Examen d'apprentissage statistique – sujet 1

Fabrice Rossi

13 janvier 2016

Exercice 1

On étudie des modèles définis sur $\mathcal{X}$ à valeurs dans $\mathcal{Y} = \{-1, 1\}$. On évalue ces modèles sur un ensemble de test comportant 100 observations. Deux modèles g_1 et g_2 sont comparés. Ils ont les matrices de confusion suivantes (en ligne la valeur prédite par le modèle, en colonne la valeur réelle) :

	$Y=-1$	$Y=1$
$g_1(X)=-1$	40	15
$g_1(X)=1$	5	40

	$Y=-1$	$Y=1$
$g_2(X)=-1$	50	5
$g_2(X)=1$	10	35

Question 1 Déterminer le modèle optimal (entre g_1 et g_2) au sens de la fonction de perte $l_0(p,v) = \mathbb{I}_{p \neq v}$ (p est la prévision, v la vraie valeur).

Question 2 Même question avec la fonction de perte l_1 définie par :

$l_1(p,v)$	$v=-1$	$v=1$
$p=-1$	0	2
$p=1$	1	0

où p désigne la valeur prédite et v la vraie valeur.

Exercice 2

On étudie un jeu de données à trois variables, X_1 à valeurs dans $\{A, B, C\}$, X_2 à valeurs dans $\{1, 2\}$ et Y, la variable cible, à valeurs dans $\{1, 2, 3\}$. Le jeu de données comporte 300 observations, avec 100 observations pour chaque valeur de Y.

Les tables suivantes indiquent pour les deux variables X_1 et X_2 le nombre d'observations prenant une valeur donnée de la variable X_i (en ligne) et une valeur de donnée de la variable Y (en colonne). Par exemple la valeur 26 dans la première case en haut à gauche de la table pour X_1 indique que sur 300 observations, 26 sont telles que $X_1 = A$ et $Y = 1$.

		1	2	3
X_1	A	26	24	23
	B	41	53	14
	C	33	23	63

		1	2	3
X_2	1	53	29	83
	2	47	71	17

Question 1 Rappeler, de façon précise et spécifique au problème étudié, les hypothèses du classifieur bayésien naïf (CBN).

Question 2 Ces hypothèses sont elles vérifiées ici ?

Question 3 Estimer les probabilités utiles pour la construction du CBN à partir des données.

Quelle que soit la réponse à la question 2, on suppose à partir de maintenant que les données vérifient l'hypothèse du CBN.

Question 4 En utilisant l'erreur $l_0(p,v) = \mathbb{I}_{p \neq v}$ (p est la prévision, v la vraie valeur), déterminer le classifieur optimal, en supposant que les probabilités estimées à partir des données sont les vraies probabilités. On donnera le classifieur complet, c'est-à-dire les valeurs prédites par le classifieur pour toutes les valeurs possibles du couple (X_1,X_2).

Question 5 Calculer le risque du classifieur optimal.

Exercice 3

On étudie dans cet exercice un ensemble de 1473 personnes décrites par 9 variables (X_1 à X_9) et une variable cible Y prenant trois valeurs possibles 1, 2 et 3. Les variables explicatives ont les caractéristiques suivantes :

- X_1 et X_4 sont numériques (valeurs entières positives uniquement) ;
- X_5, X_6 et X_9 sont nominales binaires (valeurs 0 ou 1) ;
- X_2, X_3 et X_8 sont ordinales : elles prennent les valeurs 1, 2, 3 ou 4 et l'ordre des valeurs a un sens ;
- X_7 est nominale (valeurs 1, 2, 3 et 4 sans ordre).

Dans un premier temps, les variables ordinales sont considérées comme numériques. Les effectifs des classes (les valeurs de Y) sont donnés par la table suivante :

	1	2	3
Y	629	333	511

Question 1 L'analyste commence par partitionner aléatoirement l'ensemble des données en deux sous-ensembles, A et B, d'effectifs approximativement égaux. Quelle contrainte a-t-on intérêt à imposer au tirage aléatoire ?

On construit un arbre de décision complet sur l'ensemble A. Il comporte 310 feuilles. Les matrices de confusion ci-dessous résument les performances de cet arbre sur l'ensemble A à gauche et sur l'ensemble B à droite. Les lignes des matrices correspondent aux prévisions et les colonnes aux vraies valeurs.

