# Examen d'analyse de données – sujet 1

Fabrice Rossi

2017

Les exercices sont indépendants et peuvent être traités dans n'importe quel ordre.

## Exercice 1

On étudie un ensemble de 20 observations $(X_i, Y_i)_{1 \leq i \leq 20}$, avec $Y_i \in \{1, 2, 3\}$. On construit sur cet ensemble deux modèles, $g_1$ et $g_2$ dont les prévisions sur cet ensemble sont données par le tableau suivant :

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $Y_i$ | 3 | 1 | 2 | 2 | 3 | 1 | 3 | 3 | 2 | 2 | 1 | 1 | 1 | 3 | 2 | 3 | 2 | 3 | 3 | 2 |
| $g_1(X_i)$ | 1 | 3 | 2 | 2 | 3 | 1 | 3 | 3 | 2 | 3 | 1 | 1 | 1 | 3 | 2 | 1 | 2 | 1 | 3 | 3 |
| $g_2(X_i)$ | 3 | 3 | 2 | 1 | 3 | 1 | 2 | 3 | 2 | 1 | 1 | 1 | 3 | 3 | 2 | 3 | 2 | 3 | 3 | 2 |

Dans ce tableau, chaque colonne correspond à une observation $(X_i, Y_i)$ et précise la valeur de $Y_i$ et les prévisions des deux modèles.

**Question 1** Calculer les matrices de confusion empiriques des deux modèles.

**Question 2** Déterminer le meilleur modèle (entre $g_1$ et $g_2$) au sens du risque empirique pour la fonction de perte $l_1(u,v) = |u - v|$.

**Question 3** Même question pour la fonction de perte $l_2$ donnée par la table suivante :

| $l_2(u,v)$ | | $v$ | | |
|---|---|---|---|---|
| | | 1 | 2 | 3 |
| $u$ | 1 | 0 | 2 | 1 |
| | 2 | 1 | 0 | 1 |
| | 3 | 2 | 1 | 0 |

On rappelle que par convention, le premier argument de la fonction de perte est la prévision du modèle.

## Exercice 2

On étudie un problème à trois variables explicatives binaires $X1$, $X2$ et $X3$ et à une variable à expliquer binaire $Y$ (toutes les variables sont donc à valeurs dans $\{0, 1\}$). On suppose que le modèle optimal pour la fonction de perte $l_0(u,v) = \mathbb{I}_{u \neq v}$ est donné par l'arbre de décision de la figure 1. Il s'agit du modèle optimal $g^*$ théorique.

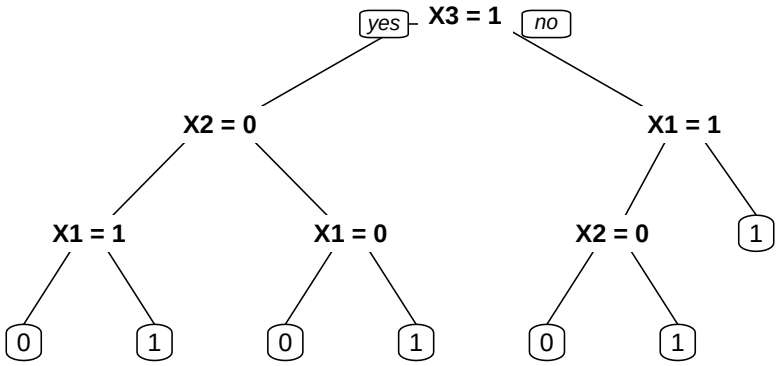


FIGURE 1 – Modèle optimal. La branche de gauche de chaque nœud correspond toujours à la réponse « oui » à la question du nœud, l'autre branche à la réponse « non ». La valeur indiquée dans chaque feuille donne la classe associée à celle-ci.

