisins

- On se donne $D_n = \{(X_i, Y_i)_{1 \leq i \leq n}\}$ un échantillon d'observations.
- $X_i \in \mathbb{R}^d$ et $Y_i \in \mathcal{Y}$.
- Si $\mathcal{Y} = \mathbb{R}$, alors il s'agit d'un problème de régression.
- Si $\mathcal{Y}$ est ensemble fini, alors il s'agit d'un problème de classification.

### Définition

Un prédicteur (ou un classificateur dans le cadre d'un problème de classification) est une fonction mesurable $g : \mathbb{R}^d \longrightarrow \mathcal{Y}$

### Problème

L'objectif est de définir un prédicteur empirique $\hat{g}$ tel que

$$\underbrace{\hat{g}(X_{n+1})}_{\text{Sortie prédite}} = \underbrace{Y_{n+1}}_{\text{Vraie sortie}} ,$$

avec $(X_{n+1}, Y_{n+1})$ est indépendante des autres observations.

## Définition

Un prédicteur empirique est une application $\hat{g} : \underbrace{x}_{\text{données}} \longmapsto \hat{g}(x)$ telle que $g(X_{n+1})$ est une prédiction de $Y_{n+1}$.

## Remarques

- Un prédicteur empirique est construit à partir des données observées.
- En classification, on parle de régle de classification à la place de prédicteur empirique.

## Définition

On appelle risque associé d'un prédicteur empirique la quantité définie par

$$\mathbb{E}\left[\ell(\hat{g}(X_{n+1}); Y_{n+1})\right],$$

où $\ell : \mathcal{Y} \times \mathcal{Y} \longrightarrow \mathbb{R}$ est une fonction de perte.

## Exemples

- Dans le cas d'une régression,

$$\ell(Y, Y') = \| Y - Y' \|^2.$$

- Dans le cas d'une classification

$$\ell(Y, Y') = \mathbf{1}_{\{Y \neq Y'\}}.$$

## Définition

On appelle prédicteur de Bayes (ou classifieur de Bayes) le prédicteur $g^*$ tel que

$$g^* \in \arg \min_{g:\mathbb{R}^d \to \mathcal{Y} mesurable} (R(g)).$$

## Remarques

- $R(g)$ est déterministe et $R(\hat{g})$ est aléatoire.
- La fonction qui minimise

$$R(g) = \mathbb{E}\left((g(x) - y)^2\right)$$

est

$$g^* = \mathbb{E}(Y \mid X = x).$$

## Proposition

- Si $\mathcal{Y} = \mathbb{R}$ et $\ell(y, y') = (y - y')^2$, alors

$$g^*(x) = \mathbb{E}(Y \mid X = x).$$

La fonction $x \longmapsto \mathbb{E}(Y \mid X = x)$ est dite la fonction de régression.

- Si $\mathcal{Y} = \{0; 1\}$ et $\ell(y, y') = \mathbf{1}_{\{y \neq y'\}}$, alors

$$g^*(x) = \mathbf{1}_{\{\mathbb{P}(Y=1|X=x) > \mathbb{P}(Y=0|X=x)\}}.$$

La fonction $x \longmapsto \mathbf{1}_{\{\mathbb{P}(Y=1|X=x) > \mathbb{P}(Y=0|X=x)\}}$ est dite la fonction de classification binaire.

## Démonstration (Régression)

- soit $g$ un prédicteur quelconque, alors

$$\begin{aligned} R(g) &= \mathbb{E}\left(\ell(g(X_{n+1}), Y_{n+1})\right) (\text{comme } \ell(y, y') = (y - y')^2) \\ &= \mathbb{E}\left((g(X_{n+1}) - Y_{n+1})^2\right). \end{aligned} \tag{1}$$

on pose

$$\eta(x) = \mathbb{E}(Y \mid X = x).$$

Montrons maintenant que pour tout prédicteur g, $R(\eta) \leq R(g)$.
D'après (**??**) et en insérant $\eta(X_{n+1})$, on peut écrire

$$\begin{aligned} R(g) &= \mathbb{E}\left((g(X_{n+1}) - \eta(X_{n+1}) + \eta(X_{n+1}) - Y_{n+1})^2\right). \\ &= \underbrace{\mathbb{E}\left((g(X_{n+1}) - \eta(X_{n+1}))^2\right)}_{\geq 0} + \underbrace{\mathbb{E}\left((\eta(X_{n+1}) - Y_{n+1})^2\right)}_{R(\eta)} \\ &\quad + 2\underbrace{\mathbb{E}\left((g(X_{n+1}) - \eta(X_{n+1}))(\eta(X_{n+1}) - Y_{n+1})\right)}_{=0?}. \end{aligned} \tag{2}$$

## Démonstration (Régression)

En effet, on a

$$\mathbb{E}\left[(g(X_{n+1}) - \eta(X_{n+1}))(\eta(X_{n+1}) - Y_{n+1})\right]$$

$$= \mathbb{E}\left[\mathbb{E}[(g(X_{n+1}) - \eta(X_{n+1})(\eta(X_{n+1}) - Y_{n+1})) \mid X_{n+1}]\right]$$

$$= \mathbb{E}[(g(X_{n+1}) - \eta(X_{n+1}))\underbrace{\mathbb{E}[(\eta(X_{n+1}) - Y_{n+1}) \mid X_{n+1}]}_{=0,\ \text{car}\ \mathbb{E}(\eta(X_{n+1}))=\mathbb{E}(Y|X=x)}].$$

Ceci avec (**??**), affirme que pour tout prédicteur g,

$$R(g) \geq R(\eta).$$

## Remarque

De plus, on a l'égalité (i.e., $R(g) = R(\eta)$) lorsque $g(X_{n+1}) = \eta(X_{n+1})$.

## Démonstration (Classification Binaire)

Soit $g : \mathbb{R}^d \longrightarrow \{0; 1\}$ un classifieur.

$$\begin{aligned} R(g) &= \mathbb{E}\left(\mathbf{1}_{\{g(X_{n+1}) \neq Y_{n+1}\}}\right). \\ &= \mathbb{E}(\underbrace{\mathbb{E}\left(\mathbf{1}_{\{g(X_{n+1}) \neq Y_{n+1}\}} \mid X_{n+1}\right)}_{\mathbb{P}((g(X_{n+1}) \neq Y_{n+1}) | X_{n+1})}). \end{aligned}$$

Calcul de $\mathbb{E}\left(\mathbf{1}_{\{g(X_{n+1}) \neq Y_{n+1}\}} \mid X_{n+1}\right) = \mathbb{P}\left((g(X_{n+1}) \neq Y_{n+1}) \mid X_{n+1}\right)$?

Comme $Y_{n+1} \in \mathcal{Y} = \{0; 1\}$, alors on peut écrire

$$(g(X_{n+1}) \neq Y_{n+1}) = (g(X_{n+1}) = 0; Y_{n+1} = 1) \cup (g(X_{n+1}) = 1; Y_{n+1} = 0).$$

## Démonstration (Classification Binaire)

Il s'en suit que

$$\mathbb{P}\left((g(X_{n+1}) \neq Y_{n+1}) \mid X_{n+1}\right)$$

$$= \mathbb{P}\left((g(X_{n+1}) = 0; Y_{n+1} = 1) \mid X_{n+1}\right) + \mathbb{P}\left((g(X_{n+1}) = 1; Y_{n+1} = 0) \mid X_{n+1}\right)$$

$$= \mathbf{1}_{\{g(X_{n+1})=0\}} \mathbb{P}\left(Y_{n+1} = 1 \mid X_{n+1}\right) + \mathbf{1}_{\{g(X_{n+1})=1\}} \mathbb{P}\left(Y_{n+1} = 0 \mid X_{n+1}\right). \tag{3}$$

On pose

$$\eta(x) = \mathbb{E}(Y_{n+1} \mid X_{n+1} = x) = \mathbb{P}(Y_{n+1} = 1 \mid X_{n+1} = x)$$

En insérant ceci dans (**??**), on peut écrire

$$\mathbb{P}\left((g(X_{n+1}) \neq Y_{n+1}) \mid X_{n+1}\right)$$

$$= \mathbf{1}_{\{g(X_{n+1})=0\}} \eta(X_{n+1}) + \mathbf{1}_{\{g(X_{n+1})=1\}} (1 - \eta(X_{n+1})).$$

## Démonstration (Classification Binaire)

Comme $\mathcal{Y} = \{0; 1\}$, alors $\mathbf{1}_{\{g(X_{n+1})=1\}} = g(X_{n+1})$. Ainsi

$$\mathbb{P}\left((g(X_{n+1}) \neq Y_{n+1}) \mid X_{n+1}\right)$$

$$= (1 - g(X_{n+1}))\eta(X_{n+1}) + g(X_{n+1})(1 - \eta(X_{n+1})).$$

Par suite

$$\begin{aligned} R(g) &= \mathbb{E}[\mathbb{P}\left((g(X_{n+1}) \neq Y_{n+1}) \mid X_{n+1}\right)] \\ &= \mathbb{E}\left[(1 - g(X_{n+1}))\eta(X_{n+1}) + g(X_{n+1})(1 - \eta(X_{n+1}))\right]. \end{aligned} \quad (4)$$

## Objectif

L'objectif maintenant est de minimiser $\arg\min_g(R(g)) = \mathbf{1}_{\{\eta(x)>1-\eta(x)\}}$.

## Démonstration (Classification Binaire)

On pose

$$g^*(x) = \mathbf{1}_{\{\eta(x) \geq 1-\eta(x)\}}.$$

### Calcul de $R(g^*)$

On a

$$g^*(X_{n+1})(1-\eta(X_{n+1})) + (1-g^*(X_{n+1}))\eta(X_{n+1})$$

$$= \begin{cases} 1-\eta(X_{n+1}) & si \quad g^*(X_{n+1}) = 1 \\ \eta(X_{n+1}) & si \quad g^*(X_{n+1}) = 0 \end{cases} = \min(1-\eta(X_{n+1}), \eta(X_{n+1})),$$

car

$$g^*(X_{n+1}) = \mathbf{1}_{\{\eta(X_{n+1}) > 1-\eta(X_{n+1})\}}.$$

Ceci, avec (**??**) donne

$$R(g^*) = \mathbb{E}[\min(1-\eta(X_{n+1}), \eta(X_{n+1}))].$$

## Démonstration (Classification Binaire)

Reste à montrer que pour tout $g$, $R(g) \geq R(g^*)$.