	1	2	3
1	308	0	4
2	1	164	6
3	6	3	246

Matrice de confusion sur A

	1	2	3
1	185	51	83
2	43	67	58
3	86	48	114

Matrice de confusion sur B

Question 2 Commenter les résultats obtenus.

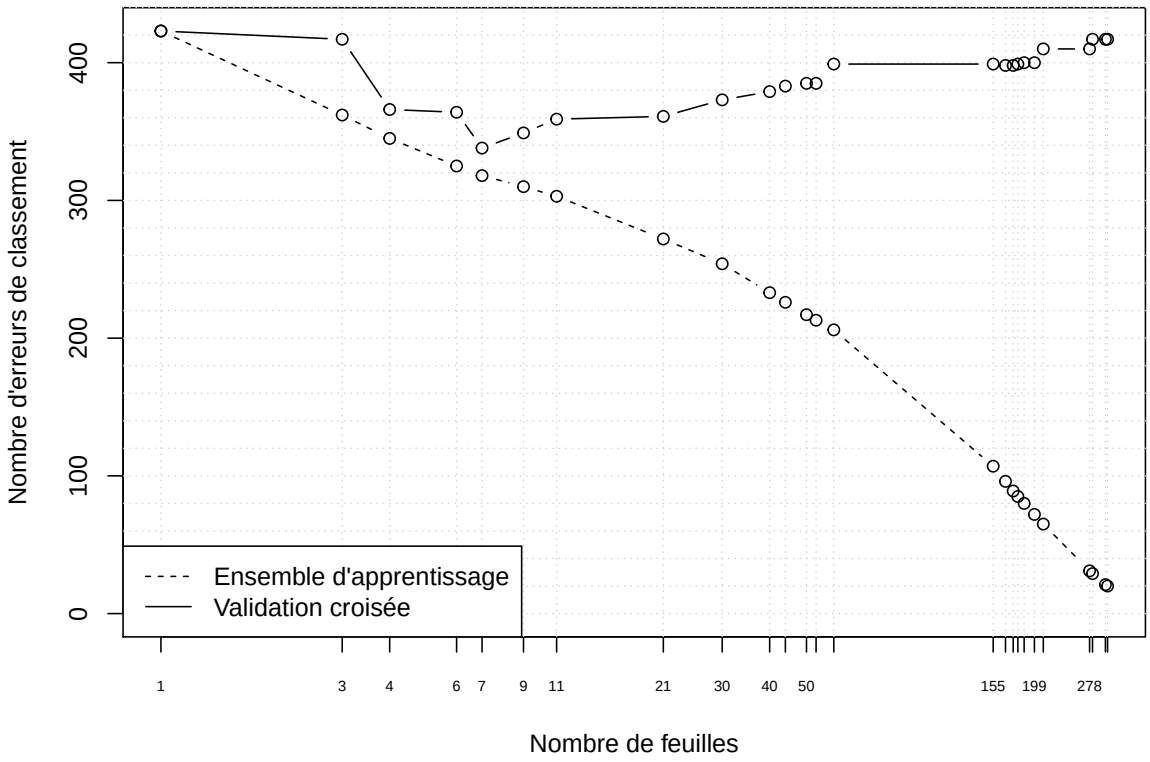

FIGURE 1 – Évolution du nombre d'erreurs de classement en fonction du nombre de feuilles dans l'arbre, sur l'ensemble A et estimé par validation croisée. L'axe des x utilise une échelle logarithmique.

Question 3 La figure 1 représente le nombre d'erreurs de classement commises par l'arbre en fonction du nombre de feuilles conservées, sur l'ensemble d'apprentissage (ligne en tirets) et grâce à une procédure de validation croisée à 10 blocs (ligne pleine). L'analyse décide de conserver un arbre à 7 feuilles. En quel sens cet arbre est il optimal ?

Question 4 La figure 2 représente l'arbre optimal. Construire un exemple fictif d'observation que l'arbre classera dans la deuxième feuille en partant de la gauche dans l'ensemble des 4 feuilles situées en bas de la figure. On donnera pour cette observation les valeurs des variables utilisées par l'arbre en respectant la nature de variables (cf le début de l'exercice).

Question 5 Déduire de la figure 2 la matrice de confusion de l'arbre optimal qu'elle représente en indiquant sur quel ensemble cette matrice est obtenue. Commenter cette matrice.

