UNIVERSITÉ CÔTE D'AZUR

Cédric Richard

Estimation non-paramétrique de densités

Exercice 1
Soit la base d'apprentissage suivante :

	ω_1			ω_2			
	$\boldsymbol{x}_1$	$\boldsymbol{x}_2$	$\boldsymbol{x}_3$	$\boldsymbol{x}_4$	$\boldsymbol{x}_5$	$\boldsymbol{x}_6$	$\boldsymbol{x}_7$
abscisse	1	0	0	0	2	0	-2
ordonnée	0	1	-1	0	0	-2	0

1. Tracer dans le plan la frontière correspondant à la règle de classification du plus proche voisin.
2. Tracer dans le plan la frontière consistant à associer le point à la classe dont il est le plus proche de la moyenne.

Exercice 2
On considère l'ensemble d'apprentissage dans $\mathbb{R}$ composé des éléments suivants :

$$-10, -8, -6, -5, -4, 0, 1, 1.5, 4, 7, 9$$

On choisit un noyau Gaussien et l'on note h la largeur de la fenêtre.

1. Tracer la densité estimée pour $h = 1$. Donner sa valeur pour $x = 0$.
2. Tracer la densité estimée pour $h = 2$. Donner sa valeur pour $x = 0$.
3. Même question pour un noyau rectangulaire de largeur $h = 1$.

Exercice 3
L'estimation de densité de probabilité par méthode à noyau peut être utilisée pour l'élaboration d'un classificateur de Bayes. On dispose de N données d'apprentissage, dont N_i pour chaque classe ω_i en compétition.

1. Représenter la fonction de densité de probabilité conditionnelle $p(x|\omega_i)$ par un modèle à noyau $\hat{p}(x|\omega_i)$.
2. Estimer la probabilité a priori $P(\omega_i)$.
3. Écrire la règle de décision de Bayes à partir des grandeurs estimées précédemment.

Exercice 4
Soit une variable aléatoire de fonction de densité de probabilité normale $p(x)$ de moyenne m et de variance σ^2. On souhaite estimer $p(x)$ à l'aide d'un estimateur à noyau sur un échantillon $x_1, \ldots, x_n$. La fonctions noyau choisie est une fonction de Gauss de moyenne nulle et de variance unité notée $K(u)$.

$$\hat{p}_n(x) = \frac{1}{nh} \sum_{k=1}^{n} K\left(\frac{x - x_k}{h}\right)$$

Montrer que $\hat{p}_n(x)$ est une fonction Gaussienne de moyenne m et de variance $\sigma^2 + h^2$. Commenter les caractéristiques de l'estimateur $\hat{p}_N(x)$ en fonction du choix de la fenêtre.
Indication : le produit de convolution de deux fonctions de Gauss est une fonction de Gauss.

Exercice 5
Soit $p(x)$ la fonction densité de probabilité exponentielle de paramètre λ définie par :

$$p(x) = \begin{cases} \lambda \exp(-\lambda x) & \text{si } x > 0 \\ 0 & \text{sinon.} \end{cases}$$

On utilise la fenêtre suivante pour estimer $p(x)$:

$$\phi(x) = \begin{cases} \exp(-x) & \text{si } x > 0 \\ 0 & \text{sinon.} \end{cases}$$

Calculer la moyenne de la distribution estimée.

Exercice 6
On considère l'estimateur de densité à noyau

$$\hat{p}_n(x) = \frac{1}{nh_n} \sum_{k=1}^{n} K\left(\frac{x - x_k}{h}\right)$$

où $K(u)$ est un noyau positif, symétrique, de volume unité, et $\{x_1, \ldots, x_n\}$ un échantillon aléatoire indépendant et identiquement distribué de densité de probabilité $p(x)$.

1. Démontrer que $\hat{p}_n(x)$ est une densité de probabilité.

2. En utilisant le théorème proposé en fin d'exercice, démontrer le théorème suivant :
   Soit $p(x)$ continue en x, et h_n une suite de réels positifs telle que $h_n \to 0$ et $nh_n \to \infty$ lorsque $n \to \infty$. Alors $\hat{p}_n(x) \xrightarrow{p} p(x)$.

