Introduction au machine learning
Exercices

François Bachoc
Université Paul Sabatier

Exercice 1

Dans une étude clinique, on observe 27 patients dont on enregistre le poids, l'age, le genre (homme/femme) et la présence de chirurgie dans leur historique médical (oui/non). Puis on mesure leur score à un test physique (noté continument entre 0 et 20, 20 désignant une performance maximale). L'objectif de l'apprentissage est de prédire le score en fonction du poids, de l'age, du genre et de la présence de chirurgie.

1) Formaliser cette étude comme un problème d'apprentissage dans le cadre du cours. Indiquer d, n, la nature des variables (quantitatives ou qualitatives) et s'il s'agit d'un problème de régression ou de classification. Expliciter aussi l'espace d'entrée $\mathcal{X}$.

2) On considère deux patients. Le premier est une femme de 40 ans pesant 60kg et n'ayant pas d'historique de chirurgie. Le second est un homme de 30 ans pesant 80kg et n'ayant pas d'historique de chirurgie. Calculer la distance D entre ces deux patients.

Exercice 2

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

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

Calculer la matrice M de taille 4×4 définie par, pour $i, j \in \{1, 2, 3, 4\}$, $M_{i,j} = D(x^{(i)}, x^{(j)})$.

Exercice 3

On considère l'espace $\mathcal{X} = \mathbb{R}^2 \times \{1, 2, 3\}$ (deux variables quantitatives et une variable qualitative). On considère $n = 4$ individus définis par

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

Calculer la matrice M de taille 4×4 définie par, pour $i, j \in \{1, 2, 3, 4\}$, $M_{i,j} = D(x^{(i)}, x^{(j)})$.

Exercice 4

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

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

On considère une problème de régression avec les observations associées

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{pmatrix} = \begin{pmatrix} -2 \\ 3 \\ 4 \\ -4 \end{pmatrix}.$$

Pour $x = (1, 0, 1)$, calculer $\hat{y}_{\text{ppv},2,4}(x)$, le prédicteur de y par 2-plus proches voisins.

Exercice 5

On considère l'espace $\mathcal{X} = \mathbb{R} \times \{1, 2\} \times \{1, 2, 3\}$ (une variable quantitative et deux variables qualitatives). On considère $n = 8$ individus définis par

$$\begin{pmatrix} x^{(1)} \\ x^{(2)} \\ x^{(3)} \\ x^{(4)} \\ x^{(5)} \\ x^{(6)} \\ x^{(7)} \\ x^{(8)} \end{pmatrix} = \begin{pmatrix} -1 & 1 & 1 \\ -2 & 1 & 1 \\ 0 & 1 & 2 \\ 2 & 1 & 2 \\ 3 & 2 & 3 \\ -1 & 2 & 3 \\ 4 & 2 & 3 \\ 2 & 2 & 3 \end{pmatrix}.$$

On considère un problème de classification avec $N = 2$ et avec les observations associées

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \\ y_7 \\ y_8 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 2 \\ 2 \\ 1 \\ 2 \\ 2 \\ 1 \end{pmatrix}.$$

Pour $x = (1, 2, 3)$, calculer $\hat{y}_{\text{ppv},3,8}(x)$, le classifieur de y par 3-plus proches voisins.

Exercice 6

On considère un problème de régression avec $d = 1$ variable quantitative et avec $x^{(1)} = 0, x^{(2)} = 1/4, x^{(3)} = 1/2, x^{(4)} = 3/4, x^{(5)} = 1$. Pour $i \in \{1, \dots, 5\}$ on a $y_i = 1 + (x^{(i)})^2$. Tracer la courbe de la fonction

$$\begin{aligned} \hat{y}_{\text{ppv},2,5} \colon [0, 1] &\to \mathbb{R} \\ x &\mapsto \hat{y}_{\text{ppv},2,5}(x). \end{aligned}$$

Exercice 7

On considère un problème de classification avec $d = 2$ variable quantitatives, $N = 2$ classes et avec $x^{(1)} = (0, 0), x^{(2)} = (1, 0), x^{(3)} = (0, 1), x^{(4)} = (1/2, 1/2), x^{(5)} = (1, 1)$. On a $y_1 = 1, y_2 = 1, y_3 = 1, y_4 = 2, y_5 = 2$. Tracer dans le plan l'ensemble

$$\{x \in [0, 1]^2, \hat{y}_{\text{ppv},1,5}(x) = 1\}.$$

Exercice 8

On considère X_1, X_2, X uniformément distribués sur $[0, 1]$ et indépendants. On pose $Y_1 = X_1 + \epsilon_1, Y_2 = X_2 + \epsilon_2, Y = X + \epsilon$, avec $\epsilon_1, \epsilon_2, \epsilon$ independants de X_1, X_2, X et identiquement distribués de loi $\mathcal{N}(0, 1)$.

