Université de Lille 2018-2019
M1 Maths/Finances - ISN 25 avril 2019

Examen de Data Mining

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

Question de cours : Qu'est-ce que le sur-apprentissage ?

Exercice 1. On considère de modèle génératif suivant sur $\mathbb{R} \times \{0,1\}$. 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

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

avec $p \in [0,1]$, et μ_0 et μ_1 deux réels. $\mathcal{N}(\mu, 1)$ désigne une loi normale de moyenne μ et de variance 1, dont on rappelle que la densité est $\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2}\right)$.

1. Justifier que le classifieur optimal, ou classifieur de Bayes, est donné par
$$g^*(x) = \mathbb{1}_{(\mathbb{P}(\boldsymbol{Y}=1|\boldsymbol{X}=x)>1/2)}$$
2. Donner l'expression la plus explicite possible de ce classifieur de Bayes.
3. Pour $p = 1/2$, $\mu_0 = -1$ et $\mu_1 = 1$, 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.
4. 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. On donne le code Python suivant.

```
from sklearn.tree import DecisionTreeClassifier
tree = DecisionTreeClassifier(max_depth=4,min_samples_leaf=5)
```

1. Ce code créé un objet `tree` qui définit un classifieur. Expliquer comment avec Python on peut entraîner ce classifieur sur une base de données et prédire l'étiquette de nouvelles données.
2. Quel est le nom de l'algorithme de classification utilisé ?
   Expliquer le rôle des deux paramètres `max_depth` et `min_samples_leaf`.
3. Quelles valeurs peut-on choisir pour ces paramètres afin de construire un arbre $\mathcal{T}$ aussi grand que possible ? Un tel arbre est appelé arbre complet.
4. En pratique, comment pouvez-vous choisir de bonnes valeurs des paramètres `max_depth` et `min_samples_leaf` ?

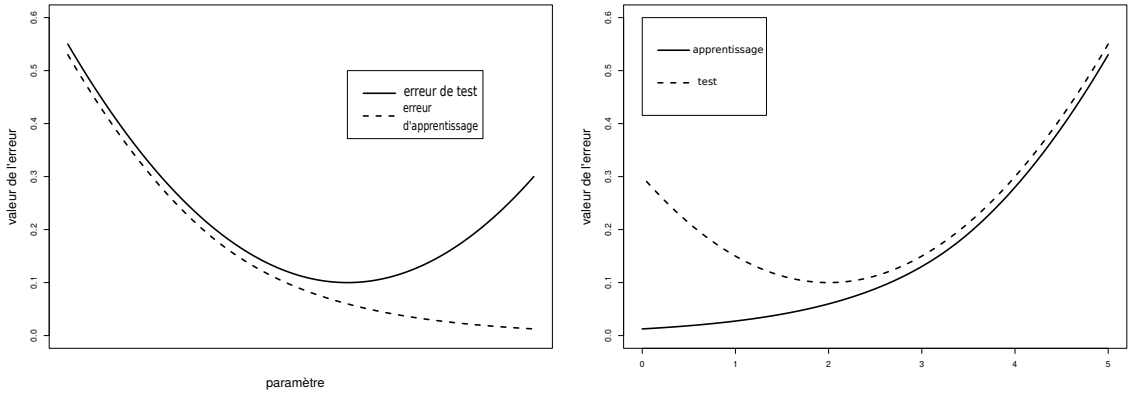
Plutôt que de sélectionner les deux paramètres `max_depth` et `min_samples_leaf`, on propose le processus suivant, appelé élagage, qui ne nécessite le choix que d'un seul paramètre, α.

A partir des données d'apprentissage, on construit d'abord l'arbre complet $\mathcal{T}$. Pour chaque valeur de $\alpha \in \mathbb{R}$, on définit alors

$$\mathcal{S}_\alpha = \underset{\mathcal{S} \subseteq \mathcal{T}}{\operatorname{argmin}} \left[\sum_{\mathcal{L} \in \text{Leaf}(\mathcal{S})} Q(\mathcal{L}) + \alpha \times \#\,|\text{Leaf}(\mathcal{S})| \right]$$

où $Q(\mathcal{N})$ désigne l'impureté d'un noeud $\mathcal{N}$, $\text{Leaf}(\mathcal{S})$ est l'ensemble des feuilles de l'arbre $\mathcal{S}$ et $\#\,|\mathcal{E}|$ désigne le cardinal de l'ensemble $\mathcal{E}$. La minimisation s'effectue sur l'ensemble des sous-arbres $\mathcal{S}$ de l'arbre complet $\mathcal{T}$, et on admet qu'elle peut être effectuée de manière efficace.