3. Démontrer le théorème suivant :
   Soit $p(x)$ trois fois différentiable et de dérivée troisième bornée au voisinage de x. Soit $K(u)$ un noyau tel que $\int |u|^3 K(u)\, du < \infty$. Si $\lim_{n\to\infty} h_n = 0$, alors :

   $$E\left[\hat{p}_n(x) - p(x)\right] = \frac{1}{2} h_n^2 p''(x) \tau^2 + o(h_n^2)$$

   avec $\tau^2 = \int u^2 K(u) du$. De plus, si $\lim_{n\to\infty} nh_n = \infty$, alors :

   $$\operatorname{var}\left[\hat{p}_n(x)\right] = \frac{1}{nh_n} p(x) \int K^2(u)\, du + o\left(\frac{1}{nh_n}\right)$$

   En déduire que :

   $$\begin{aligned} \operatorname{MSE}\left[\hat{p}_n(x)\right] &= E\left[\hat{p}_n(x) - p(x)\right]^2 \\ &= \frac{1}{4} h_n^4 [p''(x)]^2 \tau^4 + \frac{1}{nh_n} p(x) \int K^2(u)\, du + o\left(h_n^4 + \frac{1}{nh_n}\right) \end{aligned}$$

4. Calculer la largeur de bande optimum pour le noyau Gaussien au sens de la MSE à partir du résultat précédent.

5. Générer des données selon une loi normale et exponentielle. Implanter un estimateur à noyau pour ces données. Essayer différentes largeurs de bande.

Indication pour la question 2 (Théorème 1A in E. Parzen, 1962) :
Soit w une fonction bornée et intégrable satisfaisant $\lim_{u\to\infty} |u \cdot w(u)| = 0$. Soit g une fonction intégrable. Alors, pour toute suite positive h_n telle que $h_n \to 0$ pour $n \to \infty$, on a

$$\lim_{n\to\infty} \frac{1}{h_n} \int w\left(\frac{u - x}{h_n}\right) g(u)\, du = g(x) \int w(u)\, du$$

en tout point x où g est continue.

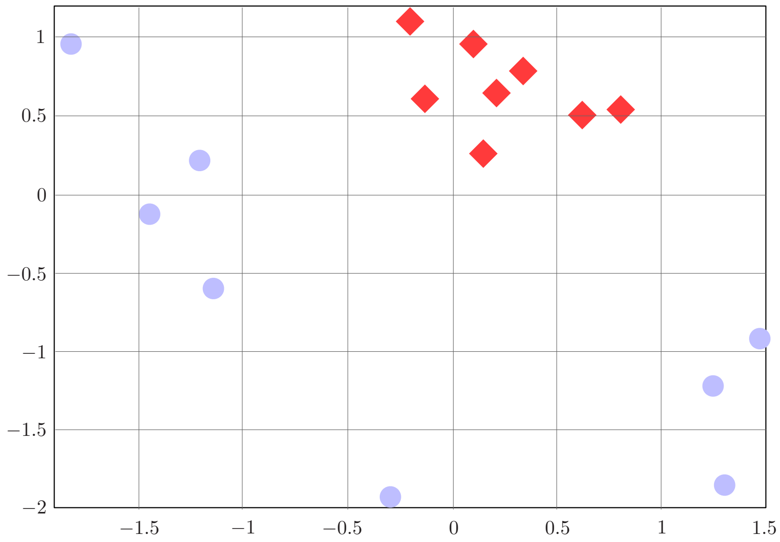

Figure 1: Nuage de points

Exercice 7
L'algorithme des k-plus-proches-voisins est un algorithme de classification intuitif, aisément paramétrable et assez performant. On considère un problème de classification à K classes. On dispose de N données étiquetées $\mathcal{A}_N = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^N$ avec $\boldsymbol{x}_i \in \mathbb{R}^D$ et $y_i \in \{1, \ldots, K\}$.

1. Donner l'algorithme des k-plus-proches-voisins sous forme de pseudo-code. Discuter du coût machine des différentes étapes. Dans le cas $k = 1$, expliquer en quoi consiste l'algorithme.

2. Dans le cas de $K = 2$ classes, proposer quelques valeurs possibles pour k. Définir la valeur de k à éviter.

3. Dans l'exemple de la Fig. 1, identifier N, D et K. Tracer approximativement la frontière de décision pour $k = 1$ sur cette figure. Expliquer en quoi consiste cette frontière.

On s'intéresse à présent à l'influence du paramètre k. On considère un point aberrant à classer en $(-\frac{1}{2}, \frac{1}{2})$.

4. Refaire le tracé de frontière. Énoncer la principale faiblesse de l'algorithme pour $k = 1$.

5. Tracer approximativement la frontière pour $k = 3$. Énoncer l'intérêt de ce paramétrage par rapport au précédent.

6. Donner la décision de l'algorithme pour $k = N$.

Afin d'améliorer la fiabilité des algorithmes de classification, on utilise parfois une technique de rejet où l'algorithme ne prend pas de décision si celle-ci n'est pas suffisamment sûre.

7. Proposer une heuristique de rejet en distance.

8. Proposer une heuristique de rejet en ambiguïté.