Université de Lille 1 2016-2017
M1 Maths/Finances - Data Mining 28 avril 2017

Devoir surveillé

Durée : 2 heures. Documents et calculatrices sont interdits.

Notations. 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 $x < 0$.

Exercice 1. *(4 points)* Un sondage téléphonique a permis de collecter la base de données dont le début est affiché ci-dessous. La variable à prédire (le vote du second tour) est isolée dans la dernière colonne.

Individu	Age	Sexe	Profession (cat.)	Revenus (k€)	Etudes (an)	Région	Vote 1	Vote 2
1	25	M	5	25	5	Hauts-de-France	1	"MLP"
2	37	M	3	27	3	Grand Est	2	"EM"
3	58	M	2	35	5	Ile-de-France	3	"EM"
4	30	F	3	25	8	Hauts-de-France	9	"EM"
5	18	F	6	17	2	Occitanie	3	"MLP"
6	45	M	1	20	0	Bretagne	1	"MLP"

1. Décrire cette base de données (nombre et types de variables et leur représentation) ainsi que la tâche d'apprentissage demandée.
2. Peut-on appliquer une Analyse en Composantes Principales ? L'algorithme k-means ?

Exercice 2. *(5 points)* On considère un problème de classification binaire où $\mathcal{X} = [0,1]$ et $\mathcal{Y} = \{0,1\}$, avec la fonction de perte $\ell(y,y') = \mathbb{1}_{(y \neq y')}$.

1. On donne la base d'apprentissage suivante

$(X_1 = 0.2, Y_1 = 0)$ $(X_2 = 0.3, Y_2 = 0)$ $(X_3 = 0.5, Y_3 = 1)$ $(X_4 = 0.6, Y_4 = 0)$ $(X_5 = 0.8, Y_5 = 1)$

   Calculer le classifieur obtenu par la méthode des k plus proches voisins pour $k = 1, 3$ et 5.

2. On suppose que les données sont générées selon le modèle suivant :

$$X \sim \mathcal{U}([0,1]) \quad \text{et} \quad \mathbb{P}(Y = 1 | X = x) = \frac{1}{3}, \quad \mathbb{P}(Y = 0 | X = x) = \frac{2}{3},$$

   où $\mathcal{U}([0,1])$ est la loi uniforme sur $[0,1]$. Calculer le classifieur de Bayes et son risque.

3. Soit $\mathcal{D}_n = (X_i, Y_i)_{1 \leq i \leq n}$ une base d'apprentissage tirée sous le modèle génératif ci-dessus et soit $\hat{f}_n$ le classifieur du plus proche voisin ($k = 1$). Justifier (on ne demande pas une démonstration précise) que pour X tiré sous le modèle génératif ci-dessus,

$$\mathbb{P}(\hat{f}_n(X) = 1 | \mathcal{D}_n) = \frac{1}{3},$$

4. Calculer le risque du classifieur $\hat{f}_n$ et comparez-le au risque minimal.

Exercice 3. *(5 points)* On se donne $\mathcal{D}_n = (X_i, Y_i)_{1\leq i\leq n}$ et $\mathcal{D}'_m = (X'_i, Y'_i)_{1\leq i\leq m}$ une base d'apprentissage et une base de test avec $X_i \in \mathbb{R}^d$ et $Y_i \in \{-1, 1\}$.

1. Rappeler l'expression de l'erreur d'apprentissage et de l'erreur de test.

Un classifieur $\hat{f}_n(p)$ qui dépend d'un paramètre p a été entraîné sur la base $\mathcal{D}_n$ pour différentes valeurs de p, et son erreur de test a également été calculée. On représente les deux types d'erreur sur le même graphique et on obtient la Figure 1.

2. Parmi les deux courbes de la Figure 1, identifier celle qui correspond à l'erreur d'apprentissage, et celle qui correspond à l'erreur de test. Justifier votre réponse.
3. Pour quelles valeurs du paramètre (grandes ou petites) observe-t-on un fort biais ? Une forte variance ? Justifier votre réponse.
4. Peut-on se servir de la Figure 1 pour sélectionner une bonne valeur du paramètre p ?
   Si oui, expliquer comment. Si non, proposer une méthode alternative.
5. Pour l'algorithme CART, quel paramètre joue le rôle du paramètre p ci-dessus et permet d'obtenir le même genre de courbes ?

Exercice 4. *(6 points)* On $\mathcal{D}_n = (X_i, Y_i)_{1\leq i\leq n}$ un jeu de données avec pour chaque i, $X_i \in \mathbb{R}^d$ et $Y_i \in \{-1, 1\}$. Soit $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}^+$ une fonction de perte.

1. Rappeler l'expression du risque d'un classifieur $f : \mathbb{R}^d \to \{-1, 1\}$ par rapport à la fonction de perte ℓ, et du risque empirique calculé sur la base de données $\mathcal{D}_n$.