Pour tout $g : \mathbb{R}^d \longrightarrow \{0; 1\}$, on a

$$\begin{aligned} R(g) &= \mathbb{E}[g(X_{n+1})(1-\eta(X_{n+1})) + (1-g(X_{n+1}))\eta(X_{n+1})] \\ &\geq \mathbb{E}[\min(1-\eta(X_{n+1}), \eta(X_{n+1}))], \end{aligned}$$

car

$$g(X_{n+1})(1-\eta(X_{n+1})) + (1-g(X_{n+1}))\eta(X_{n+1})$$

$$= \begin{cases} 1-\eta(X_{n+1}) & si \quad g(X_{n+1}) = 1 \\ \eta(X_{n+1}) & si \quad g(X_{n+1}) = 0 \end{cases} \geq \min(1-\eta(X_{n+1}), \eta(X_{n+1})).$$

Ainsi pour tout $g$,

$$R(g) \geq R(g^*).$$

## Remarques

- En pratique on ne connaît pas la loi des observations puisque $g^*$ n'est pas observable car il dérive de la fonction $\eta$ qui n'est pas observable $\eta = \mathbb{E}(Y \mid X) \Longrightarrow$ On ne connaît pas le prédicteur de Bayes.
- Le risque du prédicteur de Bayes est appelé le risque de Bayes et on le note $R^* = R(g^*)$: c'est le plus petit risque possible.

Dans le cas de la classification

$$R^* = \mathbb{E}[\min(1 - \eta(X_{n+1}), \eta(X_{n+1}))].$$

En particulier,

$$\begin{aligned} R^* = 0 \quad & \text{si, et seulement si} \quad \min(1 - \eta(X_{n+1}), \eta(X_{n+1})) = 0 \\ & \text{si, et seulement si} \quad \eta(X_{n+1}) \in \{0; 1\} \\ & \text{si, et seulement si} \quad \mathbb{E}[Y_{n+1} \mid X_{n+1}] \in \{0; 1\} \text{ p.s.} \end{aligned}$$

## Dans le cas de la régression

$$\begin{aligned} R^* &= \mathbb{E}[(\eta(X_{n+1}) - Y_{n+1})^2] \\ &= \mathbb{E}[(\mathbb{E}[Y_{n+1} \mid X_{n+1}] - Y_{n+1})^2] \\ &= \mathbb{E}[\mathbb{E}[(\mathbb{E}[Y_{n+1} \mid X_{n+1}] - Y_{n+1})^2] \mid X_{n+1}] \\ &= \mathbb{E}[\mathbb{V}[Y_{n+1} \mid X_{n+1}]]. \end{aligned}$$

En particulier,

$$\begin{aligned} R^* = 0 \quad &\text{si, et seulement si} \quad \mathbb{V}[Y_{n+1} \mid X_{n+1}] = 0 \\ &\text{si, et seulement si} \quad Y_{n+1} = \mathbb{E}[Y_{n+1} \mid X_{n+1}]. \end{aligned}$$

## Définitions

- Le prédicteur empirique $\hat{g}$ est dit faiblement consistant par rapport à une certaine distribution du couple $(X, Y)$, si

$$\mathbb{E}[R(\hat{g})] \xrightarrow[n \to +\infty]{} R^*.$$

- Le prédicteur empirique $\hat{g}$ est dit fortement consistant par rapport à une certaine distribution du couple $(X, Y)$, si

$$R(\hat{g}) \xrightarrow[n \to +\infty]{} R^* \text{ p.s.}$$

- Le prédicteur empirique $\hat{g}$ est dit faiblement (resp. fortement) universellement consistant si il est faiblement (resp. fortement) consistant quelque soit la distribution du couple $(X, Y)$.

## Proposition

Les asertions suivantes sont équivalentes

1. $\hat{g}$ est (faiblement) consistant
2. $R(\hat{g}) \xrightarrow[n\to+\infty]{L^1} R^*$.
3. $R(\hat{g}) \xrightarrow[n\to+\infty]{Proba.} R^*$.

### Rappel convergence

Dans la théorie des probabilités, il existe différentes notions de convergence de variables aléatoires.

### Définition (Convergence en norme $L^1$)

On dit que $(X_n)$ converge vers $X$ en norme $L^1$ si, pour tout $n$, $X_n$ et $X$ ont un moment d'ordre 1 fini et si

$$\lim_{n \to +\infty} \| X_n - X \|_{L^1} = 0$$

où de manière équivalente, si

$$\lim_{n \to +\infty} \mathbb{E}\left[ | X_n - X | \right] = 0.$$

Dans ce cas on note $X_n \xrightarrow{L^1} X$.

### Définition (Convergence en Probabilité)

On dit que $(X_n)$ converge vers $X$ en probabilité si, pour tout $\varepsilon > 0$

$$\lim_{n \to +\infty} \mathbb{P}\left(\mid X_n - X \mid \geq \varepsilon\right) = 0.$$

Dans ce cas on note $X_n \xrightarrow{\mathbb{P}} X$.

### Propriété ($L^1$ implique en Probabilité)

Si $(X_n)$ et $X$ sont dans $L^1$ et si $X_n \xrightarrow{L^1} X$, alors $X_n \xrightarrow{\mathbb{P}} X$.

## Démonstartion $1 \Longrightarrow 2$

$\hat{g}$ est (faiblement) consistant $\Longrightarrow R(\hat{g}) \xrightarrow[n\to+\infty]{L^1} R^*$?

D'après la définition de $g^*$, on a pour tout classifieur $\hat{g}$,

$$R(\hat{g}) \geq \inf_g \{R(g)\}.$$

Donc

$$R(\hat{g}) \geq R^*.$$

Par suite et en utilisant le fait que $\hat{g}$ est consistant, on déduit que

$$\mathbb{E}[\mid R(\hat{g}) - R^* \mid] = \mathbb{E}[R(\hat{g}) - R^*] = \mathbb{E}[R(\hat{g})] - R^* \xrightarrow[n\to+\infty]{} 0.$$

D'où

$$R(\hat{g}) \xrightarrow[n\to+\infty]{L^1} R^*.$$

## Démonstartion $2 \Longrightarrow 1$

$R(\hat{g}) \xrightarrow[n \to +infty]{L^1} R^* \Longrightarrow \hat{g}$ est (faiblement) consistant?

On a

$$\begin{aligned} \mathbb{E}[R(\hat{g})] - R^* &= \mathbb{E}[R(\hat{g}) - R^*] \\ &= \mathbb{E}[|\ R(\hat{g}) - R^*\ |] (\text{ car } R(\hat{g}) \geq R^*.) \\ &\xrightarrow[n \to +\infty]{} 0 \text{ (hypothèse).} \end{aligned}$$

Il s'en suit que

$$\mathbb{E}[R(\hat{g})] \xrightarrow[n \to +\infty]{} R^*.$$

D'où $\hat{g}$ est consistant.

## Inégalité de Markov

Soit $X$ une variable aléatoire réelle définie sur un espace probabilisé $(\Omega, \mathcal{A}, \mathbb{P})$ et supposée presque sûrement positive ou nulle. Alors pour tout $a > 0$,

$$\mathbb{P}[X \geq a] \leq \frac{\mathbb{E}(X)}{a}.$$

## Démonstartion $2 \Longrightarrow 3$

$$R(\hat{g}) \xrightarrow[n \to +\infty]{L^1} R^* \Longrightarrow R(\hat{g}) \xrightarrow[n \to +\infty]{\mathbb{P}} R^* ?$$

Soit $\varepsilon > 0$,

$$\begin{aligned} \mathbb{P}[\mid R(\hat{g}) - R^* \mid > \varepsilon] \quad &\leq \quad \frac{1}{\varepsilon}\mathbb{E}[\mid R(\hat{g}) - R^* \mid] \\ &\xrightarrow[n \to +\infty]{} \quad 0 \text{ (hypothèse).} \end{aligned}$$

D'où $R(\hat{g})$ converge en probabilité vers $R^*$.

## Démonstartion $3 \Longrightarrow 2$

$$R(\hat{g}) \xrightarrow[n\to+\infty]{\mathbb{P}} R^* \Longrightarrow R(\hat{g}) \xrightarrow[n\to+\infty]{L^1} R^* ?$$

Soit $\varepsilon > 0$, $\mathbb{E}[\mid R(\hat{g}) - R^* \mid]$

$$\begin{aligned} &= \mathbb{E}[\underbrace{\mid R(\hat{g}) - R^* \mid}_{\leq 1 \text{ car } R(\hat{g}), R^* \in [0;1]} \mathbf{1}_{\{|R(\hat{g}) - R^*| > \varepsilon\}}] + \underbrace{\mathbb{E}[\mid R(\hat{g}) - R^* \mid \mathbf{1}_{\{|R(\hat{g}) - R^*| \leq \varepsilon\}}]}_{\leq \varepsilon} \\ &\leq \underbrace{\mathbb{P}[\mid R(\hat{g}) - R^* \mid > \varepsilon]}_{\xrightarrow[n\to+\infty]{} 0 \text{ (hypothèse)}} + \varepsilon. \end{aligned}$$

Ainsi pour tout $\varepsilon > 0$,

Ceci implique que

$$\lim_{n\to+\infty} \mathbb{E}[\mid R(\hat{g}) - R^* \mid] \leq \varepsilon.$$

$$\lim_{n\to+\infty} \mathbb{E}[\mid R(\hat{g}) - R^* \mid] = 0.$$

D'où $R(\hat{g}) \xrightarrow[n\to+\infty]{L^1} R^*$.

# Plan

1 Introduction

2 Lien entre estimation de la fonction de régression et classification

3 Classifieurs plus proches voisins

# Rappel

- $Y \in \{0; 1\}$
- $\arg\min_g(R(g)) = g^*$, avec

$$g^* = \mathbf{1}_{\{\eta(x) \geq 1-\eta(x)\}} = \mathbf{1}_{\{\eta(x) \geq \frac{1}{2}\}},$$

où

$$\eta(x) = \mathbb{E}(Y \mid X = x) = \mathbb{P}[Y = 1 \mid X = x] : \text{ Fonction de régression.}$$

- Soit $\hat{\eta}$ un estimateur de la fonction de régression $\eta$. Un classifieur naturel serait

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

## Proposition

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

$$0 \leq R(\hat{g}) - R^* \leq 2\mathbb{E}[\mid \hat{\eta}(x) - \eta(x) \mid\mid D_n],$$

où $D_n = \{(X_i, Y_i)_{1 \leq i \leq n}\}$ est l'échantillon d'observations et pour tout $p \geq 1$,

$$0 \leq R(\hat{g}) - R^* \leq 2(\mathbb{E}[\mid \hat{\eta}(x) - \eta(x) \mid^p \mid D_n])^{\frac{1}{p}}. \quad (5)$$

## Démonstration

Comme $Y = \{0; 1\}$, alors

$$\begin{aligned} R(\hat{g}) &= \mathbb{P}[\hat{g}(X) \neq Y \mid D_n] \\ &= 1 - \mathbb{P}[\hat{g}(X) = Y \mid D_n] \\ &= 1 - \mathbb{P}[\hat{g}(X) = 0, Y = 0 \mid D_n] - \mathbb{P}[\hat{g}(X) = 1, Y = 1 \mid D_n]. \end{aligned}$$

Pour la suite de la démonstration, nous avons besoin du lemme suivant.

