Université de Lille 1 2016-2017
M1 Maths/Finances - Data Mining 21 juin 2017

# Examen de Data Mining - Deuxième session

*Durée : 2 heures. Documents et calculatrices sont interdits.*

**Notations.** On rappelle que le produit scalaire entre deux vecteurs de $\mathbb{R}^d$ est $\langle x|y\rangle = x^T y = y^T x = \sum_{i=1}^d x_i y_i$, où $x^T$ désigne la transposée du vecteur $x \in \mathbb{R}^d$. La norme au carré est $\|x\|^2 = \langle x|x\rangle$.

Pour toute partie $\mathcal{S}$ de $\mathbb{R}$ on rappelle que $x \mapsto \mathbb{1}_{\mathcal{S}}(x)$ est la fonction réelle qui vaut 1 si $x \in \mathcal{S}$, 0 sinon. On rappelle que $x \mapsto \mathrm{sgn}(x)$ est la fonction réelle qui vaut 1 si $x \geq 0$, $-1$ si $x < 0$.

**Questions de cours.** *(3 points)*

1. Quels sont les différents types de données qui peuvent composer une base de données ?
2. Qu'est-ce que l'algorithme K-means ? Dans quels cas et pour quel type de données peut-on l'utiliser ?

**Exercice 1.** *(8 points)* 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, ce qui vous définit un classifieur.
7. Quelle est la valeur prédite par votre classifieur pour le nouveau point $X = (3, 1)$ ?
8. Pour des jeux de données plus grands, utilise-t-on l'arbre complet comme classifieur ?

**Exercice 2.** *(5 points)* On considère le modèle génératif suivant sur $\mathbb{R}^2 \times \{0,1\}$, défini par

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

et par les lois conditionnelles de $X$ sachant $(Y=1)$ et sachant $(Y=0)$ :

$$X|(Y=1) \sim \mathcal{N}\left(\mu_1, \sigma_1^2 \mathrm{I}_2\right) \quad \text{et} \quad X|(Y=0) \sim \mathcal{N}\left(\mu_0, \sigma_0^2 \mathrm{I}_2\right)$$

où $p \in ]0,1[$, $\sigma_0$ et $\sigma_1$ sont des réels positifs et $\mu_0$ et $\mu_1$ sont deux vecteurs de $\mathbb{R}^2$. $\mathrm{I}_2$ désigne la matrice identité de taille $2 \times 2$. On rappelle que $\mathcal{N}\left(\mu, \sigma^2 \mathrm{I}_2\right)$ admet pour densité par rapport à la mesure de Lebesgue de $\mathbb{R}^2$

$$f(x) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{1}{2\sigma^2}\|x-\mu\|^2\right).$$

1. Calculez le classifieur de Bayes $f^* : \mathbb{R}^2 \to \{0,1\}$ sous ce modèle génératif.
2. Calculez la valeur de $f^*(\mu_0)$. Le classifieur de Bayes peut-il prédire la classe 1 en $x = \mu_0$ ?
3. Quelle est la forme de la frontière de décision ?
4. Donnez d'autres exemples de fontières de décision qui peuvent s'obtenir avec d'autres méthodes de classification supervisée que vous connaissez.

**Exercice 3.** *(4 points)* On se donne $\mathcal{D}_n = (X_i, Y_i)_{1 \le i \le n}$ et $\mathcal{D}'_m = (X'_i, Y'_i)_{1 \le i \le m}$ deux bases de données étiquetées dont les éléments $X_i, X'_i \in \mathbb{R}^d$ et $Y_i, Y'_i \in \{-1, 1\}$.

A partir de la base de données $\mathcal{D}_n$, on calcule le classifieur $\hat{f}_n^C(x) = \operatorname{sgn}\left(x^T \hat{\beta}_n + \hat{\beta}_0\right)$ qui dépend d'un paramètre $C > 0$ tel que

$$(\hat{\beta}_n, \hat{\beta}_0) \in \underset{\beta, \beta_0}{\operatorname{argmin}} \ \frac{1}{2}\|\beta\|^2 + C \sum_{i=1}^{n} \xi_i,$$

où la minimisation s'effectue sous les contraintes $\xi_i \ge 0$ et $Y_i(X_i^T \beta + \beta_0) \ge 1 - \xi_i$ pour $i = 1, \ldots, n$.