On dispose d'un total de 100 observations, dont 63 correspondent $Y = 1$. Les observations sont décrites de façon simplifiée par le tableau suivant :

| Valeur de Y | Nombre de X1=1 | Nombre de X2=1 | Nombre de X3=1 |
|---|---|---|---|
| 0 | 25 | 10 | 26 |
| 1 | 23 | 39 | 13 |

Chaque ligne correspond à une valeur de $Y$ (la première concerne les observations telles que $Y = 0$, la seconde celles pour lesquelles $Y = 1$). Les trois colonnes des variables explicatives indiquent le nombre d'observations avec la valeur 1 pour la variable concernée et avec simultanément la valeur de $Y$ précisée en ligne. Par exemple, la valeur 25 de la première ligne, colonne $X1$ indique que parmi les 37 observations pour lesquelles $Y = 0$, 25 ont une valeur de 1 pour la variable $X1$. De même la valeur 39 de la deuxième ligne, colonne $X2$ indique que parmi les 63 observations pour lesquelles $Y = 1$, 39 ont une valeur de 1 pour la variable $X2$.

On cherche à construire un classifieur bayésien naïf sur ces données.

**Question 1** Donner de façon précise et complète le modèle génératif du classifieur bayésien naïf (CBN) pour le problème étudié. Rappeler en particulier les différentes lois choisies et les hypothèses d'indépendance utilisées.

**Question 2** En utilisant le principe du maximum de vraisemblance, estimer grâce aux informations fournies ci-dessus les paramètres des lois utilisées par le modèle CBN.

**Question 3** Exprimer pour le modèle CBN $\log \frac{\mathbb{P}(Y=1|X=x)}{\mathbb{P}(Y=0|X=x)}$ (avec $X = (X1, X2, X3)$) sous forme d'une fonction affine en $x = (x1, x2, x3)$ faisant apparaître les paramètres du modèle.

**Question 4** Calculer sous forme d'une table la fonction correspondant au modèle optimal (pour la fonction de perte $l_0$) en supposant que les données sont bien distribuées selon un modèle CBN et en remplaçant les paramètres réels par les estimations obtenues à la question précédente.

Les matrices de confusion de l'arbre de la figure 1 et du modèle optimal calculé à la question 4 sont données ci-dessous (pour les 100 observations) :

| | 0 | 1 |
|---|---|---|
| 0 | 33 | 1 |
| 1 | 4 | 62 |

Arbre

| | 0 | 1 |
|---|---|---|
| 0 | 17 | 12 |
| 1 | 20 | 51 |

CBN

Dans ces matrices, les prévisions des modèles sont en ligne, les valeurs réelles en colonne.

**Question 5** Pourrait-on calculer ces matrices à partir des informations fournies dans l'énoncé ?

**Question 6** Représenter la fonction de décision de l'arbre optimal sous forme d'une table similaire à celle construite à la question 4.

**Question 7** Sachant que l'arbre est le modèle optimal, pouvait-on attendre des résultats satisfaisants lors de l'utilisation du modèle CBN ? Réponse à justifier soigneusement.

## Exercice 3

On étudie un ensemble de 310 patients d'un service hospitalier. Chaque patient est décrit par six variables numériques ($X1$ à $X6$) et par une variable nominale $Y$ prenant les trois valeurs $\{DH, NO, SL\}$. La valeur $NO$ désigne un « patient » sain (référence pour l'étude). Les deux autres modalités correspondent à deux pathologies.

Dans un premier temps, l'analyste applique la méthode CART aux données en utilisant l'intégralité des observations comme ensemble d'apprentissage. La figure 2 représente l'évolution du nombre d'erreurs de classement en fonction du nombre de feuilles de l'arbre, sur l'ensemble d'apprentissage et estimé par une validation croisée à 10 blocs.

**Question 1** L'analyste décide de retenir un arbre à 5 feuilles. Justifier son choix.

**Question 2** L'arbre élagué est donné par la figure 3. Déduire de la figure la matrice de confusion de l'arbre à 5 feuilles sur les données d'apprentissage.

**Question 3** L'analyste souhaite estimer le taux d'erreurs que l'arbre à 5 feuilles aura sur de nouvelles données. Peut-elle le faire à partir des éléments présentés jusqu'à présent ?