## Lemme

- $\mathbb{P}[\hat{g}(X) = 0, Y = 0 \mid D_n] = \mathbb{E}[\mathbf{1}_{\{\hat{g}(X)=0\}}(1 - \eta(X)) \mid D_n]$.
- $\mathbb{P}[\hat{g}(X) = 1, Y = 1 \mid D_n] = \mathbb{E}[\mathbf{1}_{\{\hat{g}(X)=1\}}\eta(X) \mid D_n]$.

## Démonstration du Lemme

On sait que

$$\hat{g}(X) = \hat{g}(X, D_n),$$

i.e., $\hat{g}(X)$ dépend de l'échantillon $D_n$. De plus, comme $(X, Y)$ est indépendant de $D_n$ et $\mathbf{1}_{\{Y=0\}} = 1 - Y$, alors

$$\begin{aligned}
\mathbb{P}[\hat{g}(X) = 0, Y = 0 \mid D_n = d_n] &= \mathbb{P}[\hat{g}(X, d_n) = 0, Y = 0 \mid D_n = d_n] \\
&= \mathbb{P}[\hat{g}(X, d_n) = 0, Y = 0] \text{ (indép. )} \\
&= \mathbb{E}[\mathbf{1}_{\{\hat{g}(X,d_n)=0\}}\mathbf{1}_{\{Y=0\}}] \\
&= \mathbb{E}[\mathbf{1}_{\{\hat{g}(X,d_n)=0\}}(1 - Y)] \\
&= \mathbb{E}[\mathbb{E}[\mathbf{1}_{\{\hat{g}(X,d_n)=0\}}(1 - Y) \mid X]].
\end{aligned}$$

## Démonstration du Lemme

Comme $\mathbb{E}[1-Y \mid X] = 1-\eta(X)$ (car $\eta(X) = \mathbb{E}[Y \mid X]$), alors on déduit

$$\begin{aligned}
\mathbb{P}[\hat{g}(X)=0, Y=0 \mid D_n = d_n] &= \mathbb{E}[\mathbb{E}[\mathbf{1}_{\{\hat{g}(X,d_n)=0\}}(1-Y) \mid X]] \\
&= \mathbb{E}[\mathbf{1}_{\{\hat{g}(X,d_n)=0\}}\mathbb{E}[1-Y \mid X]] \\
&= \mathbb{E}[\mathbf{1}_{\{\hat{g}(X,d_n)=0\}}(1-\eta(X))] \\
&= \mathbb{E}[\mathbf{1}_{\{\hat{g}(X,d_n)=0\}}(1-\eta(X)) \mid D_n = d_n] \\
&= \mathbb{E}[\mathbf{1}_{\{\hat{g}(X,D_n)=0\}}(1-\eta(X)) \mid D_n = d_n] \\
&= \mathbb{E}[\mathbf{1}_{\{\hat{g}(X)=0\}}(1-\eta(X)) \mid D_n = d_n].
\end{aligned}$$

Dons presque sûrement

$$\mathbb{P}[\hat{g}(X)=0, Y=0 \mid D_n] = \mathbb{E}[\mathbf{1}_{\{\hat{g}(X)=0\}}(1-\eta(X)) \mid D_n].$$

Revenons maintenant à la démonstration de la proposition.

## Démonstration de la Proposition

D'après le lemme précédent, on a

$$\begin{aligned} R(\hat{g}) &= 1 - \mathbb{P}[\hat{g}(X) = 0, Y = 0 \mid D_n] - \mathbb{P}[\hat{g}(X) = 1, Y = 1 \mid D_n] \\ &= 1 - \mathbb{E}[\mathbf{1}_{\{\hat{g}(X)=0\}}(1 - \eta(X)) \mid D_n] - \mathbb{E}[\mathbf{1}_{\{\hat{g}(X)=1\}}\eta(X) \mid D_n] \quad (6) \end{aligned}$$

De même, comme $g^*$ ne dépend pas des données $(D_n)$ et $X \perp D_n$, alors

$$\begin{aligned} R^* := R(g^*) &= 1 - \mathbb{P}[\hat{g}(X) = 0, Y = 0 \mid D_n] - \mathbb{P}[\hat{g}(X) = 1, Y = 1 \mid D_n] \\ &= 1 - \mathbb{E}[\mathbf{1}_{\{g^*(X)=0\}}(1 - \eta(X))] - \mathbb{E}[\mathbf{1}_{\{g^*(X)=1\}}\eta(X)]. \quad (7) \end{aligned}$$

Ainsi, (**??**) et (**??**) donnent

$$\begin{aligned} R(\hat{g}) - R^* &= \mathbb{E}[(\mathbf{1}_{\{g^*(X)=0\}} - \mathbf{1}_{\{\hat{g}(X)=0\}})(1 - \eta(X)) \mid D_n] \\ &\quad + \mathbb{E}[(\mathbf{1}_{\{g^*(X)=1\}} - \mathbf{1}_{\{\hat{g}(X)=1\}})\eta(X) \mid D_n]. \quad (8) \end{aligned}$$

## Démonstration de la Proposition

Le fait que

$$\mathbf{1}_{\{g^*(X)=0\}} - \mathbf{1}_{\{\hat{g}(X)=0\}} = \mathbf{1}_{\{\hat{g}(X)=1\}} - \mathbf{1}_{\{g^*(X)=1\}}$$

avec (**??**) permet d'écrire

$$R(\hat{g}) - R^* = \mathbb{E}[(\mathbf{1}_{\{g^*(X)=1\}} - \mathbf{1}_{\{\hat{g}(X)=1\}})(2\eta(X) - 1) \mid D_n]. \qquad (9)$$

## Démonstration de la Proposition

D'autre part,

$$\mathbf{1}_{\{g^*(X)=1\}} - \mathbf{1}_{\{\hat{g}(X)=1\}} = \begin{cases} -1 & \text{si} \quad g^*(X) = 0 \quad \text{et} \quad \hat{g}(X) = 1 \\ 1 & \text{si} \quad g^*(X) = 1 \quad \text{et} \quad \hat{g}(X) = 0 \\ 0 & \text{si} \quad g^*(X) = \hat{g}(X). \end{cases}$$

En utilisant le fait que

$$g^*(X) = \mathbf{1}_{\{\eta(X) \geq \frac{1}{2}\}},$$

on aura

$$\mathbf{1}_{\{g^*(X)=1\}} - \mathbf{1}_{\{\hat{g}(X)=1\}} = \begin{cases} -1 & \text{si} \quad \eta(X) \leq \frac{1}{2} \quad \text{et} \quad g^*(X) \neq \hat{g}(X) \\ 1 & \text{si} \quad \eta(X) > \frac{1}{2} \quad \text{et} \quad g^*(X) \neq \hat{g}(X) \\ 0 & \text{si} \quad g^*(X) = \hat{g}(X). \end{cases}$$

Il s'en suit que

$$\mathbf{1}_{\{g^*(X)=1\}} - \mathbf{1}_{\{\hat{g}(X)=1\}} = \textit{signe}(2\eta(X) - 1)\mathbf{1}_{\{g^*(X) \neq \hat{g}(X)\}}.$$

## Démonstration de la Proposition

Ceci, avec (**??**) donne

$$R(\hat{g}) - R^* = \mathbb{E}[signe(2\eta(X) - 1)\mathbf{1}_{\{g^*(X)\neq\hat{g}(X)\}}(2\eta(X) - 1) \mid D_n].$$

Comme

$$signe(x)x = \mid x \mid,$$

alors

$$R(\hat{g}) - R^* = \mathbb{E}[\mid 2\eta(X) - 1 \mid \mathbf{1}_{\{g^*(X)\neq\hat{g}(X)\}} \mid D_n]. \tag{10}$$

Reste à montrer maintenant que

$$\mid 2\eta(X) - 1 \mid \mathbf{1}_{\{g^*(X)\neq\hat{g}(X)\}} \leq 2 \mid \hat{\eta}(X) - \eta(X) \mid \text{ p.s},$$

avec

$$\hat{g}(X) = \mathbf{1}_{\{\hat{\eta}(X)\geq\frac{1}{2}\}} \text{ et } g^*(X) = \mathbf{1}_{\{\eta(X)\geq\frac{1}{2}\}}.$$

## Démonstration de la Proposition

Cas 1. Si $g^*(X) = \mathbf{1}_{\{\eta(X) \geq \frac{1}{2}\}} = 1$ et $\hat{g}(X) = \mathbf{1}_{\{\hat{\eta}(X) \geq \frac{1}{2}\}} \neq g^*(X)$.

$$
\begin{aligned}
| 2\eta(X) - 1 | \mathbf{1}_{\{g^*(X) \neq \hat{g}(X)\}} &= (2\eta(X) - 1)\mathbf{1}_{\{g^*(X)=1\}}\mathbf{1}_{\{\hat{g}(X)=0\}} \\
&= (2\eta(X) - 1)\mathbf{1}_{\{\eta(X) \geq \frac{1}{2}\}}\mathbf{1}_{\{\hat{\eta}(X) < \frac{1}{2}\}} \\
(\text{car } \hat{\eta}(X) < \frac{1}{2}) &\leq 2(\eta(X) - \hat{\eta}(X))\mathbf{1}_{\{\eta(X) \geq \frac{1}{2}\}}\mathbf{1}_{\{\hat{\eta}(X) < \frac{1}{2}\}} \\
&\leq 2 | \hat{\eta}(X) - \eta(X) | .
\end{aligned}
$$

Cas 2. Si $g^*(X) = \mathbf{1}_{\{\eta(X) \geq \frac{1}{2}\}} = 0$ et $\hat{g}(X) = \mathbf{1}_{\{\hat{\eta}(X) \geq \frac{1}{2}\}} \neq g^*(X)$.

