Méthode des k plus proches voisins (k-NN)
Approche théorique et mathématique

1 Cadre mathématique

Soit $(\mathcal{X}, d)$ un espace métrique, où :
— $\mathcal{X} \subset \mathbb{R}^p$ est l'espace des observations,
— $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$ est une distance.
On observe un échantillon d'apprentissage i.i.d. :

$$\mathcal{D}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$$

où $X_i \in \mathcal{X}$ et $Y_i \in \mathcal{Y}$, avec :

$$\mathcal{Y} = \begin{cases} \mathbb{R} & \text{(régression)} \\ \{1, \ldots, C\} & \text{(classification)} \end{cases}$$

2 Distances utilisées dans k-NN

2.1 Distance euclidienne

$$d_2(x, x') = \|x - x'\|_2 = \sqrt{\sum_{j=1}^{p} (x_j - x'_j)^2}$$

2.2 Distance de Manhattan

$$d_1(x, x') = \|x - x'\|_1 = \sum_{j=1}^{p} |x_j - x'_j|$$

2.3 Distance de Minkowski

Pour $r \geq 1$:

$$d_r(x, x') = \left(\sum_{j=1}^{p} |x_j - x'_j|^r \right)^{1/r}$$

2.4 Distance de Chebyshev

$$d_\infty(x, x') = \|x - x'\|_\infty = \max_{1 \leq j \leq p} |x_j - x'_j|$$

2.5 Distance de Mahalanobis

Soit Σ une matrice de covariance symétrique définie positive :

$$d_M(x, x') = \sqrt{(x - x')^\top \Sigma^{-1} (x - x')}$$

2.6 Distance cosinus

Pour $x, x' \neq 0$:

$$d_{\cos}(x, x') = 1 - \frac{\langle x, x' \rangle}{\|x\|_2 \|x'\|_2}$$

3 Définition des k plus proches voisins

Pour un point $x \in \mathcal{X}$, on ordonne les observations :

$$d(x, X_{(1)}) \leq d(x, X_{(2)}) \leq \cdots \leq d(x, X_{(n)})$$

Définition 1. *L'ensemble des* k plus proches voisins *de x est :*

$$\mathcal{N}_k(x) = \{X_{(1)}, \ldots, X_{(k)}\}$$

4 k-NN en régression

On cherche à estimer la fonction de régression :

$$m(x) = \mathbb{E}[Y \mid X = x]$$

Définition 2. *L'estimateur k-NN de $m(x)$ est :*

$$\hat{m}_k(x) = \frac{1}{k} \sum_{i=1}^{k} Y_{(i)}(x)$$

Il s'agit d'un estimateur non paramétrique, local et adaptatif.

5 k-NN en classification

Le classifieur de Bayes est défini par :

$$g^*(x) = \arg\max_{c \in \mathcal{Y}} \mathbb{P}(Y = c \mid X = x)$$

Définition 3. *Le classifieur k-NN est :*

$$\hat{g}_k(x) = \arg\max_{c \in \mathcal{Y}} \sum_{i=1}^{k} \mathbb{1}_{\{Y_{(i)}(x) = c\}}$$

Il correspond à une estimation empirique des probabilités conditionnelles :

$$\hat{P}_k(Y = c \mid X = x) = \frac{1}{k} \sum_{i=1}^{k} \mathbb{1}_{\{Y_{(i)}(x) = c\}}$$

6 Interprétation statistique

La méthode k-NN peut être vue comme un estimateur à noyau :

$$\hat{m}_k(x) = \frac{\sum_{i=1}^{n} Y_i \mathbb{1}_{\{d(X_i,x)\leq h(x)\}}}{\sum_{i=1}^{n} \mathbb{1}_{\{d(X_i,x)\leq h(x)\}}}$$

où $h(x)$ est choisi tel que la boule $B(x, h(x))$ contienne exactement k points.

7 Propriétés asymptotiques

Théorème 1 (Consistance de Stone)**.** *Si*

$$k \to \infty \quad et \quad \frac{k}{n} \to 0$$

alors l'estimateur k-NN est consistant :

$$\hat{m}_k(x) \xrightarrow{\mathbb{P}} m(x)$$

et, en classification,

$$\mathbb{P}(\hat{g}_k(X) \neq Y) \to \mathbb{P}(g^*(X) \neq Y)$$

8 Limites théoriques

— Malédiction de la dimension
— Sensibilité au choix de la distance
— Coût computationnel en prédiction