L'analyste décide de partitionner aléatoirement l'ensemble des données en deux sous-ensembles, $A$ et $B$, d'effectifs approximativement égaux.

**Question 4** L'analyste impose une (des) contrainte(s) sur la partition : laquelle (lesquelles) ?

En appliquant une méthode de validation croisée respectant la (les) même(s) contrainte(s) que la partition, l'analyste obtient un arbre « optimal » avec 5 feuilles. La matrice de confusion de l'arbre obtenu appliqué à l'ensemble $B$ est

| | DH | NO | SL |
|---|---|---|---|
| DH | 21 | 15 | 1 |
| NO | 8 | 32 | 2 |
| SL | 1 | 3 | 72 |

où les prévisions sont en ligne.

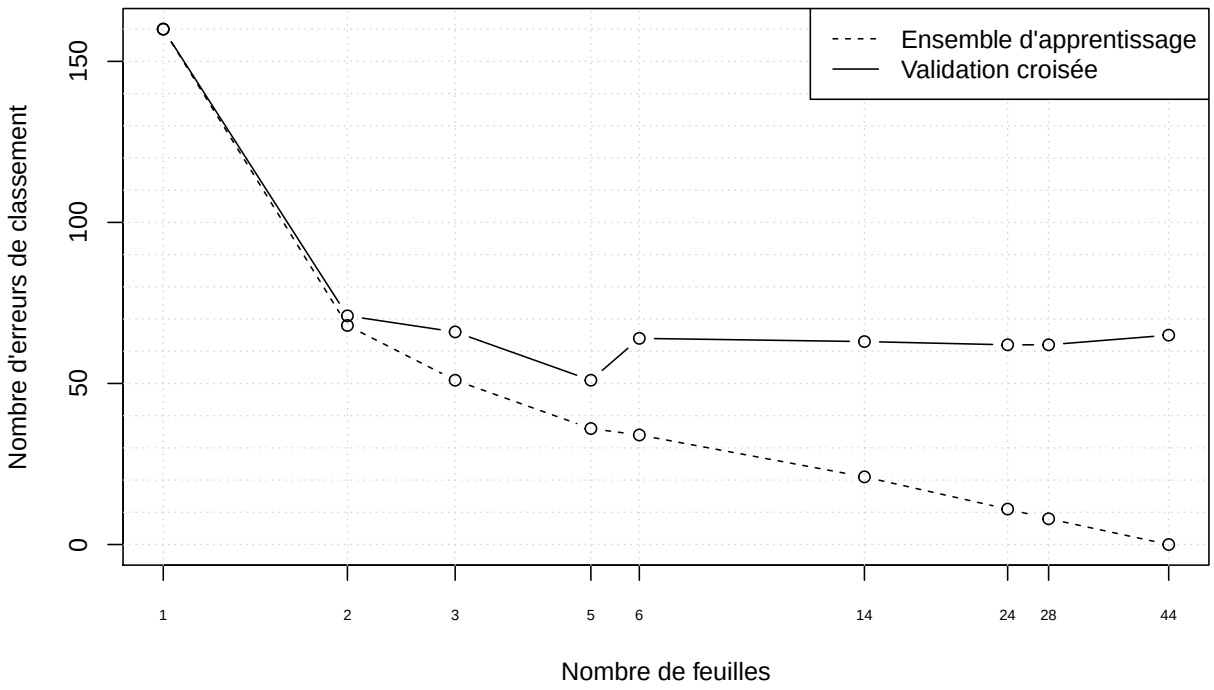


FIGURE 2 – Évolution du nombre d'erreurs de classement en fonction du nombre de feuilles dans l'arbre, sur l'ensemble d'apprentissage et estimé par validation croisée. L'axe des x utilise une échelle logarithmique.