5. A quoi correspond l'arbre $\mathcal{S}_\alpha$ pour $\alpha = 0$? Pour α très grand ?
6. Parmi les deux graphiques ci-dessous, laquelle pourrait représenter l'erreur de test et l'erreur d'apprentissage en fonction du paramètre α ? Justifier votre réponse.



Exercice 3. Dans cet exercice, $\langle x|y \rangle = x_1y_1 + x_2y_2$ désigne le produit scalaire sur $\mathbb{R}^2$ des vecteurs $x = (x_1, x_2)$ et $y = (y_1, y_2)$. On considère un problème de classification binaire à partir de la base de données $\mathcal{D}_n = \{(X_i, Y_i)\}_{1 \leq i \leq 8}$ ci-dessous où $X_i = (X_i^1, X_i^2) \in \mathbb{R}^2$ et $Y_i \in \{1, -1\}$.

i	X_i^1	X_i^2	Y_i
1	-2	-2	1
2	-2	-1	1
3	1	2	1
4	2	1	1
5	-2	2	-1
6	0	2	-1
7	0	-1	-1
8	2	-1	-1

1. Représenter graphiquement les données dans $\mathbb{R}^2$. Sont-elles linéairement séparables ?
2. Proposer deux algorithmes d'apprentissage construisant des séparateurs linéaires.
   Vous paraît-il pertinent de les appliquer ici ?

On définit l'application $\Phi : \mathbb{R}^2 \to \mathbb{R}^2$ par $\Phi(u, v) = (u^2, uv)$ ainsi que la base de données transformée

$$\mathcal{D}_n^\Phi = \{(\Phi(X_i), Y_i)\}_{1 \leq i \leq 8}.$$

3. Représenter graphiquement les données de $\mathcal{D}_n^\Phi$ dans $\mathbb{R}^2$. Sont-elles linéairement séparables ?
4. On définit la solution du problème d'optimisation suivant

$$(w^*, b^*) \in \underset{\substack{w \in \mathbb{R}^d, b \in \mathbb{R}: \\ \forall i, Y_i(\langle \phi(X_i) | w \rangle + b) \geq 1}}{\operatorname{argmin}} \frac{1}{2}\|w\|^2.$$

   Donner l'expression du SVM linéaire calculé à partir de $\mathcal{D}_n^\Phi$ en fonction de w^* et b^*.

On rappelle que w^* s'écrit comme combinaison linéaire d'un petit nombre d'observations appelées *vecteurs supports*, satisfaisant $Y_i(\langle \phi(X_i) | w^* \rangle + b^*) = 1$. On appelle hyperplans de marge les deux hyperplans d'équations

$$\langle x | w^* \rangle + b^* = 1 \quad \text{et} \quad \langle x | w^* \rangle + b^* = -1.$$

5. En admettant que les vecteurs supports sont $\Phi(X_3)$, $\Phi(X_6)$ et $\Phi(X_8)$, ajouter les deux hyperplans de marge sur la figure de la question 3.
6. Calculer w^* et b^* et donner l'équation de l'hyperplan séparateur. L'ajouter sur la figure précédente.
7. Donner l'équation du séparateur associé au SVM de noyau $k(x, x') = \langle \Phi(x) | \Phi(x') \rangle$ calculé à partir de la base de données initiale $\mathcal{D}_n$.

Exercice 4. On souhaite résoudre à l'aide d'une méthode d'apprentissage le problème de reconnaissance d'objets suivants. On dispose d'un grand nombre de "dessins" pixelisé de taille 10 par 10, où chaque pixel peut être blanc ou noir. Ces dessins représentent soit des animaux, soit des humains, soit des véhicules. On souhaite développer une méthode automatique qui étant donné un dessin, identifie s'il représente un animal, un humain ou un véhicule.