a) Prouver soigneusement que, pour tout $x, x^{(1)}, x^{(2)} \in [0, 1]$,

$$\mathbb{E}\left(\left(Y - \hat{Y}_{\text{ppv},1,2}(X) \right)^2 \middle| X = x, X^{(1)} = x^{(1)}, X^{(2)} = x^{(2)} \right) = \left(x - x^{(I_{\text{ppv},1}(x))} \right)^2 + 2.$$

b) Montrer que pour $0 \leq t \leq 1/2$,

$$\mathbb{P}\left(\left| X - X^{(I_{\text{ppv},1}(x))} \right| \geq t \middle| X = \frac{1}{2} \right) = (1 - 2t)^2.$$

Exercice 9

On considère un problème de régression avec d variables quantitatives et les entrées aléatoires $X^{(1)}, \dots, X^{(n)}, X$, indépendantes et presque surement contenues dans $[0,1]^d$. On définit $Y = \sum_{a=1}^{d} X_a + \epsilon$ et, pour $i = 1, \dots, n$, $Y_i = \sum_{a=1}^{d} X_a^{(i)} + \epsilon_i$ avec $\epsilon, \epsilon_1, \dots, \epsilon_n$ indépendants, de loi $\mathcal{N}(0,1)$, et indépendants de $X, X^{(1)}, \dots, X^{(n)}$.

On admet que pour $u \in [0,1]^d$, $\mathbb{E}(Y|X=u) = \sum_{a=1}^{d} u_a$ et $\mathrm{Var}(Y|X=u) = 1$.

a) Prouver que pour $u, v \in [0,1]^d$,

$$|\mathbb{E}(Y|X=u) - \mathbb{E}(Y|X=v)| \leq \sqrt{d}||u-v||.$$

On pourra éventuellement utiliser Cauchy-Schwarz.

b) Montrer que pour $p \in \{1, \dots, n\}$,

$$\mathbb{E}\left(\left(Y - \hat{Y}_{\mathrm{ppv},p,n}(X)\right)^2\right) \leq \left(1 + \frac{1}{p}\right) + d\mathbb{E}\left(D\left(X - X^{(I_{\mathrm{ppv},p}(x))}\right)^2\right).$$

On admettra que l'inégalité finale de la section III du cours est vraie dans le cadre de l'exercice.

Exercice 10

On considère l'espace $\mathcal{X} = \mathbb{R} \times \{1,2\} \times \{1,2,3\}$ (une variable quantitative et deux variables qualitatives). On considère $n = 4$ individus définis par

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

On considère un problème de classification avec $N = 2$ et avec les observations associées

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 2 \\ 2 \end{pmatrix}.$$

On considère l'ensemble d'entrainement $E = \{1,2\}$ et l'ensemble de validation $V = \{3,4\}$. On considère la classification par plus proches voisins avec $p = 1$. Calculer $\mathrm{TE}_{E,V}$.

Exercice 11

On considère l'espace $\mathcal{X} = \mathbb{R}^2$ (deux variables quantitatives). On considère $n = 6$ individus définis par

$$\begin{pmatrix} x^{(1)} \\ x^{(2)} \\ x^{(3)} \\ x^{(4)} \\ x^{(5)} \\ x^{(6)} \end{pmatrix} = \begin{pmatrix} 1 & 2 \\ 0 & 2 \\ 1 & 1 \\ 2 & 2 \\ 0 & 0 \\ 2 & -2 \end{pmatrix}.$$

On considère un problème de régression avec les observations associées

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \\ 2 \\ 1 \\ 2 \\ 0 \end{pmatrix}.$$

On considère la validation croisée k-fold avec $k = 3$ et avec $V_1 = \{1,2\}$, $V_2 = \{3,6\}$ et $V_3 = \{4,5\}$. On considère la régression par plus proches voisins avec $p = 1$. Calculer $\mathrm{EQM}_{V_1,V_2,V_3}$.