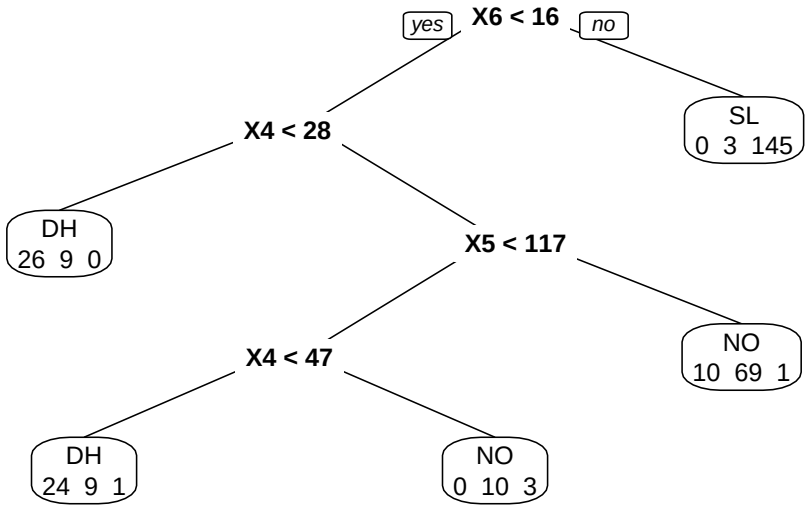


FIGURE 3 – Arbre de classification optimal. La branche de gauche de chaque nœud correspond toujours à la réponse « oui » à la question du nœud, l'autre branche à la réponse « non ». Les lettres indiquées sur la première ligne de chaque feuille donnent la classe associée à la feuille. Les nombres de la deuxième ligne d'une feuille indiquent les effectifs des trois classes dans la feuille, dans l'ordre DH, NO et SL.

**Question 5** Comparer les résultats avec ceux obtenus sur l'ensemble complet.

**Question 6** Quelles sont les performances attendues pour l'arbre obtenu sur l'ensemble $A$ si on l'appliquait à un nouvel ensemble d'observations issu de patients similaires à ceux étudiés ici ?

## Exercice 4

Soit $\mathcal{G}$ un espace vectoriel de fonctions de $\mathbb{R}^d$ dans $\mathbb{R}$ de dimension $m$. On rappelle que la dimension de Vapnik-Chervonenkis ($VC - dim$) de l'ensemble $\mathcal{F}$ de fonctions de $\mathbb{R}^d$ dans $\{0, 1\}$ défini par

$$\mathcal{F} = \{f : \mathbb{R}^d \to \{0, 1\} \mid \exists g \in \mathcal{G}, \forall x \in \mathbb{R}^d, f(x) = \mathbb{I}_{g(x)\geq 0}\},$$

est telle que $VC - dim(\mathcal{F}) \leq m$.

On s'intéresse à la classe de fonctions $\mathcal{H}$ définie par

$$\mathcal{H} = \{h : \mathbb{R}^d \to \{0, 1\} \mid \exists a \in \mathbb{R}^d, b \in \mathbb{R}, \forall x \in \mathbb{R}^d, f(x) = \mathbb{I}_{\|x-a\|^2\leq b^2}\}.$$

En termes moins formels, il s'agit des fonctions indicatrices des boules fermées de $\mathbb{R}^d$ ($a$ est le centre de la boule, $b$ est son rayon, $h(x)$ vaut 1 si et seulement si $x$ est dans la boule). En utilisant le résultat rappelé ci-dessus, on cherche à montrer que $VC - dim(\mathcal{H}) \leq d + 2$.

**Question 1** Pour $a \in \mathbb{R}^d$ et $b \in \mathbb{R}$ fixés, écrire $b^2 - \|x - a\|^2$ sous la forme d'une combinaison linéaire de fonctions de $\mathbb{R}^d$ dans $\mathbb{R}$ indépendantes de $a$ et $b$ : les coefficients peuvent dépendre de $a$ et $b$, mais pas l'expression des fonctions.

**Question 2** Montrer que les fonctions obtenues sont linéairement indépendantes. On pourra procéder par l'absurde.

**Question 3** Conclure.