Chapitre II
k plus proches voisins
Cycle pluridisciplinaire d'études supérieures
Université Paris sciences et lettres

Nous présentons dans ce chapitre l'algorithme des k plus proches voisins qui se décline aussi bien en classification qu'en régression. Dans le cas de la classification, on peut le présenter informellement de la façon suivante. Soit $n \geqslant 1$ un entier et $S = (x_i, y_i)_{i\in[n]}$ un échantillon d'apprentissage [1]. k est un paramètre à choisir au préalable. Pour $k = 1$, le prédicteur correspondant prédit, pour toute entrée $x \in \mathscr{X}$, la sortie $y_{i(x)}$ où $i(x) \in [n]$ est un indice tel que $x_{i(x)}$ est l'entrée la *plus proche* de x parmi tous les x_i ($i \in [n]$). Lorsque $k \geqslant 2$, la généralisation est naturelle : le prédicteur donne la classe majoritaire [2] parmi les k plus proches voisins de x.

En pratique, l'algorithme est efficace en faible dimension (la dimension étant le nombre de variables explicatives) et pour de petits échantillions d'apprentissage. À l'inverse, l'algorithme requiert un long temps de calcul lorsque l'échantillon d'apprentissage est grand, et est peu performant en grande dimension.

I. cadre

On se donne les éléments suivants.

1. On rappelle que $[n]$ est une notation compacte qui désigne l'ensemble $\{1, 2, \dots, n\}$.
2. c'est-à-dire la valeur qui revient le plus souvent parmi les y_i concernés

— Un ensemble d'entrées $\mathcal{X}$ muni d'une fonction *distance* $\rho : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$. Par exemple, lorsque $\mathcal{X} = \mathbb{R}^d$, on peut considérer la distance euclidienne : $\rho(x, x') = \|x - x'\|_2 = \sqrt{\sum_{j=1}^{d}(x'_j - x_j)^2}$.
— Un ensemble de sorties $\mathcal{Y}$ fini ou égal à $\mathbb{R}$. Par exemple, $\mathcal{Y} = \{0, 1\}$ (ensemble fini) ou encore $\mathcal{Y} = \mathbb{R}$ (convexe).
— Soit $n \geqslant 1$ et $S = (x_i, y_i)_{i\in[n]} \in \mathcal{S}(\mathcal{X}, \mathcal{Y})$ un échantillon d'apprentissage.

2. STATISTIQUES DE RANG

On introduit ci-dessous les statistiques de rang, qui vont permettre de définir formellement et de façon unique les *k plus proches voisins* d'un point $x \in \mathcal{X}$.

Soit $x \in \mathcal{X}$. La *première statistique de rang* est définie par :

$$r_1(x) = \min \operatorname*{Arg\,min}_{i\in[n]} \rho(x, x_i).$$

Autrement dit, parmi les indices $i \in [n]$ qui minimisent la distance $\rho(x, x_i)$, $r_1(x)$ est défini comme étant le plus petit (indice). On pose ensuite $I_1(x) = [n] \setminus \{r_1(x)\}$. Puis par récurrence pour $2 \leqslant k \leqslant n$, on définit la *k-ème statistique de rang* par :

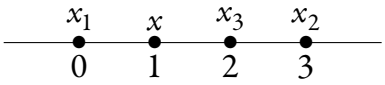
$$r_k(x) = \min \operatorname*{Arg\,min}_{i\in I_{k-1}(x)} \rho(x, x_i),$$

et on pose $I_k(x) = I_{k-1}(x) \setminus \{r_k(x)\}$.

REMARQUE. — Le fait de prendre le minimum de chaque ensemble de minimiseur est une convention arbitraire qui permet de départager les cas où il existe plusieurs minimiseurs.

REMARQUE. — La première statistique de rang $r_1(x)$ est bien définie car n n'étant pas nul, $[n]$ est non vide, donc les valeurs $\rho(x, x_i)$ pour $i \in [n]$ sont en nombre fini et non nul. Il existe donc bien au moins un minimum, et donc l'Argmin est non vide. De même, $r_k(x)$ est bien défini pour tout $2 \leqslant k \leqslant n$ car on voit facilement par récurrence que $I_{k-1}(x)$ est de cardinal $n - k + 1 > 0$, et l'Argmin correspondant est non vide.

EXEMPLE. — On considère $\mathcal{X} = \mathbb{R}$ et pour $x, x' \in \mathbb{R}$, $\rho(x, x') = |x' - x|$. On se donne des entrées $x_1 = 0, x_2 = 3, x_3 = 2$, ainsi que $x = 1$, qu'on peut représenter comme suit.