$$
\begin{aligned}
| 2\eta(X) - 1 | \mathbf{1}_{\{g^*(X) \neq \hat{g}(X)\}} &= (1 - 2\eta(X))\mathbf{1}_{\{g^*(X)=0\}}\mathbf{1}_{\{\hat{g}(X)=1\}} \\
&= (1 - 2\eta(X))\mathbf{1}_{\{\eta(X) < \frac{1}{2}\}}\mathbf{1}_{\{\hat{\eta}(X) \geq \frac{1}{2}\}} \\
(\text{car } \hat{\eta}(X) \geq \frac{1}{2}) &\leq 2(\hat{\eta}(X) - \eta(X))\mathbf{1}_{\{\eta(X) < \frac{1}{2}\}}\mathbf{1}_{\{\hat{\eta}(X) \geq \frac{1}{2}\}} \\
&\leq 2 | \hat{\eta}(X) - \eta(X) | .
\end{aligned}
$$

## Démonstration de la Proposition

En conclusion

$$| 2\eta(X) - 1 | \mathbf{1}_{\{g^*(X) \neq \hat{g}(X)\}} \leq 2 | \hat{\eta}(X) - \eta(X) | \quad \text{p.s.} \tag{11}$$

Ceci avec (**??**) donne

$$0 \underbrace{\leq}_{\text{Par déf. } (\arg\min_g(R(g)) = g^*)} R(\hat{g}) - R^* \leq 2\mathbb{E}[| \hat{\eta}(x) - \eta(x) || D_n] \text{ p.s.} \tag{12}$$

De plus, d'après l'inégalité de Hölder, on a pour tout $p \geq 1$,

$$\mathbb{E}[| \hat{\eta}(x) - \eta(x) || D_n] \leq (\mathbb{E}[| \hat{\eta}(x) - \eta(x) |^p | D_n])^{\frac{1}{p}}.$$

Ceci achève la démonstration de la proposition.

## Inégalité de Hölder

Soit $p$ et $q$ deux réels strictement positifs conjugués ($\frac{1}{p}+\frac{1}{q}=1$), $f \in L^p$ et $g \in L^q$. Le produit $fg$ appartient à $L^1$. De plus

$$\| fg \|_1 = \int | fg | \leq \| f \|_p \| g \|_q = \left( \int | f |^p \right)^{\frac{1}{p}} \left( \int | g |^q \right)^{\frac{1}{q}}.$$

## Corollaire (Inégalité de Hölder et espérance)

Si $X$ et $Y$ sont deux variables aléatoires réelles sur l'espace probabilisé fini $(\Omega, \mathcal{A}, \mathbb{P})$, alors pour tout réels strictement positifs $p$ et $q$ vérifiant $\frac{1}{p}+\frac{1}{q}=1$, on a

$$\mathbb{E}[| XY |] \leq \mathbb{E}[| X |^p)^{\frac{1}{p}} \mathbb{E}[| Y |^q)^{\frac{1}{q}}.$$

### Remarque

On applique l'espérance aux deux côtés de (**??**), si $\hat{g}(X) = \mathbf{1}_{\{\hat{\eta}(X) \geq \frac{1}{2}\}}$, alors

$$0 \leq \mathbb{E}[R(\hat{g}) - R^*] \leq 2\mathbb{E}[\mid \hat{\eta}(x) - \eta(x) \mid]. \tag{13}$$

### Proposition

Le prédicteur empirique

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

avec

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

est universellement consistant.

## Démonstration

On a

$$\hat{\eta}(X) = \frac{\sum_{\{i,\ X_i=x\}} Y_i}{card\{i;\ X_i = x\}} = \frac{\sum_i Y_i \mathbf{1}_{\{X_i=x\}}}{\sum_i \mathbf{1}_{\{X_i=x\}}} = \frac{\frac{1}{n}\sum_i Y_i \mathbf{1}_{\{X_i=x\}}}{\frac{1}{n}\sum_i \mathbf{1}_{\{X_i=x\}}}. \tag{14}$$

En utilisant la loi forte des grands nombres, on déduit que

$$\frac{1}{n}\sum_i Y_i \mathbf{1}_{\{X_i=x\}} \xrightarrow[n\to+\infty]{} \mathbb{E}[Y\mathbf{1}_{\{X=x\}}] \tag{15}$$

et

$$\frac{1}{n}\sum_i \mathbf{1}_{\{X_i=x\}} \xrightarrow[n\to+\infty]{} \mathbb{E}[\mathbf{1}_{\{X=x\}}] = \mathbb{P}[X=x]. \tag{16}$$

En insérant (**??**) et (**??**) dans (**??**), on obtient

$$\hat{\eta}(X) \xrightarrow[n\to+\infty]{p.s.} \frac{\mathbb{E}[Y\mathbf{1}_{\{X=x\}}]}{\mathbb{P}[X=x]} = \mathbb{E}[Y \mid X=x] = \eta(x).$$

Ceci, avec (**??**) implique que

$$R(\hat{g}) \xrightarrow[n\to+\infty]{p.s.} R^* \ \forall (X, Y)$$

Ainsi, le prédicteur empirique $\hat{g}$ est fortement universellement consistant.

De plus, en utilisant (**??**), on peut facilement déduire que le prédicteur empirique $\hat{g}$ est faiblement universellement consistant.

### Définition (convergence presque sûre)

On dit que $(X_n)$ converge presque sûrement vers $X$ si

$$\mathbb{P}\left[\lim_{n\to+\infty} X_n = X\right] = 1$$

ou de manière équivalente, s'il existe un ensemble négligeable $N \subset \Omega$ tel que pour tout $\omega \in \Omega \backslash N$,

$$X_n(\omega) \xrightarrow[n\to+\infty]{} X(\omega).$$

Dans ce cas on note $X_n \xrightarrow{p.s.} X$.

# Plan

1 Introduction

2 Lien entre estimation de la fonction de régression et classification

3 Classifieurs plus proches voisins

- $D_n = \{(X_i; Y_i)_{1 \leq i \leq n}\}$.
- $X_i \in \mathbb{R}^d$.
- $Y_i \in \{0; 1\}$.

Soit $x \in \mathbb{R}^d$.

Exemple: d=2

Soit

$$i(x) = (i_1(x), i_2(x), \ldots, i_n(x))$$

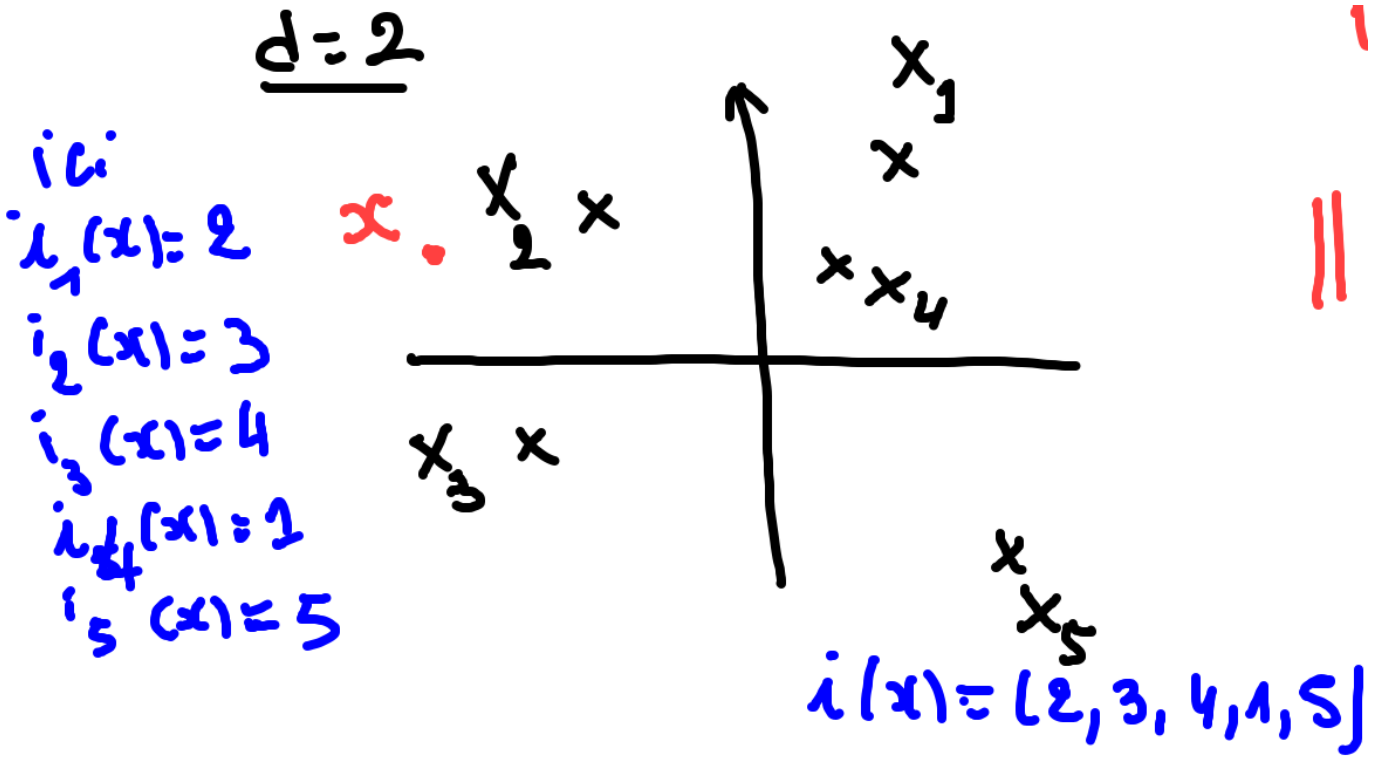
une permutation de $\{1; 2; \ldots; n\}$ telle que

$$\| X_{i_1(x)} - x \|_2 \leq \| X_{i_2(x)} - x \|_2 \leq \ldots \leq \| X_{i_n(x)} - x \|_2,$$

où $i_1$ représente l'indice du plus proche voisin de $x$

et $i_n$ représente l'indice du voisin le plus éloigné de $x$.

## Exemple: d=2



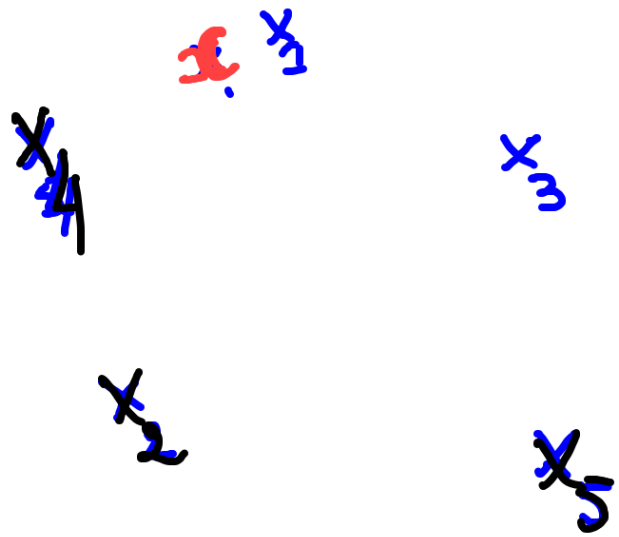
On fixe un entier $k \in \{1; 2; \ldots; n\}$.