L'arbre optimal a la matrice de confusion suivante sur l'ensemble B :

	1	2	3
1	210	42	73
2	16	41	19
3	88	83	163

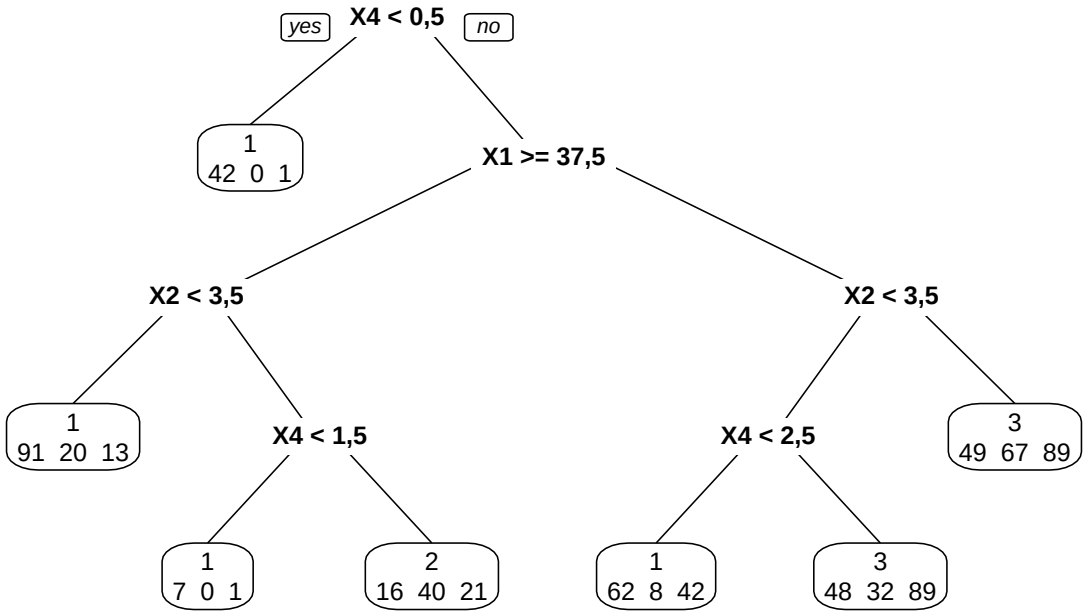

FIGURE 2 – Arbre de classification optimal. La branche de gauche de chaque nœud correspond toujours à la réponse « oui » à la question du nœud, l'autre branche à la réponse « non ». La lettre indiquée sur la première ligne de chaque feuille donne la classe associée à la feuille. Les nombres de la deuxième ligne d'une feuille indiquent les effectifs des trois classes dans la feuille, dans l'ordre 1, 2 et 3.

Question 6 Commenter les résultats obtenus, en comparant notamment avec ceux de l'arbre complet.

L'analyste reproduit la démarche ci-dessus en représentant les variables ordinales X_2, X_3 et X_8 sous forme nominale plutôt que numérique, c'est-à-dire en considérant les valeurs 1, 2, 3 et 4 comme des symboles abstraits plutôt que des nombres.

Question 7 Décrire avec précision les questions binaires qui sont possibles sur la variable X_2 considérée comme numériques et celles qui sont possibles quand elle est considérée comme nominale.

Question 8 Indiquer les avantages et les inconvénients des deux représentations (dans le contexte des arbres de décision).

L'analyste constate que les résultats obtenus ne sont pas meilleurs qu'avec le codage initial. Elle décide alors de tester un classifieur bayésien naïf (CBN) avec le codage nominal.

Question 9 En tenant compte des caractéristiques de X_1 et X_4, comment pourrait-on modéliser leur lois conditionnelles ?

Après construction du CBN sur l'ensemble A, on obtient les matrices de confusion suivantes :

	1	2	3
1	157	29	65
2	78	98	78
3	80	40	113

Matrice de confusion sur A

	1	2	3
1	143	26	79
2	75	89	67
3	96	51	109

Matrice de confusion sur B

Question 10 Commenter les résultats. Comparer notamment les résultats du CBN et de l'arbre optimal en justifiant la pertinence de la méthode de comparaison employée.

Question 11 Quelle(s) autre(s) méthode(s) pourrait-on utiliser pour construire un modèle sur ces données ?

Exercice 4

On étudie un jeu de données comportant 972 observations. Chaque observation correspond à des mesures d'expression de protéines chez des souris. Une observation est décrite par 68 variables numériques et une variable de classe qui prend deux valeurs, Control et Ts65Dn.

