Université de Lille 2017-2018
M1 Maths/Finances - Data Mining 16 avril 2018

Examen de Data Mining

Durée : 2 heures. Documents et calculatrices sont interdits.

Exercice 1. La base de données dont le début est affiché ci-dessous a été collectée par un médecin qui a vu de nombreux patients. Il voudrait s'en servir pour prédire si un patient est malade ou non, en fonction de l'observation de symptômes.

Patient	Toux	Fièvre	Nez qui coule	Malade
1	+	+	+	+
2	+	+	-	+
3	+	-	-	+
4	-	-	-	-
5	+	-	+	-
6	-	+	+	-

1. Formuler l'objectif du médecin comme un problème d'apprentissage (on précisera les espaces $\mathcal{X}$ et $\mathcal{Y}$, et ce que l'on cherche à faire).

On propose les hypothèses suivantes sur le modèle qui a généré ces données. Soit Y la variable aléatoire donnant le diagnostic, T, F, N les variables aléatoires indicant respectivement si le patient à la toux, de la fièvre et le nez qui coule. On suppose qu'il existe des paramètres $p, p_T^{\pm}, p_F^{\pm}$ et $p_N^{\pm}$ dans $[0,1]$ tels que

$$\begin{aligned} \mathbb{P}(Y=+) = p,\ \mathbb{P}(T=+|Y=+) &= p_T^+,\ \mathbb{P}(F=+|Y=+) = p_F^+,\ \mathbb{P}(N=+|Y=+) = p_N^+, \\ \mathbb{P}(T=+|Y=-) &= p_T^-,\ \mathbb{P}(F=+|Y=-) = p_F^-, \text{ et } \mathbb{P}(N=+|Y=-) = p_N^-. \end{aligned}$$

De plus, on suppose que les différents symptômes sont indépendants conditionnellement à Y, et que chaque variable aléatoire ne peut prendre que deux valeurs, + et −.

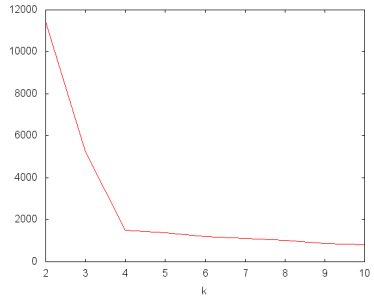
2. Donner la définition du classifieur de Bayes g^* et son expression pour la classification binaire.
3. Exprimer $g^*((-,+,-))$ en fonction des paramètres ci-dessus.
4. Dans le cas où on ne connait pas les paramètres du modèle génératif, comment peut-on s'appuyer sur le classifieur des Bayes pour construire un classifieur à partir des données ?
5. Proposer une prédiction $\hat{g}_n((-,+,-))$ en se basant sur les 6 patients ci-dessus.

Exercice 2. On se réfèrera à l'annexe 1 pour les figures à compléter. On considère un problème de classification où $\mathcal{X} = (\mathbb{R}^+) \times (\mathbb{R}^+)$ et $\mathcal{Y} = \{1,2,3\}$. La figure 1 en annexe 1 donne la représentation graphique d'une base d'apprentissage $\mathcal{D}_n = \{(X_i, Y_i)\}_{1 \le i \le n}$. On donne la légende $\times : 1, \bigcirc : 2, \triangle : 3$.

1. La figure 2 présente un arbre de décision construit à partir de ces données. A quelle partition de l'espace correspond-il ? Représenter cette partition sur la figure 1.
2. Donner l'expression du classifieur associé $\hat{g}_n : \mathcal{X} = (\mathbb{R}^+) \times (\mathbb{R}^+) \to \{1,2,3\}$.
3. Que vaut $\hat{g}_n((50,2))$?

Exercice 3. Soit $\mathcal{D}_n = \{X_i\}_{1 \leq i \leq n}$ une base de données avec $X_i \in \mathbb{R}^d$.

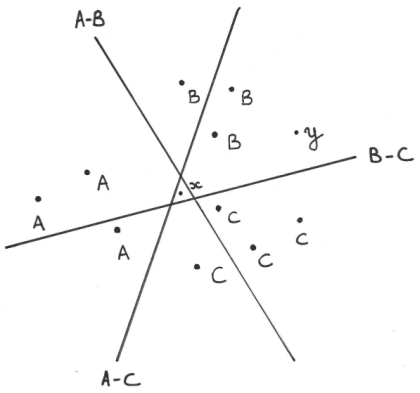
1. Etant donnée une segmentation S des données, donner l'expression du crière $\mathcal{W}_K(S)$ que l'algorithme K-means cherche à minimiser.
2. Le figure A en annexe 2 représente les groupes (clusters) à un certain stade de l'algorithme K-means. Ajouter (approximativement) les centroïdes sur la figure. L'algorithme a-t-il convergé ?
3. L'algorithme K-means a été effectué pour différentes valeur de $K \in \{2, \ldots, 10\}$, retournant une segmentation S_K. La figure ci-dessous représente la valeur de $\mathcal{W}_K(S_K)$ en fonction de K. Parmi les segmentations $S_2, \ldots, S_K$, laquelle vous paraît la plus pertinente ? Justifier.