## Définition

Le classifieur des $k$ plus proches voisins (k-PPV ou k-NN: k- Nearest Neighbors) est défini par

$$\hat{g}_k(x) = \begin{cases} 1 & \text{si} \quad \frac{1}{k}\sum_{j=1}^{k} Y_{i_j(x)} > \frac{1}{2} \\ 0 & \text{sinon} \end{cases}$$

## Exemple d=2




$Y = 0$ et $Y = 1$.

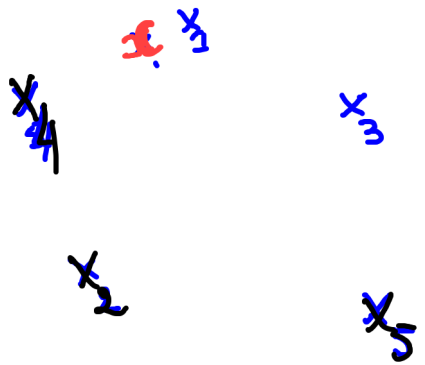
Le point le plus proche à $x$ est $X_1$, donc $i_1(x) = 1$ avec $Y_{i_1(x)} = 1$. D'où

$$\frac{1}{k}\sum_{j=1}^{k} Y_{i_j(x)} = \frac{1}{1}\sum_{j=1}^{1} Y_{i_j(x)} = Y_{i_1(x)} = 1 > \frac{1}{2}.$$

Par suite

$$\hat{g}_1(x) = 1$$

## Exemple d=2




$Y = 0$ et $Y = 1$.

Les deux points les plus proches à $x$ sont

- $X_1$ ($\rightarrow i_1(x) = 1$ donc $Y_{i_1(x)} = 1$).
- $X_4$ ($\rightarrow i_2(x) = 4$ donc $Y_{i_2(x)} = 0$).

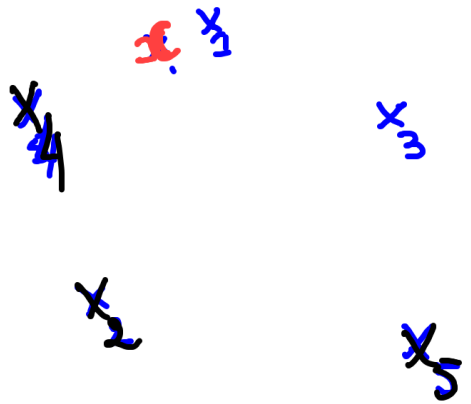
Il s'en suit que

$$\frac{1}{k}\sum_{j=1}^{k} Y_{i_j(x)} = \frac{1}{2}\sum_{j=1}^{2} Y_{i_j(x)} = \frac{1}{2}(Y_{i_1(x)} + Y_{i_2(x)}) = \frac{1}{2}.$$

Par suite

$$\hat{g}_2(x) = 0.$$

## Exemple d=2



$Y = 0$ et $Y = 1$.

Les trois points les plus proches de $x$ sont

- $X_1$ $(\rightarrow i_1(x) = 1$ donc $Y_{i_1} = 1)$
- $X_4$ $(\rightarrow i_2(x) = 4$ donc $Y_{i_2(x)} = 0)$
- $X_3$ $(\rightarrow i_3(x) = 3$ donc $Y_{i_3(x)} = 1)$.

Ainsi

$$\hat{g}_3(x) = \mathbf{1}_{\{\frac{1}{3}\sum_{j=1}^{3} Y_{i_j}(x) > \frac{1}{2}\}} = \mathbf{1}_{\{\frac{2}{3} > \frac{1}{2}\}} = 1.$$

De même on prouve que

$$\hat{g}_4(x) = \mathbf{1}_{\{\frac{2}{4} > \frac{1}{2}\}} = 0.$$

## Définition (fonction lipschitzienne)

Soit $f : \mathbb{R}^d \longrightarrow \mathbb{R}$ une application et $L$ un réel positif. on dit que $f$ est $L-$lipschitzienne ($L-$lipschitz) si pour tout $x$, $y \in \mathbb{R}^d$,

$$| f(x) - f(y) | \leq L \| x - y \|_2 .$$

## Remarque

$f$ est dite contractante si $f$ est $L-$lischitzienne, avec $L \in [0; 1]$.

## Théorème

- Si $\eta$ est une fonction $L-$lipschitz.
- Si le nombre de k plus proches voisins dépend de $n$ de la façon suivante:

$$\frac{k}{n} \xrightarrow[n \to +\infty]{} 0 \text{ et } k \xrightarrow[n \to +\infty]{} +\infty,$$

alors le classifieur $\hat{g}_k$ est universellement consistant.

## Remarques

- L'hypothèse sur $\eta$ n'est pas nécessaire.
- Les hypothèses sur $k$ sont nécessaires.

## Démonstration

Il suffit de montrer que quelque soit la loi du couple $(X, Y)$, on a

$$\lim_{n \longrightarrow +\infty} \mathbb{E}[R(\hat{g}_k)] = R^*,$$

ou bien

$$R(\hat{g}_k) \xrightarrow[n \to +\infty]{L^1} R^*,$$

ou bien

$$R(\hat{g}_k) \xrightarrow[n \to +\infty]{Proba.} R^*.$$

## Démonstration

On sait que

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

où

$$\hat{\eta}_k(x) = \frac{1}{k} \sum_{j=1}^{k} Y_{i_j(x)}.$$

De plus d'après (**??**), ona pour tout $p \geq 1$,

$$0 \leq R(\hat{g}_k) - R^* \leq 2(\mathbb{E}[\mid \hat{\eta}_k(x) - \eta(x) \mid^p \mid D_n])^{\frac{1}{p}}.$$

Si on choisit $p = 2$, alors le théorème est démontré si

$$\mathbb{E}[\mathbb{E}[\mid \hat{\eta}_k(x) - \eta(x) \mid^2 \mid D_n]^{\frac{1}{2}}] \xrightarrow[n \to +\infty]{} 0.$$

## Démonstration

D'après Hölder, on a

$$\mathbb{E}[Z^{\frac{1}{2}}] \leq (\mathbb{E}[Z])^{\frac{1}{2}}.$$

On pose

$$Z = \mathbb{E}[(\hat{\eta}_k(x) - \eta(x))^2 \mid D_n].$$

Ainsi

$$\mathbb{E}[Z] = \mathbb{E}[\mathbb{E}[(\hat{\eta}_k(x) - \eta(x))^2 \mid D_n]] = \mathbb{E}[(\hat{\eta}_k(x) - \eta(x))^2].$$

Donc le théorème est démontré si on montre que

$$\lim_{n \to +\infty} \mathbb{E}[(\hat{\eta}_k(x) - \eta(x))^2] = 0.$$

## Démonstration

On a

$$\begin{aligned}\mathbb{E}[(\hat{\eta}_k(x)-\eta(x))^2] &= \mathbb{E}[(\frac{1}{k}\sum_{j=1}^{k} Y_{i_j(x)}-\eta(x))^2] \\ = \mathbb{E}[(\underbrace{\frac{1}{k}\sum_{j=1}^{k}(Y_{i_j(x)}-\eta(X_{i_j(x)}))}_{a} &+ \underbrace{\frac{1}{k}\sum_{j=1}^{k}(\eta(X_{i_j(x)}))-\eta(x))^2}_{b}] \\ &\leq 2A+2B, \end{aligned} \tag{17}$$

où

$$A=\mathbb{E}[(\frac{1}{k}\sum_{j=1}^{k}(Y_{i_j(x)}-\eta(X_{i_j(x)})))^2] \text{ et } B=\mathbb{E}[(\frac{1}{k}\sum_{j=1}^{k}(\eta(Y_{i_j(x)}))-\eta(x))^2].$$

On veut, maintenant, montrer que $A \xrightarrow[n\to+\infty]{} 0$ et $B \xrightarrow[n\to+\infty]{} 0$.

## Démonstration

**Majoration de $A$.** En utilisant le fait que

$$(\sum_{j=1}^{k} \alpha_j)^2 = \sum_{j=1}^{k} \alpha_j^2 + 2 \sum_{j \neq j'} \alpha_j \alpha_{j'},$$

on déduit que

$$\begin{aligned} A &= \frac{1}{k^2} \sum_{j=1}^{k} \underbrace{\mathbb{E}[(Y_{i_j(x)} - \eta(X_{i_j(x)}))^2]}_{\leq 1} \\ &+ \frac{1}{k^2} \underbrace{\sum_{j \neq j'} \mathbb{E}[(Y_{i_j(x)} - \eta(X_{i_j(x)}))(Y_{i_{j'}(x)} - \eta(X_{i_{j'}(x)}))]}_{=0?} \\ &\leq \frac{k}{k^2} = \frac{1}{k} \xrightarrow[n \to +\infty]{} 0 \text{ (hypothèse). Donc } A \xrightarrow[n \to +\infty]{} 0 \end{aligned} \tag{18}$$

## Démonstration

**Montrons que** $\sum_{j\neq j'} \mathbb{E}[(Y_{i_j(x)} - \eta(X_{i_j(x)}))(Y_{i_{j'}(x)} - \eta(X_{i_{j'}(x)}))] = 0.$

$$\sum_{j\neq j'} \mathbb{E}[(Y_{i_j(x)} - \eta(X_{i_j(x)}))(Y_{i_{j'}(x)} - \eta(X_{i_{j'}(x)}))] =$$

$$\sum_{j\neq j'} \sum_{i\neq i'} \mathbb{E}[\mathbf{1}_{\{i_{j(x)}=i\}} \mathbf{1}_{\{i_{j'(x)}=i'\}} (Y_i - \eta(X_i))(Y_{i'} - \eta(X_{i'}))] =$$

$$\sum_{j\neq j'} \sum_{i\neq i'} \mathbb{E}[\mathbf{1}_{\{i_{j(x)}=i\}} \mathbf{1}_{\{i_{j'(x)}=i'\}} \mathbb{E}[(Y_i - \eta(X_i))(Y_{i'} - \eta(X_{i'})) \mid X]].$$

