Contrôle de synthèse du 23 Mai 2023

Durée 1h30-Notes de cours et calculatrices autorisées
Dans toute l'épreuve, la distance utilisée est la distance euclidienne classique

1 Régression linéaire simple

On considère le modèle linéaire

$$Y_i = \theta_0^* + \theta_1^* x_i + \varepsilon_i,\ i = 1, \cdots, 4,$$

où $x_1 = x_3 = 1$ et $x_2 = x_4 = -1$ et les variables aléatoires $\varepsilon_i,\ i = 1, \cdots, 4$ sont i.i.d. de loi $\mathcal{N}(0, \sigma_*^2)$.

1. Montrer, en utilisant les résultats du cours, que les estimateurs du maximum de vraisemblance sont :

$$\begin{aligned}
\widehat{\theta}_0 &= \overline{Y} = \frac{1}{4}\left(Y_1 + Y_3 + Y_2 + Y_4\right), \\
\widehat{\theta}_1 &= \frac{1}{4}\left(Y_1 + Y_3 - Y_2 - Y_4\right), \\
S^2 &= \sum_{i=1}^{4}\left(Y_i - \widehat{\theta}_0 - \widehat{\theta}_1 x_i\right)^2 = \frac{1}{4}\left(Y_1^2 + Y_3^2 + Y_2^2 + Y_4^2 - \frac{1}{4}\left[Y_1 + Y_3 + Y_2 + Y_4\right]^2 - \frac{1}{4}\left[Y_1 + Y_3 - Y_2 - Y_4\right]^2\right).
\end{aligned}$$

2. On a observé la réalisation des variables Y_1, Y_2, Y_3, Y_4 suivantes : $y_1 = 1$, $y_3 = 1.6$, $y_2 = -1$, $y_4 = -1.2$. Calculer les estimations des paramètres θ_0^*, θ_1^* et σ_*^2.
3. On souhaite prédire la valeur de Y si $x = 0$. Proposer et calculer un prédicteur.

2 Quantification

Soit U une variable aléatoire de loi uniforme sur $[0, 1]$. L'objectif de cet exercice et de montrer l'assertion : *Pour $n \in \mathbb{N}^*$, la quantification optimale à n points de U est la loi uniforme sur l'ensemble Ω_n défini par,*

$$\Omega_n := \left\{\frac{2i-1}{2n},\ i = 1, \cdots, n\right\}.$$

1. Vérifier que l'assertion est vraie pour $n = 1, 2, 3$.
2. Montrer que l'assertion est vraie pour tout $n \in \mathbb{N}^*$.
3. On se place dans le cas $n = 2$. Montrer que si l'on initialise les centres aux valeurs 0, 1, l'algorithme de Loyd converge en seulement une itération.

3 Plus proche voisin

On considère un problème de classification supervisée binaire (label L valant 0 ou 1). La covariable X utilisée est un point du carré $ABCD$ centré en O et de côté 2. L'échantillon d'apprentissage constitué de 5 observations est $(O, 0)$, $(A, 1)$, $(B, 1)$, $(C, 1)$, $(D, 1)$. Pour prédire le label $L(X)$ associé à un point du carré $ABCD$ on va utiliser la méthode du plus proche voisin. C'est-à-dire que le prédicteur $\widehat{L}(X)$ est le label du point de l'échantillon d'apprentissage le plus proche de X.

1. On appelle A' (resp. B', C', D') le milieu du segment AB (resp., BC, CD, DA). Faire une figure.
2. On rappelle que les cellules de Voronoï entre deux points du plan sont séparées par la médiatrice (droite perpendiculaire au segment passant par son milieu). Montrer que la cellule de Voronoï associée à O pour la collection de points $\Omega := \{O, A, B, C, D\}$ est le carré $A'B'C'D'$. On considère un point X à l'intérieur de ce carré. Que vaut $\widehat{L}(X)$?
3. Dessiner, dans le carré $ABCD$, la tesselation de Voronoï associée à Ω.
4. Représenter graphiquement en bicolore sur le carré $ABCD$ le prédicteur $\widehat{L}$.

4 Last but not least

On considère l'espace $\mathcal{X} = \mathbb{R}^3$ (trois variables quantitatives). On considère $n = 4$ individus définis par

$$\begin{pmatrix} x^{(1)} \\ x^{(2)} \\ x^{(3)} \\ x^{(4)} \end{pmatrix} = \begin{pmatrix} 2 & -1 & 1 \\ -1 & 1 & 0 \\ 1 & -1 & 3 \\ 2 & 1 & 3 \end{pmatrix}.$$

On considère une problème de régression avec les observations associées

$$\begin{pmatrix} y^{(1)} \\ y^{(2)} \\ y^{(3)} \\ y^{(4)} \end{pmatrix} = \begin{pmatrix} 2 \\ 3 \\ -3 \\ 0 \end{pmatrix}.$$

Pour $x = (2, 0, 2)$, calculer $\hat{y}(x)$, le prédicteur de y par 2-plus proches voisins.