Université de Lille 2018-2019
M1 Maths/Finances - ISN 27 juin 2019

Examen de Data Mining, 2ème session

Durée : 2h. Documents, calculatrices et téléphones portables sont interdits.

Question de cours : Pour quels problèmes d'apprentissage peut-on utiliser l'algorithme K-means ?

Exercice 1. On considère de modèle génératif suivant sur $\mathbb{R}^2 \times \{0,1\}$. La loi de $\boldsymbol{Y}$ est donnée par

$$\mathbb{P}(\boldsymbol{Y} = 1) = p \quad \text{et} \quad \mathbb{P}(\boldsymbol{Y} = 0) = 1 - p$$

et les loi conditionnelles de $\boldsymbol{X}$ sachant $(\boldsymbol{Y} = 1)$ et $(\boldsymbol{Y} = 0)$ sont telles que

$$\boldsymbol{X}|(\boldsymbol{Y} = 1) \sim \mathcal{N}(\mu_1, I_2) \quad \text{et} \quad \boldsymbol{X}|(\boldsymbol{Y} = 0) \sim \mathcal{N}(\mu_0, I_2),$$

avec $p \in [0,1]$, et $\mu_0 \in \mathbb{R}^2$ et μ_1^2 deux vecteurs. $\mathcal{N}(\mu, I_2)$ désigne une loi normale de moyenne $\mu \in \mathbb{R}^2$ et de matrice de covariance I_2, dont la densité sur $\mathbb{R}^2$ est $f(x) = \frac{1}{2\pi} \exp\left(-\frac{(x-\mu)^T(x-\mu)}{2}\right)$, avec y^T qui désigne la transposée du vecteur y de $\mathbb{R}^2$.

1. Donner l'expression la plus explicite possible du classifieur de Bayes sous ce modèle.
2. Pour $p = 1/2$, $\mu_0 = (1,1)$ et $\mu_1 = (-1,-1)$, représenter graphiquement un jeu de données possible de $n = 10$ observations indépendantes tirées sous le modèle génératif.
   Ajouter le classifieur de Bayes sur ce graphique.
3. Comment utiliser la connaissance du classifieur de Bayes pour proposer un classifieur $\hat{g}_n$ à partir d'une base de données $\mathcal{D}_n = \{(X_i, Y_i)\}_{1 \le i \le n}$ dont les observations sont tirées de manière indépendantes sous un modèle génératif de la forme ci-dessus ?

Exercice 2. Après des élections aux Etats-Unis, une base de données qui contenant le vote d'un certain nombre d'individus ainsi que des éléments démographiques, a été collectée. Une (petite) partie de celle-ci est donnée ci-dessous. L'objectif est de s'en servir pour prédire pour quel parti un individu va voter en fonction de ses caractéristiques démographiques.

Individu	Age	Sexe	Profession	Revenus (k$)	Etudes (an)	Etat	Vote
1	25	M	5	30	5	New-York	Democrate
2	37	M	3	28	3	Californie	Républicatin
3	58	M	2	40	5	Washington	Démocrate
4	30	F	3	35	8	Illinois	Démocrate
5	18	F	6	20	2	Floride	Républicain
6	45	M	1	25	0	Nebraska	Républicain

1. De quelle tâche d'apprentissage s'agit-il ?
2. Pour chacune des variables explicatives, donner son type (qualitative, quantitative).

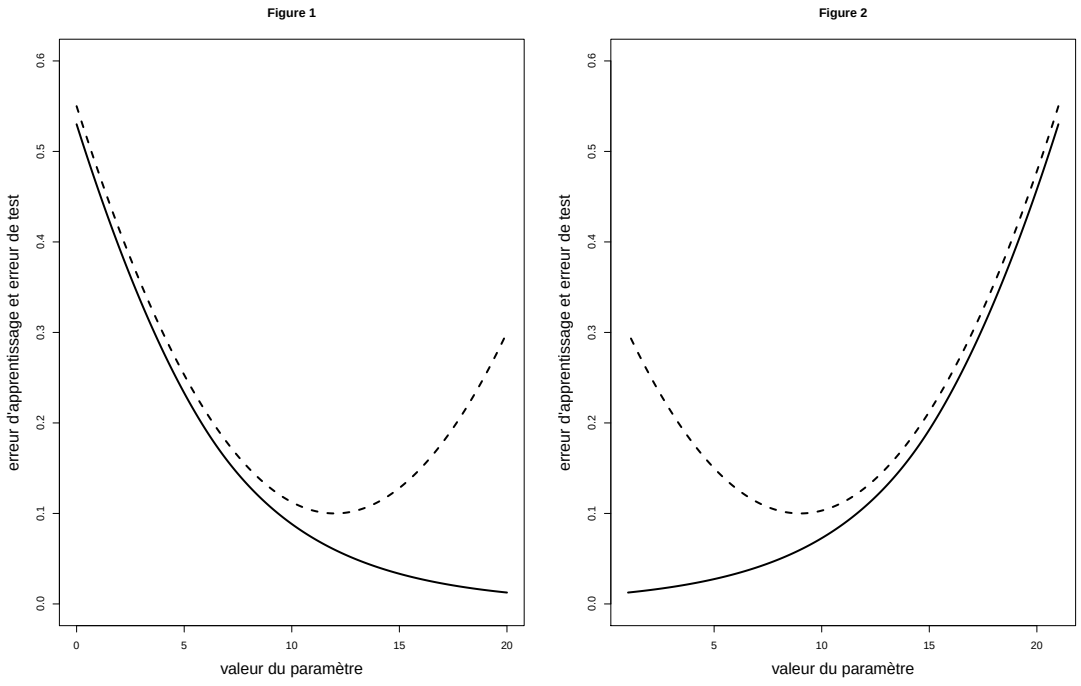
3. Comment peut-on transformer la base de données pour que les variables explicatives soient représentées par un vecteur dans $\mathbb{R}^d$?
4. Si cette dimension d est trop grande, quelle méthode peut-on utiliser pour réduire la dimension ? Pourquoi cherche-t-on à faire cela ?