L'analyste souhaite classer automatiquement les souris dans les deux classes. Elle teste une SVM (avec un noyau linéaire) sur les données. La figure 3 représente l'estimation des performances de la SVM par une validation croisée à 5 blocs en fonction du paramètre de régularisation.

Question 1 Quelle valeur faut-il choisir pour C pour garantir de bonnes performances ?

	Control	Ts65Dn
Control	511	12
Ts65Dn	3	446

TABLE 1 – Matrice de confusion de la SVM choisie (prédictions en ligne, vraies valeurs en colonne)

La table 1 donne la matrice de confusion de la SVM retenue par l'analyste, évaluée sur l'intégralité des données.

Question 2 Les performances obtenues sont elles compatibles avec celles estimées par validation croisée ?

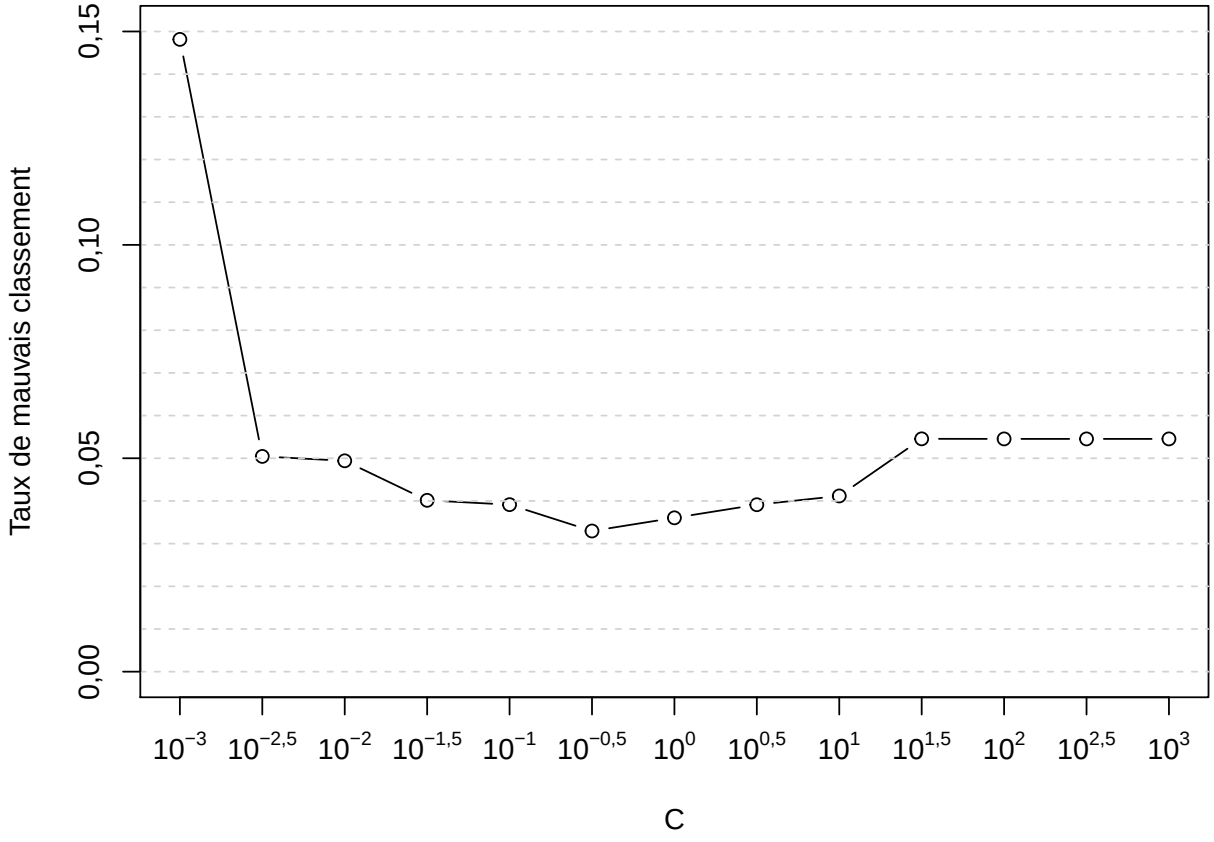

FIGURE 3 – Estimation des performances d'une SVM linéaire en fonction du paramètre de régularisation C par validation croisée