1. Comment s'appelle ce classifieur ?

On calcule ce classifieur pour plusieurs valeurs de $C$ et on l'évalue sur la base $\mathcal{D}_n$ et sur la base $\mathcal{D}'_m$ en calculant la fraction d'erreurs de prédiction sur chaque base de données, appelée $E(\mathcal{D}_n)$ pour la base $\mathcal{D}_n$ et $E(\mathcal{D}'_m)$ pour la base $\mathcal{D}'_m$.

2. Parmi les graphiques des figures 1 et 2, lequel vous paraît correspondre à ce qui peut se passer ? Justifiez votre réponse.
3. Sur le graphique choisi, pour quelles valeurs de $C$ a-t-on un fort biais et pour quelles valeurs de $C$ a-t-on une forte variance ? Justifiez votre réponse.
4. Quelle est selon vous la "meilleure" valeur de $C$ ?
   En appelant $\hat{C}$ cette valeur, comment évaluer la performance du classifieur $\hat{f}^{\hat{C}}$ ?

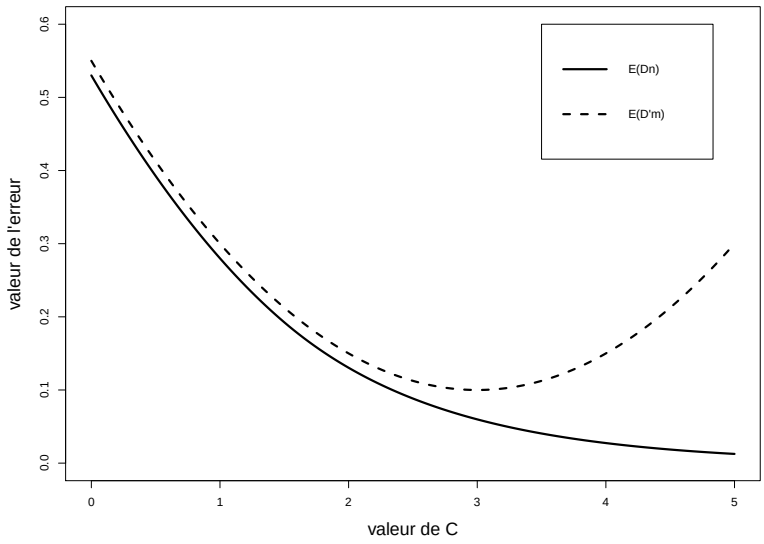


FIGURE 1 – Erreur sur $\mathcal{D}_n$ (trait plein) et $\mathcal{D}'_m$ (pointillés)

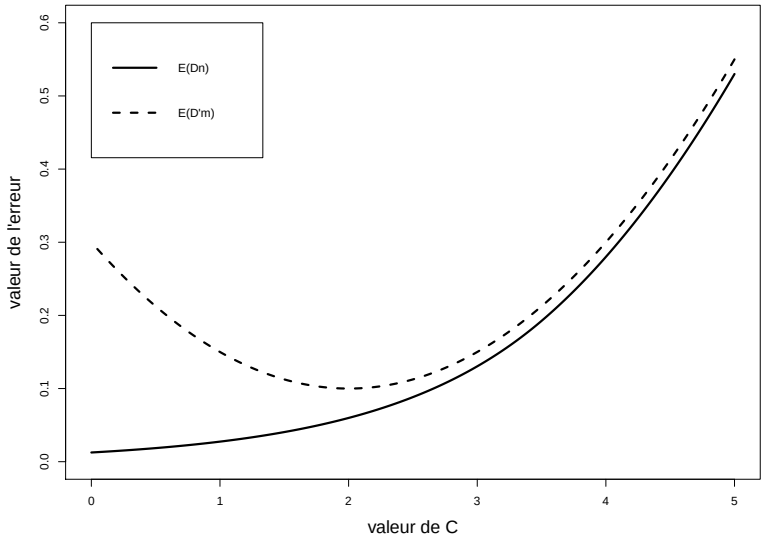


FIGURE 2 – Erreur sur $\mathcal{D}_n$ (trait plein) et $\mathcal{D}'_m$ (pointillés)