Université de Lille 2017-2018
M1 Maths/Finances - Data Mining 26 juin 2018

Examen de Data Mining (2ème session)

Durée : 2 heures. Documents et calculatrices sont interdits.

Question de cours. Donner un exemple de noyau qui peut être utilisé avec la méthode SVM pour de la classification binaire avec $\mathcal{X} = \mathbb{R}^d$ et $\mathcal{Y} = \{-1, 1\}$.

Exercice 1. On s'intéresse à un problème de classification binaire avec $\mathcal{X} = [0, 1]$ et $\mathcal{Y} = \{-1, +1\}$. On cherche à construire un classifieur à partir de la base de données

$$(X_1 = 0.2, Y_1 = -1)\ (X_2 = 0.3, Y_2 = -1)\ (X_3 = 0.6, Y_3 = 1)\ (X_4 = 0.7, Y_4 = -1)\ (X_5 = 0.8, Y_5 = 1)$$

1. Représenter graphiqumement ces données puis donner l'expression du classifieur obtenu par la méthode des k plus proches voisins pour $k = 1, 3$ et 5.
2. Comment peut-on procéder pour sélectionner une bonne valeur de k ?

On suppose maintenant que ces données ont été générées selon le modèle suivant, pour une certaine valeur $p \in]0, 1[$:

$$\begin{cases} \mathbb{P}(Y = 1) = p \text{ et } \mathbb{P}(Y = -1) = 1 - p \\ X|(Y = -1) \sim \mathcal{U}([0, 3/4]) \\ X|(Y = 1) \sim \mathcal{U}([1/2, 1]) \end{cases}$$

où $\mathcal{U}([a, b])$ désigne la loi uniforme sur le segment $[a, b]$.

3. Donner l'expression du classifieur de Bayes en fonction de p.
4. En se servant de cette expression, proposer alors un classifieur $\hat{f}_n$ construit à partir de la base de données $\mathcal{D}_n$.

Exercice 2. Un vendeur d'articles pour enfants souhaite répartir ses clients en différents groupes afin d'appliquer des stratégies de marketing différentes pour chaque groupe. Pour chaque client, il dispose des informations suivantes : l'âge de son enfant et le montant moyen dépensé chaque mois dans le magasin. Ces données sont présentées dans le tableau suivant.

âge de l'enfant	2	2	3	**3**	3	3	4	4	6	6	7	7	**8**	8	9	10
montant mensuel	50	150	25	**75**	125	175	50	150	100	175	75	125	**100**	150	200	75

1. Décrivez le problème d'apprentissage auquel le vendeur fait face.
2. Représenter ces données graphiquement. Le vendeur décide d'appliquer l'algorithme K-means avec $K = 2$. Représentez les centroïdes et les groupes obtenus après la première itération de cet algorithme en prenant pour centroïdes initiaux les deux points en gras dans le tableau de données.
3. Combien d'étapes sont nécessaires à la convergence de l'algorithme ?
4. La valeur $K = 2$ vous paraît-elle bien choisie ?
5. Quelle(s) autre(s) méthode(s) le vendeur pourrait-il utiliser pour grouper les données ?

Exercice 3. On se réfèrera à l'annexe pour les figures à compléter. On s'intéresse à de la classification binaire de $\mathbb{R}^2$ dans $\{0, 1\}$, où les labels $\times$ correspondent aux 1 et les $\bigcirc$ aux 0.

1. Ajouter sur la figure A un séparateur linéaire qui vous paraît bien classifier les données, et donner son erreur d'apprentissage.
2. Donner l'expression générale d'un séparateur linéaire $\hat{g}_n : \mathbb{R}^2 \to \{0, 1\}$. Quelles sont les méthodes vues en cours qui construisent de tels classifieurs ?
3. Représenter graphiquement sur la figure B un classifieur qui a une erreur d'apprentissage égale à zéro. Est-il un meilleur classifieur que le précédent ?
4. Parmi les algorithmes que vous connaissez, lesquels peuvent avoir de bonnes performances pour les données de la Figure C en annexe ? Ajouter une frontière de décision possible sur cette figure.

Exercice 4. La figure D en annexe représente l'erreur de test et l'erreur d'apprentissage pour l'algorithme CART en fonction d'un paramètre d'élagage, c'est-à-dire le paramètre k ou le paramètre best utilisé dans le package R `tree`.