Plusieurs méthodes de construction de classifieur peuvent s'écrire sous la forme $\hat{f}_n = \operatorname{sgn}(\hat{g}_n(x))$ où

$$\hat{g}_n \in \operatorname*{argmin}_{g\in\mathcal{G}} \sum_{i=1}^{n} \ell(g(X_i), Y_i),$$

avec $\mathcal{G}$ est en ensemble de fonctions (par exemple : toutes les fonctions linéaires). On dit que le classifieur est obtenu par minimisation du risque empirique pour la fonction de perte ℓ.

On propose ci-dessous des exemples de fonctions de perte :

$$\begin{array}{lcllcl} \ell_1(u,v) & = & \mathbb{1}_{(u\neq v)} & \ell_2(u,b) & = & (1-uv)\mathbb{1}_{(1-uv>0)} \\ \ell_3(u,v) & = & (u-v)^2 & \ell_4(u,v) & = & \mathbb{1}_{(uv<0)} \\ \ell_5(u,v) & = & \log\left(1+e^{-uv}\right) & & & \end{array}$$

2. Parmi les fonction de pertes ci-dessus, identifier les deux fonctions qui sont égales lorsque u et v appartienent à $\{-1, 1\}$.
3. Pour quelle fonction de perte (et quel ensemble $\mathcal{G}$ de fonctions) la régression linéaire peut-elle être vue comme une minimisation du risque empirique ? Même question pour la régression logistique.
4. Quelle méthode de classification supervisée peut être vue comme une minimisation du risque empirique pour la fonction de perte ℓ_4 ?
5. Le graphique de la Figure 2 représente les fonctions de perte qui ne dépendent que du produit uv. Identifier les fonctions correspondant à Perte 1, Perte 2 et Perte 3.
6. Serait-il facile de calculer le classifieur par minimisation du risque empirique associé à la fonction Perte 3 de la Figure 2 ?

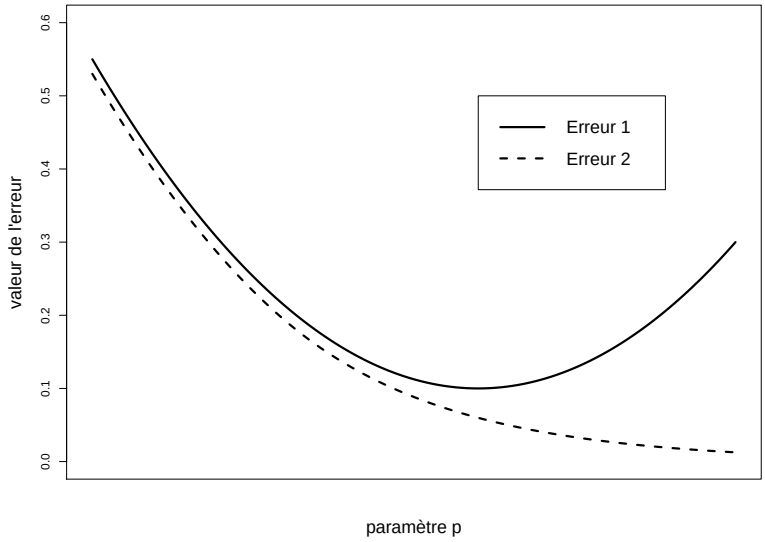

FIGURE 1 – Deux types d'erreurs.

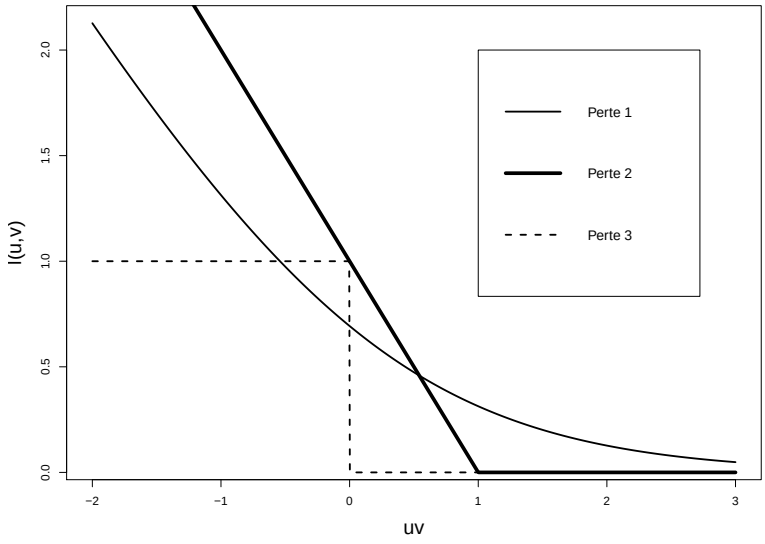

FIGURE 2 – Plusieurs fonctions de perte.