On détermine alors les trois statistiques de rang.

$$r_1(x) = \min \overbrace{\operatorname*{Arg\,min}_{i\in[3]} \rho(x, x_i)}^{\{1,3\}} = 1 \qquad \mathrm{I}_1(x) = [3] \setminus \{1\} = \{2, 3\}$$

$$r_2(x) = \min \overbrace{\operatorname*{Arg\,min}_{i\in \mathrm{I}_1(x)} \rho(x, x_i)}^{\{3\}} = 3 \qquad \mathrm{I}_2(x) = \mathrm{I}_1(x) \setminus \{3\} = \{2\}$$

$$r_3(x) = \min \overbrace{\operatorname*{Arg\,min}_{i\in \mathrm{I}_2(x)} \rho(x, x_i)}^{\{2\}} = 2$$

3. Découpage de l'espace d'entrées

Pour $1 \leqslant k \leqslant n$, on note $\mathscr{P}_k([n])$ l'ensemble des sous-ensembles de $[n]$ à k éléments. Pour $\mathrm{J} \in \mathscr{P}_k([n])$, on définit :

$$\mathrm{A}_{\mathrm{J}}^{(k)} = \{x \in \mathscr{X} \mid \{r_1(x), \dots, r_k(x)\} = \mathrm{J}\}\,.$$

Autrement dit, $\mathrm{A}_j^{(k)}$ est l'ensemble des $x \in \mathscr{X}$ pour lesquels les k premières statistiques de rang correspondent aux éléments de J.

Exemple. — En reprenant l'exemple précédent, on a pour $k = 1$:

$$\mathrm{A}_{\{1\}}^{(1)} =]-\infty, 1]\,, \quad \mathrm{A}_{\{2\}}^{(1)} = [5/2, +\infty[\,, \quad \mathrm{A}_{\{3\}}^{(1)} =]1, 5/2[\,,$$

et pour $k = 2$:

$$\mathrm{A}_{\{1,2\}}^{(2)} = \varnothing, \quad \mathrm{A}_{\{1,3\}}^{(2)} =]-\infty, 3/2]\,, \quad \mathrm{A}_{\{2,3\}}^{(2)} =]3/2, +\infty[\,.$$

PROPOSITION. — *Pour tout* $1 \leqslant k \leqslant n$ *:*

$$\mathscr{X} = \bigsqcup_{J \in \mathscr{P}_k([n])} A_J^{(k)},$$

où $\sqcup$ *désigne l'union disjointe.*

Autrement dit, pour k donné, les ensembles $A_J^{(k)}$ pour J parcourant $\mathscr{P}_k([n])$ forment une partition [3] de $\mathscr{X}$. La démonstration de la proposition est laissée en exercice.

Pour $x \in \mathscr{X}$, on note $J^{(k)}(x)$ l'unique élément de $\mathscr{P}_k([n])$ tel que $x \in A^{(k)}_{J^{(k)}(x)}$. Autrement dit, $J^{(k)}(x)$ est l'ensemble des indices des k plus proches voisins de x (selon les statistiques de rang).

4. LE PRÉDICTEUR DES k PLUS PROCHES VOISINS (ALIAS kNN)

En *régression* (c'est-à-dire lorsque $\mathscr{Y} = \mathbb{R}$), le régresseur des k plus proches voisins est défini par :

$$\forall x \in \mathscr{X}, \quad \widehat{f}^{(k)}(x) = \frac{1}{k} \sum_{i \in J^{(k)}(x)} y_i.$$

Autrement dit, $\widehat{f}^{(k)}$ prédit pour l'entrée x la moyenne des sorties des k plus proches voisins de x.

En *classification* (c'est-à-dire lorsque $\mathscr{Y}$ est fini), on suppose, quitte à renommer les étiquettes, que $\mathscr{Y} = [N]$ où N est le cardinal de $\mathscr{Y}$. Le classifieur des k plus proches voisins est défini par :

$$\forall x \in \mathscr{X}, \quad \widehat{f}^{(k)}(x) = \min \operatorname*{Arg\,max}_{m \in [N]} \operatorname{Card} \left\{ i \in J^{(k)}(x) \,\middle|\, y_i = m \right\}.$$

Autrement dit, $\widehat{f}^{(k)}(x)$ est le plus petit $m \in [N]$ parmi ceux qui sont majoritaires parmi les k plus proches voisins de x.

REMARQUE. — Le choix du plus petit indice m parmi ceux qui sont majoritaires parmi les valeurs de y_i concernés est une convention arbitraire qui permet de départager les cas où plusieurs valeurs sont majoritaires à égalité.

3. avec un abus de langage car certains des ensembles $A_J^{(k)}$ peuvent être vides tandis qu'une *partition* à proprement parler ne contient pas d'ensemble vide.