De plus, comme $(X_i, Y_i)$ indep. de $X$ et $(X_{i'}, Y_{i'})$ indep. de $X$, alors

$$\begin{aligned}
\mathbb{E}[(Y_i - \eta(X_i))(Y_{i'} - \eta(X_{i'})) \mid X] &= \mathbb{E}[(Y_i - \eta(X_i))(Y_{i'} - \eta(X_{i'}))] \\
(X_i, Y_i) \text{ indep. } (X_{i'}, Y_{i'}) &= \underbrace{\mathbb{E}[(Y_i - \eta(X_i))]}_{=0} \mathbb{E}[(Y_{i'} - \eta(X_{i'}))] \\
&= 0.
\end{aligned}$$

## Démonstration

**Majoration de $B$.**

$$
\begin{aligned}
B &= \mathbb{E}[(\frac{1}{k}\sum_{j=1}^{k}\eta(X_{i_j(x)}) - \eta(X))^2] \\
&= \frac{1}{k^2}\mathbb{E}[(\sum_{j=1}^{k}(\eta(X_{i_j(x)}) - \eta(X)))^2] \ (\text{car}\frac{1}{k}\sum_{j=1}^{k}\eta(X) = \eta(X)) \\
&\leq \frac{1}{k^2}k\sum_{j=1}^{k}\mathbb{E}[(\eta(X_{i_j(x)}) - \eta(X))^2] \\
&= \frac{1}{k}\sum_{j=1}^{k}\mathbb{E}[(\eta(X_{i_j(x)}) - \eta(X))^2(\mathbf{1}_{\{\|X_{i_j(x)}-X\|_2\geq\delta\}} + \mathbf{1}_{\{\|X_{i_j(x)}-X\|_2<\delta\}})] \\
&\leq \underbrace{\frac{1}{k}\sum_{j=1}^{k}\mathbb{P}[\| X_{i_j(x)} - X \|_2\geq \delta]}_{B_1} + \frac{L^2\delta^2}{k} \ (\eta \text{ est } L-\text{lipschitz}). \qquad (19)
\end{aligned}
$$

## Démonstration

**Contrôle de** $B_1 = \frac{1}{k}\sum_{j=1}^{k} \mathbb{P}[\| X_{i_j(x)} - X \|_2 \geq \delta]$.

On a pour tout $1 \leq j \leq k$,

$$\| X_{i_j(x)} - X \|_2 \leq \| X_{i_k(x)} - X \|_2 .$$

Ainsi

$$B_1 \leq \mathbb{P}[\| X_{i_k(x)} - X \|_2 \geq \delta] = \int \mathbb{P}[\| X_{i_k(x)} - x \|_2 \geq \delta] dP_X(x).$$

## Démonstration

**On montre maintenant que**

$$\mathbb{P}[\| X_{i_k(x)} - x \|_2 \geq \delta] \xrightarrow[n \to +\infty]{} 0 \ P_X \ \textbf{p.p.}.$$

## Idée

On introduit la quantité

$$S_n(x) = \mathbf{1}_{\{\|X_i - x\|_2 \leq \delta\}}.$$

## Démonstration

Notons que $S_n(X)$ est une variable aléatoire qui suit une loi Binomiale de paramètres $n$ et $p_\delta(x) = \mathbb{P}[\| X_i - X \|_2 \leq \delta]$, i.e.,

$$S_n(X) \sim B(n, p_\delta(x)).$$

De plus $\| X_{i_k(x)} - X \|_2 > \delta$ signifie que le $k^{ieme}$ plus proche voisin de $x$ est à une distance supérieure strictement à $\delta$. Ceci est équivalent à dire que le nombre d'éléments de l'échantillon à une distance inférieure ou égale à $\delta$ de $X$ est strictement inférieur à $k$. i.e.,

$$\| X_{i_k(x)} - X \|_2 > \delta \iff S_n(X) < k.$$

Il s'en suit que

$$B_1 \leq \mathbb{P}[\| X_{i_k(x)} - X \|_2 \geq \delta] = \mathbb{P}[S_n(X) < k]. \quad (20)$$

## Démonstration

Comme $S_n(X) \sim B(n, p_\delta(x))$, alors il existe $Z_1, Z_2, \ldots, Z_n$ iid $Ber(p_\delta(x))$ telle que

$$S_n(X) = Z_1 + Z_2 + \ldots + Z_n$$

et puisque $\mathbb{E}[Z_1] = p_\delta(x) < +\infty$, alors d'après la loi forte des grands nombres

$$\frac{S_n(X)}{n} = \frac{Z_1 + Z_2 + \ldots + Z_n}{n} \xrightarrow[n\to+\infty]{p.s.} \mathbb{E}[Z_1] = p_\delta(x). \tag{21}$$

D'autre part, par hypothèse, on a

$$\frac{k}{n} \xrightarrow[n\to+\infty]{} 0. \tag{22}$$

En combinant (**??**) et (**??**), on déduit que

$$\mathbb{P}[S_n(X) < k] = \mathbb{P}\left[\frac{S_n(X)}{n} < \frac{k}{n}\right] \xrightarrow[n\to+\infty]{p.s.} \mathbb{P}[p_\delta(X) \leq 0] = 0.$$

## Démonstration

Ceci avec (**??**) prouve que

$$B_1 \xrightarrow[n\to+\infty]{} 0. \tag{23}$$

Finalement, si on insére(**??**), (**??**) et (**??**) dans (**??**), on déduit que

$$0 \leq \mathbb{E}[R(\hat{g}_k) - R^*] \leq \mathbb{E}[(\hat{\eta}_k(x) - \eta(x))^2] \leq 2A + 2B \xrightarrow[n\to+\infty]{} 0.$$

Ainsi le classifieur $\hat{g}_k = \mathbf{1}_{\{\hat{\eta}_k(x) > \frac{1}{2}\}}$, où $\hat{\eta}_k(x) = \frac{1}{k}\sum_{j=1}^{k} Y_{i_j(x)}$ est universellement consistant.

## Théorème admis

- Si $X$ adet une densité sur $\mathbb{R}^d$,
- Si le nombre de k plus proches voisins vérifie

$$\frac{k}{n} \xrightarrow[n\to+\infty]{} 0 \text{ et } k \xrightarrow[n\to+\infty]{} +\infty,$$

alors pour tout $\varepsilon > 0$, $\exists n_0$, $\exists \gamma_d$ tque pour tout $n \geq n_0$,

$$\mathbb{P}[\mid R(\hat{g}_k) - R^* \mid \geq \varepsilon] \leq 4 \exp\left(-\frac{n\varepsilon^2}{\gamma_d}\right). \tag{24}$$

### Remarque

Le théorème précédent entraine que le classifieur des $k$ plus proches voisins est fortement consistant.

En effet, d'après (**??**) la série $\sum_n \mathbb{P}[\mid R(\hat{g}_k) - R^* \mid \geq \varepsilon]$ est convergente. Il s'en suit que $R(\hat{g}_k)$ converge presque sûrement vers $R^*$.

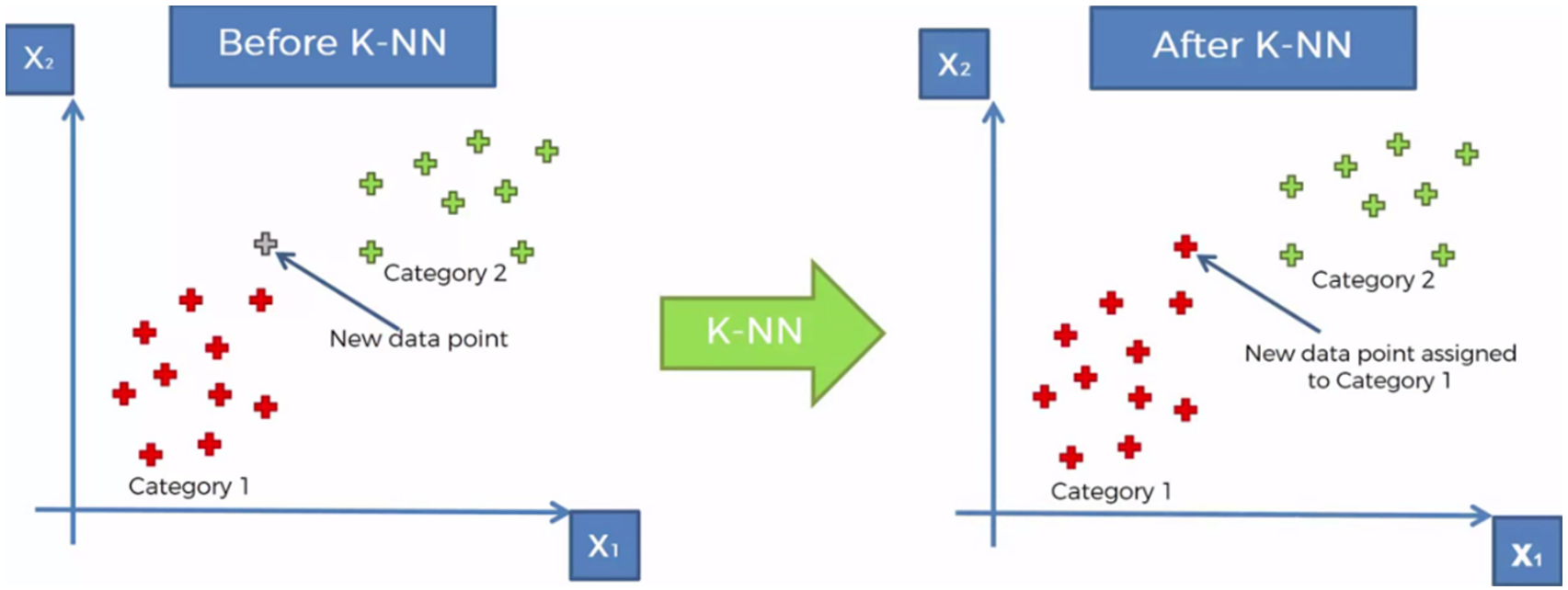
# Choix de $k$

- **Si $k$ est grand**, alors $\hat{g}_k = \mathbf{1}_{\{\hat{\eta}_k(x) > \frac{1}{2}\}}$, où $\hat{\eta}_k(x) = \frac{1}{k}\sum_{j=1}^{k} Y_{i_j(x)}$.
- **Cas extrême $k = n$**, alors $\hat{\eta}_n(x) = \frac{1}{n}\sum_{j=1}^{k} Y_i = \overline{Y}$ et $\hat{g}_k = \mathbf{1}_{\{\overline{Y} > \frac{1}{2}\}}$.
- **Si $k = 1$**, alors $\hat{g}_1(x)$ ne dépend que d'une seule observation, il est donc sensible aux fluctuations de l'échantillon $\Longrightarrow$ Sur-apprentissage.

