

## 5. ESTIMATEUR DE LA DENSITÉ PAR K-PLUS-PROCHES-VOISINS

## ESTIMATEUR PAR K-PLUS-PROCHES-VOISINS

- Dans le cas des estimateurs non-paramétriques de densité vus jusqu'à présent nous avons toujours du choisir/ajuster la taille de la fenêtre (via le paramètre  $\nu$ ), avec des conséquences très importantes sur la qualité de l'estimation. Et si on pouvait ajuster cette taille ?
- Une solution est de laisser l'échantillon définir cette taille, qui peut être variable en fonction des zones où les observations de l'échantillon même sont plus ou moins denses



Estimateur pas k-plus-proches-voisins (KNN)

Soit  $\mathcal{D}_N = \{x_1, \dots, x_N\} \subset \mathbb{R}$  un  $N$ -échantillon. Soit  $x \in \mathbb{R}$ . Pour estimer la densité de  $\mathcal{D}_N$  au point  $x$  par la méthode des  $k$ -plus-proches-voisins,  $\hat{f}_k^{KNN}(x)$ , il faut :

1. Déterminer la distance  $d_i$  de tout point dans  $\mathcal{D}_N$  à  $x$  (on peut utiliser la distance Euclidienne) :

$$Dist(x) := \{d_1, \dots, d_N\}, \text{ où, } d_i := ||x - x_i||$$

2. Ordonner l'ensemble  $Dist(x)$  en ordre croissant,  $Dist^{\uparrow}(x)$ , de telle manière que le premier élément de  $Dist^{\uparrow}(x)$  correspondra à la distance du premier plus proche voisin de  $x$ , le deuxième élément celle du deuxième plus proche voisin et ainsi de suite
3. Déterminer  $R_k(x)$ , la distance du  $k$ -ème plus proche voisin de  $x$  (c.a.d. le  $k$ -ème élément de  $Dist^{\uparrow}(x)$ )

Nous sommes maintenant prêts pour calculer  $\hat{f}_k^{KNN}(x)$  :

$$\hat{f}_k^{KNN}(x) = \frac{k}{N} \cdot \frac{1}{V \cdot R_k(x)}$$

Où  $V = 2$  car on s'est placé dans un cas unidimensionnel (c.a.d. nous avons considéré un échantillon à valeurs dans  $\mathbb{R}$ , et non dans  $\mathbb{R}^d$ )

Pour fixer les idées :

- Considérons l'échantillon suivant, composé de 10 observations :

$$\mathcal{D}_{10} = \{3, 9, 15, 13, 11, 5, 4, 7, 10, 6\}$$

- Soit  $x = 8$ .

**EXERCICE.** Déterminer  $\hat{f}_k^{KNN}(x)$  pour  $k = 2$  et pour  $k = 5$

Pour fixer les idées :

- Considérons l'échantillon suivant, composé de 10 observations :

$$\mathcal{D}_{10} = \{3, 9, 15, 13, 11, 5, 4, 7, 10, 6\}$$

- Soit  $x = 8$ .

**SOLUTION.**

$$Dist(8) = \{5, 1, 7, 5, 3, 3, 4, 1, 2, 2\} \Rightarrow Dist^{\uparrow}(8) = \{1, 1, 2, 2, 3, 3, 4, 5, 5, 7\}$$

$$D'où : R_2(8) = 1, R_5(8) = 3$$

Et finalement :

$$\hat{f}_2^{KNN}(8) = \frac{2}{10} \cdot \frac{1}{2 \cdot 1} = 0,1 ; \hat{f}_{25}^{KNN}(8) = \frac{5}{10} \cdot \frac{1}{2 \cdot 3} = 0,083$$

A noter, nous avons vu la définition de l'estimateur par k-plus-proches-voisins pour une variable à valeurs dans  $\mathbb{R}$ . Cette définition peut s'étendre très simplement à des variables multidimensionnelles, à valeurs dans  $\mathbb{R}^d$  :

$$\hat{f}_k^{KNN}(x) = \frac{k}{N} \cdot \frac{1}{V_d \cdot R_k(x)}$$

Où  $V_d$  est le volume d'une sphère d-dimensionnel unitaire.

En particulier :

$$V_2 = \pi, V_3 = \frac{4}{3}\pi$$

## Quels paramètre avons-nous du fixer ?

- ~~La valeur (ou coordonnée)  $m$~~
- ~~Le nombre total d'intervalles  $b$ /la longueur (ou volume) de chaque intervalle  $v$~~
- ~~Le noyau  $K$~~
- Le nombre  $k$  de voisins

Nous allons voir empiriquement l'effet de ces choix à l'aide d'un exemple. Nous allons aussi observer à nouveau combien la taille de l'échantillon va jouer sur notre estimation.

- Télécharger le fichier TP3\_KNN.ipynb : [ibalelli.github.io](https://ibalelli.github.io) → Teaching → Modélisation statistique avancée
- Ouvrir un terminal, aller dans le dossier où vous avez enregistré le fichier → jupyter notebook

## Quelques observations :

- $k$  définit détermine le « lissage » de l'estimation. Si  $k$  est choisi trop petit on obtiens une estimation beaucoup trop bruité.
- En dimension 1,  $\hat{f}_k^{KNN}$  ne définit pas une densité !
- Trouver le  $k$  optimale n'est pas un problème facile à résoudre. Une analyse asymptotique du MSE peut nous permettre de suggérer  $k = CN^{\frac{4}{5}}$  pour une certaine constante  $C$ .