Exercice 4. On se réfèrera à l'annexe 2 pour les figures à compléter. On s'intéresse à de la classification binaire de $\mathbb{R}^2$ dans $\{-1, 1\}$, où les labels $\times$ correspondent aux 1 et les $\bigcirc$ aux -1.

1. Ajouter sur la figure B 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 $\hat{g}_n : \mathbb{R}^2 \to \{-1, 1\}$. Quelles sont les méthodes vues en cours qui construisent de tels classifieurs ?
3. Représenter graphiquement sur le figure C 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 D en annexe ? Ajouter une frontière de décision possible sur cette figure.

La figure ci-dessous représente les données d'un problème de classification multi-classes avec $\mathcal{Y} = \{A, B, C\}$, ainsi que trois séparateurs linéaires ayant été entraînés en se basant sur les données ne correspondant qu'à deux classes. Les points x et y sont des points dont on ne connait pas l'étiquette.



5. Expliquer comment combiner ces séparateurs linéaires pour construire un classifieur $\hat{g}_n$ à valeurs dans $\{A, B, C\}$. Donner les valeurs de $\hat{g}_n(x)$ et $\hat{g}_n(y)$.

Exercice 5. On considère une base de données $\mathcal{D}_n = \{(X_i, Y_i)\}_{1 \leq i \leq n}$ avec $X_i \in \mathbb{R}^d$ et $Y_i \in \mathbb{R}$.

1. Si on collecte une base de données qui comprend des variables explicatives qualitatives, comment peut-on se ramener à $X_i \in \mathbb{R}^d$?
2. On définit le prédicteur $\hat{g}_n : \mathbb{R}^d \to \mathbb{R}$ tel que $\hat{g}_n(x) = \hat{\theta}_n^T x + \hat{b}_n$, avec $\hat{\theta}_n \in \mathbb{R}^d$ et $\hat{b}_n \in \mathbb{R}$, où

$$(\hat{\theta}_n, \hat{b}_n) \in \underset{(\theta,b)\in\mathbb{R}^d\times\mathbb{R}}{\operatorname{argmin}} \sum_{i=1}^{n} \left(Y_i - \theta^T X_i - b\right)^2$$

   De quel prédicteur s'agit-il ?
3. Si $Y_i \subseteq \{0, 1\}$ comment convertir ce prédicteur en un classifieur $\hat{h}_n : \mathbb{R}^d \to \{0, 1\}$?
4. Justifier qu'il existe $\hat{\beta}_n \in \mathbb{R}^{d+1}$ tel que $\hat{g}_n(x) = \hat{\beta}_n{}^T \underline{x}$ où $\underline{x} \in \mathbb{R}^{d+1}$ s'obtient à partir de x en rajoutant un 1 à ce vecteur.

On fixe $\lambda \geq 0$ et on définit le prédicteur $\hat{g}_n^\lambda(x) = (\hat{\beta}_n^\lambda)^T \underline{x}$ avec

$$\hat{\beta}_n^\lambda \in \underset{\beta\in\mathbb{R}^{d+1}}{\operatorname{argmin}} \; F_{n,\lambda}(\beta) \quad \text{où} \quad F_{n,\lambda}(\beta) = \sum_{i=1}^{n} \left(Y_i - \beta^T \underline{X}_i\right)^2 + \lambda \|\beta\|^2$$

5. Calculer le gradient de $F_{n,\lambda}$ en un point $\beta \in \mathbb{R}^{d+1}$ et montrer que

$$-\sum_{i=1}^{n} Y_i \underline{X}_i + \left(\sum_{i=1}^{n} \underline{X}_i^T \underline{X}_i\right) \hat{\beta}_n^\lambda + \lambda \hat{\beta}_n^\lambda = 0.$$

6. Justifier que la matrice $\sum_{i=1}^n \underline{X}_i^T \underline{X}_i + \lambda \mathrm{I}_{d+1}$ est toujours inversible.
7. Montrer que $\hat{\beta}_n^\lambda = \left(\sum_{i=1}^n \underline{X}_i^T \underline{X}_i + \lambda \mathrm{I}_{d+1}\right)^{-1} \left(\sum_{i=1}^n Y_i \underline{X}_i\right)$.
8. Si d est très grand, l'inversion de la matrice pour le calcul de $\hat{\beta}_n^\lambda$ est très coûteuse. Quelle(s) méthode(s) pouvez-vous mettre en œuvre en pratique pour réduire la dimension ?
9. Quel peut être l'intérêt du prédicteur $\hat{g}_n^\lambda$ par rapport au prédicteur $\hat{g}_n$?