Remarque

En pratique, on choisit le paramètre par validation croisée.

# Ce que k-NN fait pour vous



## Algorithme des k-NN

- Etape 1: Choisir le nombre $k$ des voisins.
- Etape 2: Prendre les $k$ plus proches voisins de la nouvelle observation en utilisant la distance Euclidienne.

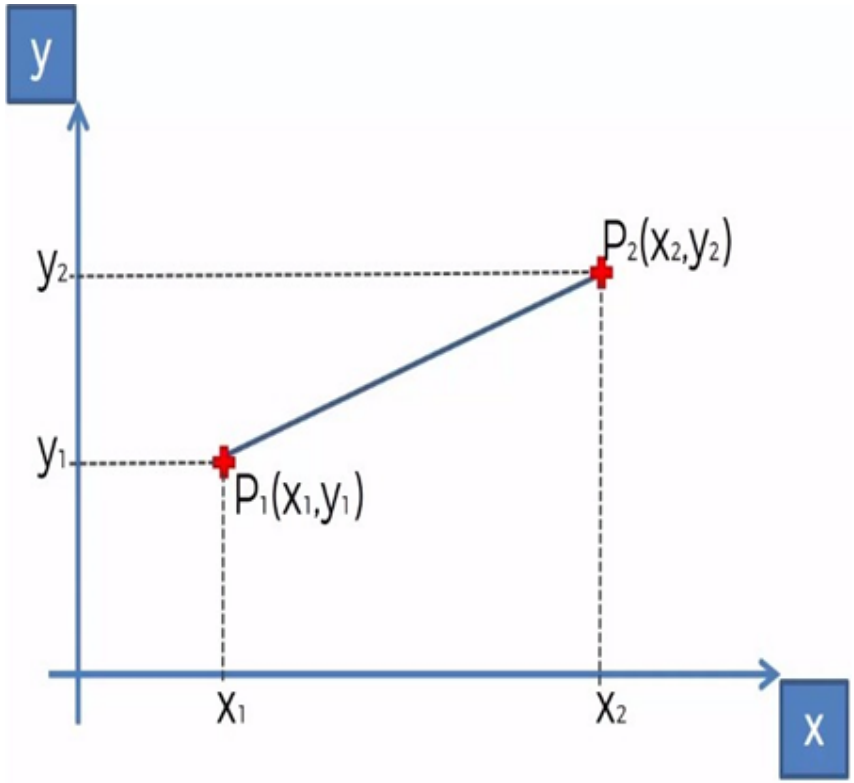


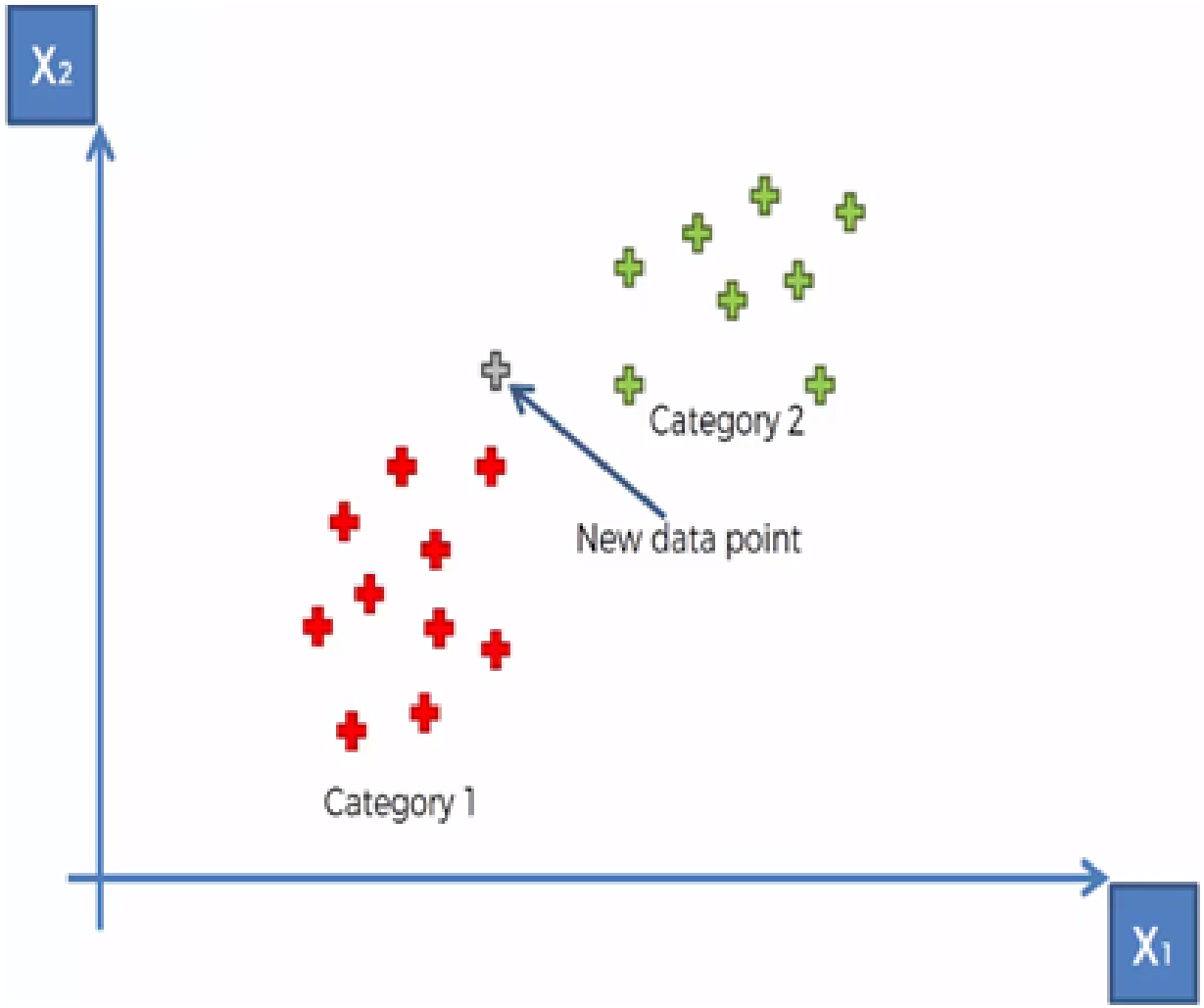

Figure 1: Distance Euclidienne entre $P_1$ et $P_2 = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$.

## Algorithme des k-NN

- Etape 3: Compter parmi ces $k$ voisins le nombre de points par catégorie.
- Etape 4: Affecter la nouvelle observation à la catégorie qui posséde le plus de voisins.

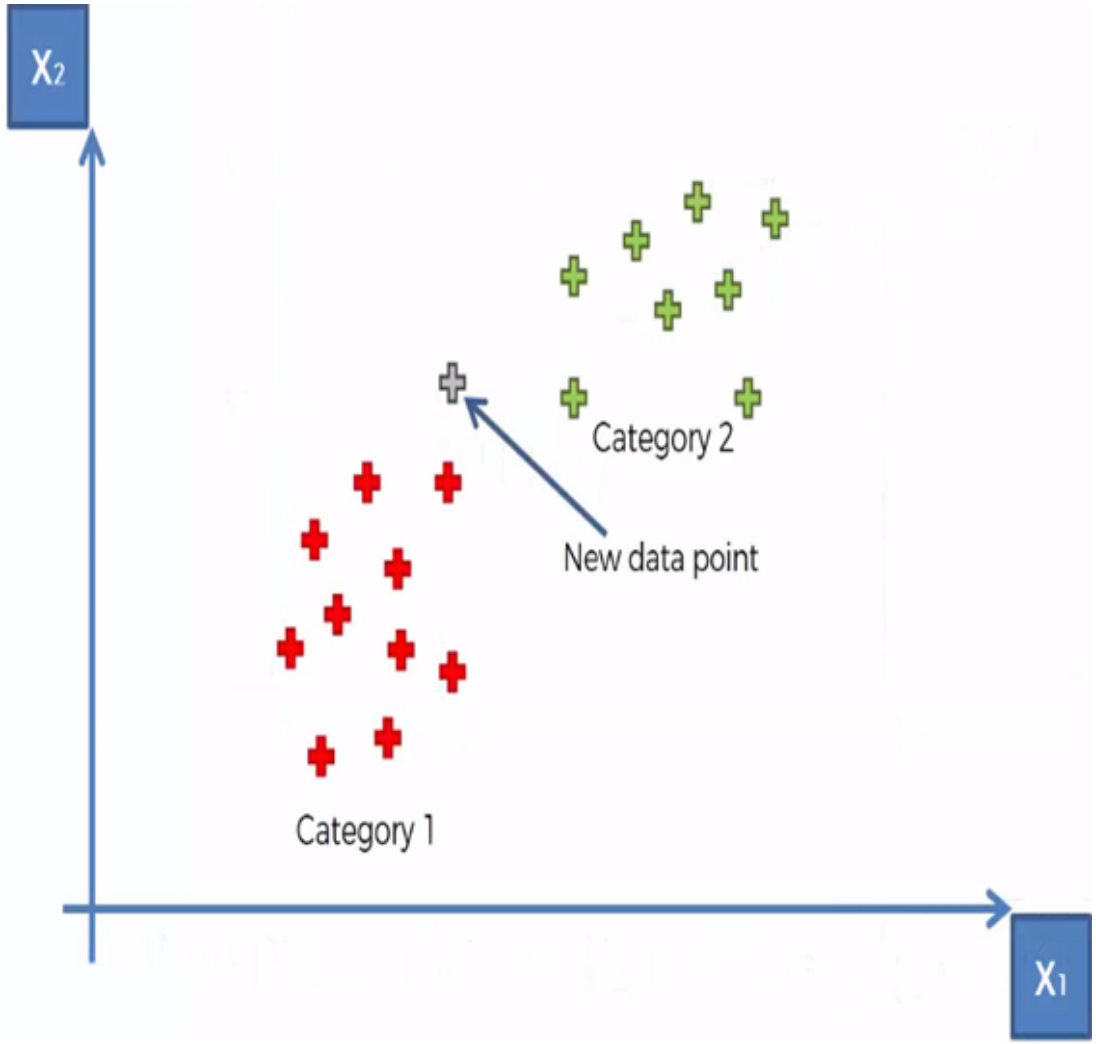
$\Longrightarrow$ Votre modèle est prêt.