1. Proposer une formulation mathématique du problème à résoudre : définir l'espace des variable explicatives et de la variable à prédire, et formaliser l'objectif.
2. On souhaite représenter graphiquement une projection pertinente des variables explicatives en dimension 2. Comment peut-on procéder ?

La figure au verso présente cette projection pour $n = 16$ observations appartenant à deux des trois catégories, représentées par des $\circ$ et des $\times$. On souhaite appliquer l'algorithme K-means avec $K = 2$ à ces données, pour obtenir une partition de celles-ci en deux groupes.

3. En prenant pour centroïdes initiaux les observations notées C_1 et C_2, indiquer sur la figure les clusters et les centroïdes (approximatifs) obtenus après une étape de l'algorithme.
4. En combien d'étapes l'algorithme va-t-il converger ?
5. L'utilisation de l'algorithme K-means vous paraît-elle appropriée pour effectuer un clustering de ces données ? Quel(s) autre(s) algorithme(s) auriez-vous pu utiliser ?
6. Expliquer comment le résultat fourni par un algorithme de clustering peut être utile pour résoudre le problème initial de reconnaissance d'objets.

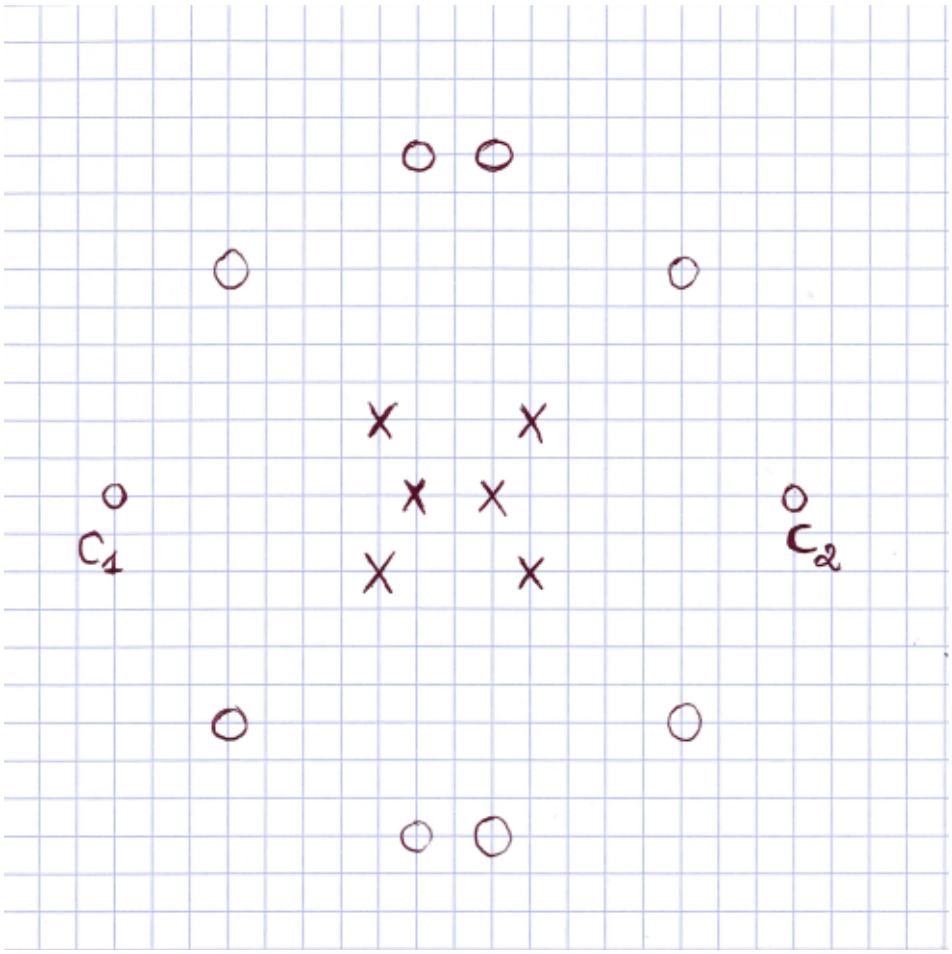


FIGURE 1 – Figure à rendre avec la copie