On considère maintenant plus généralement un problème de classification binaire où $\mathcal{D}_n = \{(X_i, Y_i)\}_{1 \leq i \leq n}$ avec $X_i \in \mathbb{R}^d$ et $Y_i \in \{-1, 1\}$. On construit le prédicteur $\hat{g}_n(x) = \operatorname{sgn}\left(\langle \hat{w}_n, x \rangle + \hat{b}_n\right)$ où

$$(\hat{w}_n, \hat{b}_n) \in \underset{\substack{w,b:\forall i,\xi_i \geq 0 \\ \forall i, Y_i(\langle w, X_i \rangle + b) \geq 1 - \xi_i}}{\operatorname{argmin}} \frac{1}{2}\|w\|^2 + C \sum_{i=1}^{n} \xi_i \tag{1}$$

5. Quel est ce prédicteur ?
6. Donner les deux propriétés importantes de la solution du problème d'optimisation (1) qui permettent notamment d'étendre cette méthode via l'introduction d'un noyau.

Les deux figures ci-dessous représentent l'erreur d'apprentissage et l'erreur de test d'un algorithme d'apprentissage en fonction d'un paramètre dont il dépend.



7. Entre la courbe pleine et la courbe en pointillée, laquelle correspond à l'erreur de test et laquelle correspond à l'erreur d'apprentissage ? Justifier.
8. Identifier quel titre pourrait correspondre à la Figure 1 et à la Figure 2 :
   (a) erreurs de l'algorithme des k-plus proches voisins en fonction du paramètre k
   (b) erreurs de l'algorithme CART en fonction de paramètre `max_depth`
9. Comment sélectionner une bonne valeur du paramètre d'un de ces algorithmes ?
10. Pour un SVM de noyau fixé et de paramètre de coût C, tracer une allure possible de l'erreur de test et d'apprentissage en fonction du paramètre C. Justifier.

Exercice 3. On considère un problème de classification binaire avec $\mathcal{Y} = \{0, 1\}$, pour lequel on cherche à construire un classifieur $f : \mathcal{X} \to \mathcal{Y}$.
Pour les questions 1 à 3, on pose $\mathcal{X} = [0, 1]$ et on se donne la base d'apprentissage

$$(X_1 = 0.1, Y_1 = 0)\ \ (X_2 = 0.3, Y_2 = 1)\ \ (X_3 = 0.6, Y_3 = 0)\ \ (X_4 = 0.7, Y_4 = 1)\ \ (X_5 = 0.8, Y_5 = 1)$$

1. Expliquez comment le classifieur des k-plus proches voisins est défini.
2. Représentez la base d'apprentissage ci-dessus et calculez le classifieur des 3-plus proches voisins.
3. Que vaut le classifieur des 5-plus proches voisins ?

Dans toute la suite, on pose $\mathcal{X} = [0, 5] \times [0, 5]$ et on se donne la base d'apprentissage suivante, où chaque ligne correspond à une observation $X_i \in \mathbb{R}^2$ avec $X_i = (X_i^1, X_i^2)$.

On cherchera à appliquer l'algorithme CART.

	X^1	X^2	Y
X_1	1	4	1
X_2	5	1	0
X_3	4	5	0
X_4	2	1	1
X_5	4	2	1
X_6	3	4	0

4. Représentez graphiquement ce jeu de données étiqueté.
   Combien y-a-t-il de séparations possible sur chaque coordonnée ?
5. Proposez un critère d'impureté que peut utiliser l'algorithme CART. Evaluez chaque séparation possible pour ce critère, et en déduire la première séparation effectuée par l'algorithme.
6. Donnez l'arbre de décision complet retourné par l'algorithme.
7. De manière générale, est-il pertinent d'utiliser l'arbre complet comme classifieur ?