1. Laquelle des deux courbes correspond à l'erreur de test ? Justifier.
2. Parmi k et best, quel est le paramètre qui figure en abscisses ? Justifier.
3. Commenter ce qu'on observe sur la figure.

Exercice 5. On considère la base d'apprentissage $\mathcal{D}_n$ suivante où $n = 8$, $X_i \in \mathbb{R}^2$ (deux variables explicatives) et $Y_i \in \{0, 1\}$.

	X^1	X^2	Y		X^1	X^2	Y
X_1	1	1	1	X_5	2	2	1
X_2	1	2	1	X_6	3	1	0
X_3	1	3	0	X_7	3	2	0
X_4	2	1	1	X_8	3	4	0

1. Représenter graphiquement ces données.
2. Proposer un critère d'impureté utilisable dans l'algorithme CART et rappeler comment le calculer.
3. A la première étape de l'algorithme CART, combien de séparations sont possibles sur chaque coordonnée ? Pour chacune d'elle, calculer la valeur du critère d'impureté choisi après séparation, et donner la première séparation construite par l'algorithme.
4. Construire l'arbre de décision complet.
5. Donner l'expression du classifieur associé $\hat{f}_n : \mathbb{R}^2 \to \{0, 1\}$.

Annexe à rendre avec la copie

Numéro de Candidat :

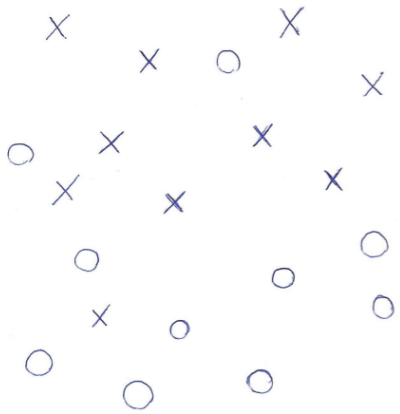

FIGURE A – données d'apprentissage pour un problème de classification binaire

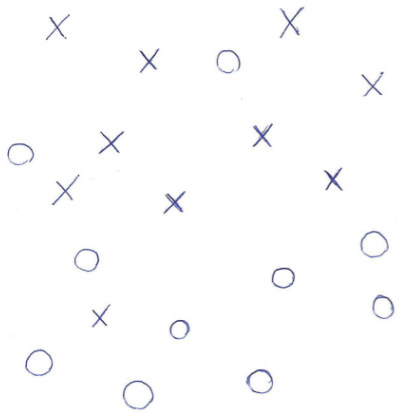

FIGURE B – données d'apprentissage pour un problème de classification binaire

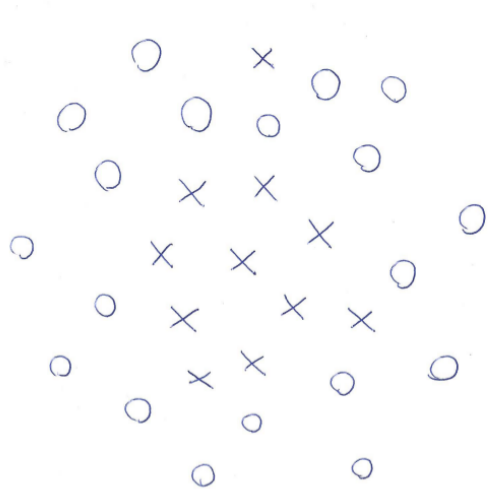

FIGURE C – données d'apprentissage pour un problème de classification binaire

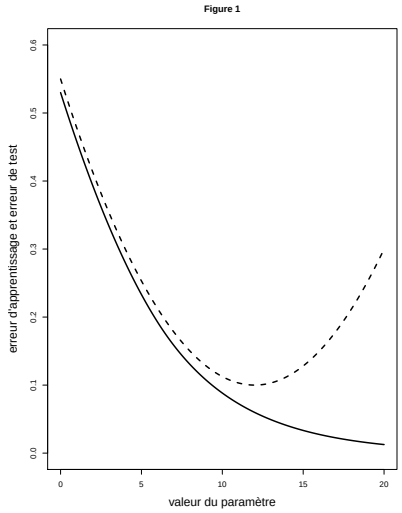

FIGURE D – erreurs de test et d'apprentissage en fonction d'un paramètre d'élagage