# Exemple k-NN




Etape 1: Choisir le nombre $k$ des voisins $k = 5$.

# Exemple k-NN

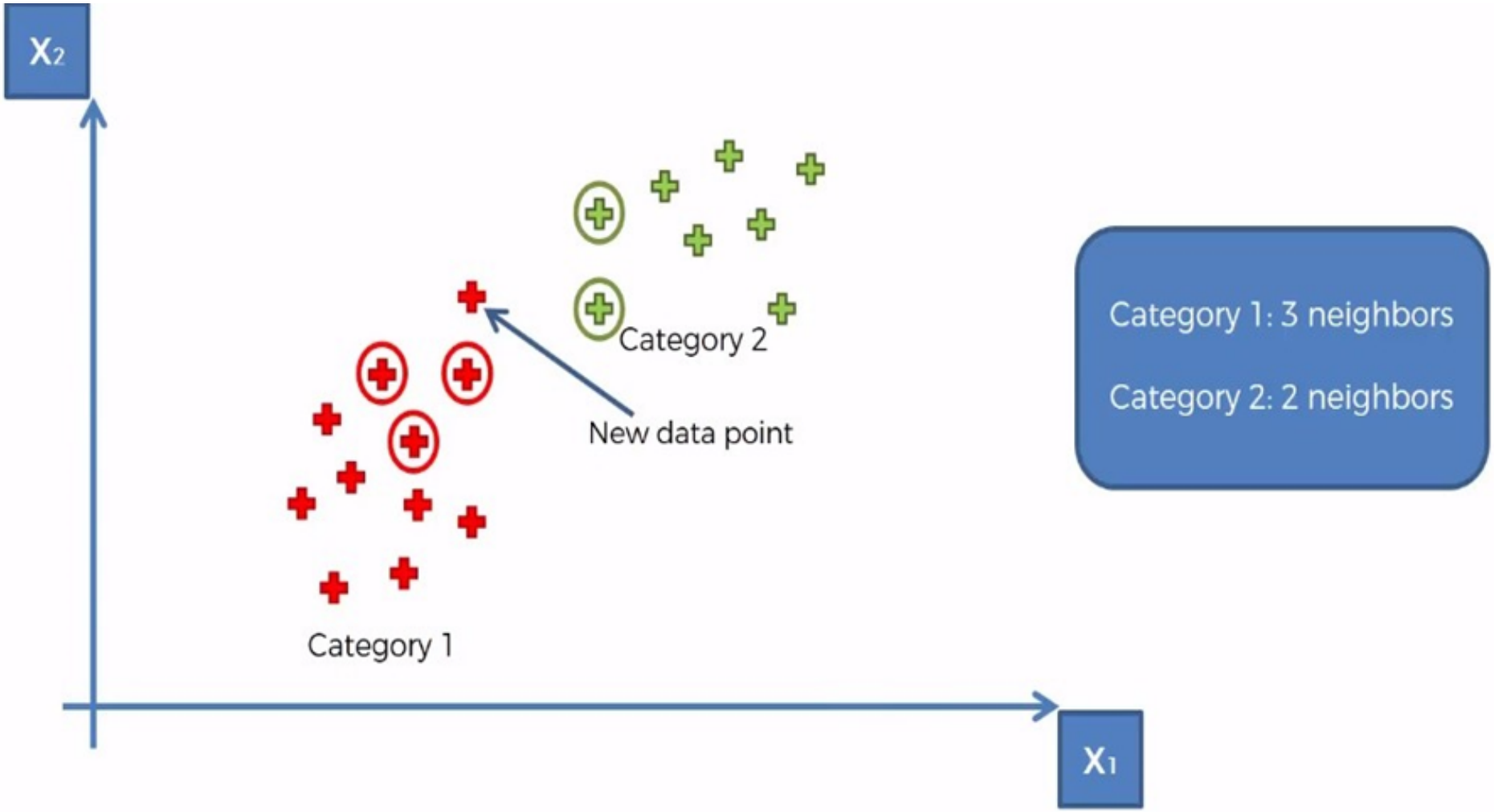



Etape 2: Prendre les $k$ plus proches voisins de la nouvelle observation.

# Exemple k-NN


Etape 3: Compter parmi ces $k$ voisins le nombre de points par catégorie.

# Exemple k-NN




Etape 4: Affecter la nouvelle observation à la catégorie correspondante.

# Exemple k-NN

| Index | User ID | Gender | Age | EstimatedSalary | Purchased |
|---|---|---|---|---|---|
| 0 | 15624510 | Male | 19 | 19000 | 0 |
| 1 | 15810944 | Male | 35 | 20000 | 0 |
| 2 | 15668575 | Female | 26 | 43000 | 0 |
| 3 | 15603246 | Female | 27 | 57000 | 0 |
| 4 | 15804002 | Male | 19 | 76000 | 0 |
| 5 | 15728773 | Male | 27 | 58000 | 0 |
| 6 | 15598044 | Female | 27 | 84000 | 0 |
| 7 | 15694829 | Female | 32 | 150000 | 1 |
| 8 | 15600575 | Male | 25 | 33000 | 0 |
| 9 | 15727311 | Female | 35 | 65000 | 0 |
| 10 | 15570769 | Female | 26 | 80000 | 0 |
| 11 | 15606274 | Female | 26 | 52000 | 0 |

Figure 2: Jeu de données: Social Network.

## k-NN avec Python

```
import pandas as pd
dataset=pd.read_csv('Social_Network_Ads. csv')
X=dataset.iloc[:, [2,3]].values
Y=dataset.iloc[:,4].values
from sklearn.model_selection import train_test_split
X_train,X_test,Y_train,Y_test=train_test_split(X,Y,test_size=0.25)
from sklearn.neighbors import KNeighborsClassifier
classifier = KNeighborsClassifier(n_neighbors = 5, metric = 'minkowski', p =2)
classifier.fit(X_train, y_train)
Y_pred=classifier.predict(X_test)
from sklearn.metrics import confusion_matrix
cm=confusion_matrix(Y_test,Y_pred)
```

On obtient la matrice de confusion suivante:

$$cm = \begin{pmatrix} 64 & 4 \\ 3 & 29 \end{pmatrix}.$$

$\Longrightarrow$ Soit une précision (accuracy) de $64 + 29 = 93\%$.

## k-NN avec Python

```
from matplotlib.colors import ListedColormap
X_set, Y_set = X_train, Y_train
X1, X2 = np.meshgrid(np.arange(start = X_set[:, 0].min() - 1, stop = X_set[:, 0].max() + 1, step = 0.01),
                     np.arange(start = X_set[:, 1].min() - 1, stop = X_set[:, 1].max() + 1, step = 0.01))
plt.contourf(X1, X2, classifier.predict(np.array([X1.ravel(), X2.ravel()]).T).reshape(X1.shape),
             alpha = 0.4, cmap = ListedColormap(('red', 'green')))
plt.xlim(X1.min(), X1.max())
plt.ylim(X2.min(), X2.max())
for i, j in enumerate(np.unique(Y_set)):
    plt.scatter(X_set[Y_set == j, 0], X_set[Y_set == j, 1],
                c = ListedColormap(('red', 'green'))(i), label = j)
plt.title('KNN(Training set)')
plt.xlabel('Age')
plt.ylabel('Estimated Salary')
plt.legend()
```

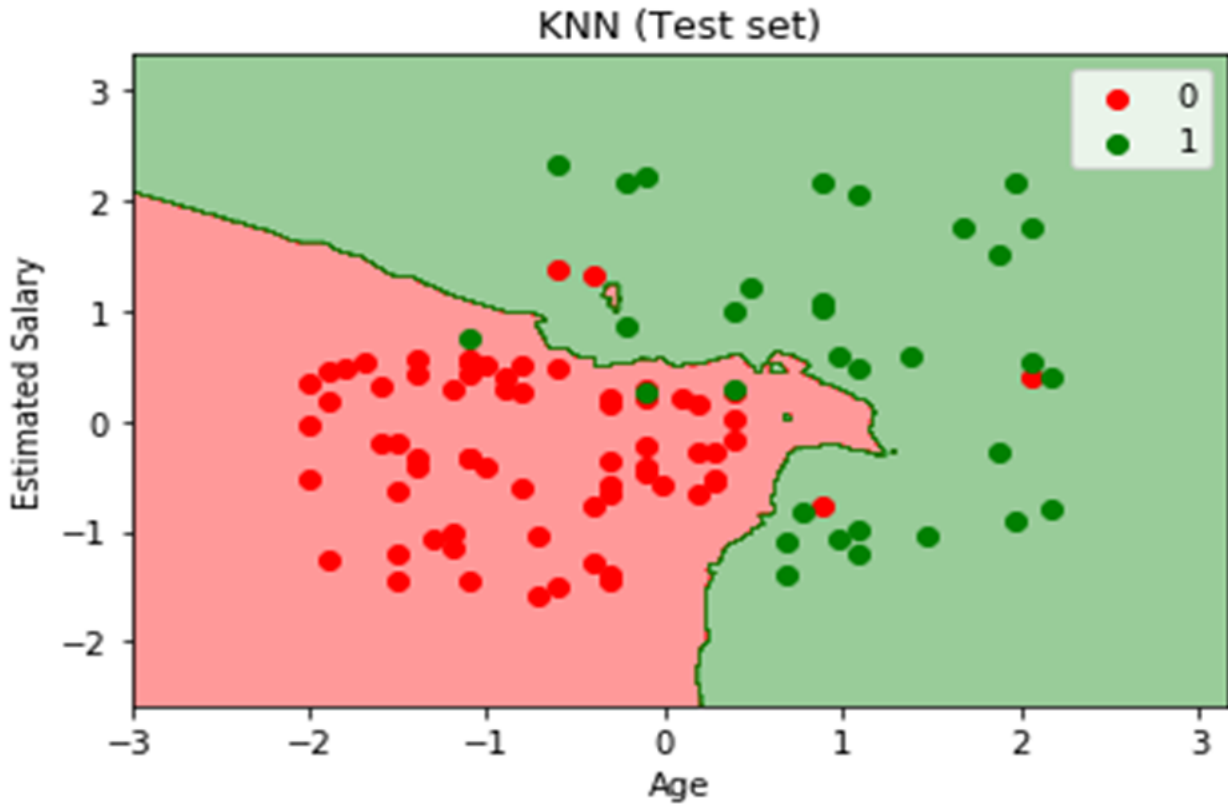



Figure 3: k-NN test set