I.3. 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 ν), 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)

ESTIMATEUR PAR K-PLUS-PROCHES-VOISINS

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) :

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

2. Ordonner l'ensemble $Dist(x)$ en ordre croissant, $Dist^{\nearrow}(x)$, de telle manière que le premier élément de $Dist^{\nearrow}(x)$ 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 $Dist^{\nearrow}(x)$)

ESTIMATEUR PAR K-PLUS-PROCHES-VOISINS

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 $\mathbb{R}^d$)

ESTIMATEUR PAR K-PLUS-PROCHES-VOISINS

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 $k = 5$

ESTIMATEUR PAR K-PLUS-PROCHES-VOISINS

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.

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

D'où : $\mathrm{R}_2(8) = 1, \mathrm{R}_5(8) = 3$

Et finalement :

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

ESTIMATEUR PAR K-PLUS-PROCHES-VOISINS

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 → Teaching → Modélisation statistique avancée
- Ouvrir un terminal, aller dans le dossier où vous avez enregistré le fichier → jupyter notebook

ESTIMATEUR PAR K-PLUS-PROCHES-VOISINS

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.

ESTIMATEUR PAR K-PLUS-PROCHES-VOISINS

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 ω_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.