Pour toutes ces raisons, l'estimateur par  $k$ -plus-proches-voisins est rarement utilisé en pratique pour estimer la densité.

Par contre il est très largement utilisé pour des problèmes de classification. L'idée sous-jacente, très intuitive, est celle d'associer à chaque nouvelle donnée  $x$  la classe  $\omega_i$  majoritairement représentée parmi ses  $k$  voisins.

On va tester cela sur un exemple → retour au TP

## CONCLUSIONS

- Nous avons vu trois approches non paramétriques pour estimer la densité d'une distribution à partir d'un échantillon observé.
- Nous avons appris comment utiliser, implémenter et tester ces trois méthodes.
- Nous en avons commenté les avantages et les limites, ainsi que les effets de nos choix en terme de hyperparamètres.

## 6. LA RÉGRESSION NON PARAMÉTRIQUE

# RÉGRESSION NON PARAMÉTRIQUE

Vous avez déjà parlé de régression dans la première partie de ce cours : l'objectif est d'étudier le comportement d'une variable aléatoire, disons  $Y$ , qui est liée à une deuxième variable aléatoire,  $X$ , selon une fonction  $g(X) : Y = g(X)$ .

L'approche paramétrique consiste à avancer des hypothèses à propos de la fonction  $g$ , et ensuite estimer les paramètres inconnus de cette fonction, à partir d'un échantillon observé.

E.g. régression linéaire :

$$Y = \beta X + \varepsilon$$

↑                      ↑  
Paramètre inconnu    Erreur ou résidu



# RÉGRESSION NON PARAMÉTRIQUE

Comme nous l'avons commenté pour l'estimation non paramétrique de la densité, malgré l'approche paramétrique aie des avantages (parcimonie, prédiction, calcul, poids de l'échantillon), les même problèmes liés au choix parfois erronée du modèle se posent ici également, d'où l'intérêt à explorer des méthodes non paramétriques.



# RÉGRESSION NON PARAMÉTRIQUE

Le problème peut être formaliser comme suit :

Soit  $\mathcal{D}_N = \{(x_1, y_1), \dots, (x_N, y_N)\}$  un  $N$ -échantillon, réalisation des variables  $X$  et  $Y$ , nous souhaitons modéliser le comportement de  $X$  comme fonction de  $Y$ , sans avancer des hypothèse sur la forme de la fonction  $g$  :

$$Y = g(X) + \varepsilon$$

$$g(X) = \mathbb{E}(Y|X) \quad \text{Centré, de variance } \sigma^2 : \text{bruit}$$

# 7. RÉGRESSION PAR K-PLUS- PROCHES-VOISINS

## RÉGRESSION NON PARAMÉTRIQUE : K-PLUS-PROCHES-VOISINS

Un des approches plus intuitifs consiste à utiliser les k-plus-proches-voisins pour déterminer la valeur de  $g$  dans un point  $x$  donnée.

- Soit  $k$  le nombre des voisins que l'on souhaite considérer, choisi a priori.
- Pour un  $x$  donnée, soit  $N_k(x) := \{x_i \in \mathcal{D}_N \mid x_i \in \text{Dist}^\uparrow(x)[:k]\}$ , l'ensemble des premiers  $k$  plus proches voisins de  $x$ . Dans la suite on omettra  $x$  de la notation et écrira simplement  $N_k$ .
- Pourriez-vous imaginer comment peut on définir  $\hat{g}_k^{KNN}(x)$  à l'aide de cette notation ?

Idée : pensez à l'exemple de l'utilisation de k-plus-proches-voisins pour la classification, le même type de raisonnement s'applique ici pour la régression.

# RÉGRESSION NON PARAMÉTRIQUE : K-PLUS-PROCHES-VOISINS

$$\hat{g}_k^{KNN}(x) = \frac{1}{k} \sum_{\{i: x_i \in N_k\}} y_i$$

On considère toutes les observations dans notre échantillon d'entraînement qui sont dans la boule de rayon  $k$  centrée sur  $x$

# RÉGRESSION NON PARAMÉTRIQUE : K-PLUS-PROCHES-VOISINS

$$\hat{g}_k^{KNN}(x) = \frac{1}{k} \sum_{\{i: x_i \in N_k\}} y_i$$

Pour chacune des observations  $x_i$  choisies, on va récupérer l'observation  $y_i$  correspondante et on en fait la moyenne

# RÉGRESSION NON PARAMÉTRIQUE : K-PLUS-PROCHES-VOISINS

$$\hat{g}_k^{KNN}(x) = \frac{1}{k} \sum_{\{i: x_i \in N_k\}} y_i$$

On répète ce procédé  
pour tout  $x$  dans  
l'intervalle considéré

$$\hat{g}_k^{KNN}(x) = \frac{1}{k} \sum_{\{i: x_i \in N_k\}} y_i$$

Télécharger le fichier TP5\_Regression\_partiel.ipynb :  
[ibalelli.github.io](https://ibalelli.github.io) → Teaching → Modélisation statistique avancée