Université de Montpellier Année 2016-2017
HMMA303 - Discrimination et scoring

Examen final - 9 novembre 2016
Durée 2h00 - Documents interdits

Exercice 1 (6 pts) On se place dans le contexte de la classification supervisée binaire.

1) (2 pts) Expliquer les étapes de construction d'un arbre de classification (CART).

2) (2 pts) Expliquer les étapes de construction d'une forêt aléatoire (random forest).

3) (2 pts) Supposons que le jeu de données `d` contient le facteur binaire `y` et un ensemble de variables explicatives. Donner le code `R` permettant de comparer la qualité des classifieurs CART et random forest.

Exercice 2 (8 pts) Soient Y un facteur binaire à expliquer $Y \in \{0,1\}$, X_1 un facteur explicatif prenant les modalités `a` et `b` et X_2 un deuxième facteur explicatif prenant les modalités `c` et `d`. Nous disposons d'un échantillon de 100 observations dont les résultats sont synthétisés dans les deux tableaux ci-dessous

— lorsque $y = 0$

	$x_1 = a$	$x_1 = b$
$x_2 = c$	15	5
$x_2 = d$	5	25

— lorsque $y = 1$

	$x_1 = a$	$x_1 = b$
$x_2 = c$	0	30
$x_2 = d$	10	10

1) (2 pts) Donner le classifieur obtenu à l'aide de la méthode de l'analyse discriminante non paramétrique.

2) (2 pts) Proposer un classifieur alternatif dont la construction n'est pas basée sur une hypothèse d'indépendance conditionnelle.

3) (4 pts) Nous supposons que
$\mathbb{P}(Y = 0) = 1/2$, $\mathbb{P}((X_1, X_2) = (i,j)|Y = 0) = \mathbb{P}(X_1 = i|Y = 0)\mathbb{P}(X_2 = j|Y = 0)$,
$\mathbb{P}(X_1 = a|Y = 0) = \mathbb{P}(X_2 = c|Y = 0) = 1/3$ et enfin que

$$\mathbb{P}((X_1, X_2) = (i,j)|Y = 1) = \begin{cases} 0 & (i,j) = (a,c) \\ 2/3 & (i,j) = (b,c) \\ 1/6 & (i,j) = (a,d) \\ 1/6 & (i,j) = (b,d) \end{cases}.$$

Pour la fonction de coût élémentaire $h(y,d) = \mathbf{1}_{y \neq d}$, donner le coût moyen des deux classifieurs mis en évidence questions 1) et 2).

Exercice 3 (6 pts) Commenter en détails le fichier R Markdown fourni en annexes. Les données sont celles du challenge kaggle sur le naufrage du Titanic.

Correction examen final 09/11/2016 HMMA303

Exercice 1

Voir le cours

Exercice 2

1) La technique de l'analyse discriminante non paramétrique consiste à estimer les lois $(X_1, X_2)|Y=0$ et $(X_2, X_2)|Y=1$ sous les hypothèses $X_1 \perp\!\!\!\perp X_2 \,|Y=0$ $X_1 \perp\!\!\!\perp X_2 \,|Y=1$

2) D'après la table de contingence pour $y=0$, nous avons

$$\widehat{\mathbb{P}}(X_1=a\,|Y=0) = \frac{2}{5} \text{ et } \widehat{\mathbb{P}}(X_2=c\,|Y=0) = \frac{2}{5}$$

Nous obtenons,

$$\widehat{\mathbb{P}}((X_1,X_2)=(i,j)\,|\,Y=0) = \begin{cases} \frac{4}{25} & (i,j)=(a,c) \\ \frac{6}{25} & (i,j)=(b,c) \\ \frac{6}{25} & (i,j)=(a,d) \\ \frac{9}{25} & (i,j)=(b,d) \end{cases}$$

De même,

$$\widehat{\mathbb{P}}(X_1=a\,|\,Y=1)=\frac{1}{5} \quad \text{et} \quad \widehat{\mathbb{P}}(X_2=c\,|\,Y=1)=\frac{3}{5}$$

$$\widehat{\mathbb{P}}((X_1,X_2)=(i,j)\,|\,Y=1) = \begin{cases} \frac{3}{25} & (i,j)=(a,c) \\ \frac{12}{25} & (i,j)=(b,c) \\ \frac{2}{25} & (i,j)=(a,d) \\ \frac{8}{25} & (i,j)=(b,d) \end{cases}$$

Par ailleurs,

$$\widehat{\mathbb{P}}(Y=0)=\frac{1}{2}=\widehat{\mathbb{P}}(Y=1)$$

Nous obtenons,

$$\widehat{\mathbb{P}}(Y=0 \mid (X_1,X_2)=(a,c)) = \frac{\widehat{\mathbb{P}}(Y=0)\,\widehat{\mathbb{P}}((X_1,X_2)=(a,c) \mid Y=0)}{\widehat{\mathbb{P}}(Y=0)\,\widehat{\mathbb{P}}((X_1,X_2)=(a,c) \mid Y=0) + \widehat{\mathbb{P}}(Y=1)\,\widehat{\mathbb{P}}((X_1,X_2)=(a,c) \mid Y=1)}$$

$$= \frac{\frac{4}{25}}{\frac{4}{25}+\frac{3}{25}} = \frac{4}{7}$$

De même,

$$\widehat{\mathbb{P}}(Y=0 \mid (X_1,X_2)=(b,c)) = \frac{\frac{6}{25}}{\frac{6}{25}+\frac{12}{25}} = \frac{1}{3}$$

$$\widehat{\mathbb{P}}(Y=0 \mid (X_1,X_2)=(a,nl)) = \frac{\frac{6}{25}}{\frac{6}{25}+\frac{2}{25}} = \frac{3}{4}$$

$$\widehat{\mathbb{P}}(Y=0 \mid (X_1,X_2)=(b,nl)) = \frac{\frac{9}{25}}{\frac{9}{25}+\frac{8}{25}} = \frac{9}{17}$$

On en déduit la classifieur

$$g_1(n_1,n_2) = \begin{cases} 0 & (n_1,n_2)=(a,c) \\ 1 & (n_1,n_2)=(b,c) \\ 0 & (n_1,n_2)=(a,nl) \\ 0 & (n_1,n_2)=(b,nl) \end{cases}$$

2] On peut estimer les lois $(X_1, X_2)|Y=0$ et $(X_1, X_2)|Y=1$ sans utiliser l'hypothèse d'indépendance conditionnelle.

$$\widehat{\mathbb{P}}((X_1,X_2)=(a,c)|Y=0) = \frac{15}{50} = \frac{3}{10}$$
$$\widehat{\mathbb{P}}((X_1,X_2)=(b,c)|Y=0) = \frac{5}{50} = \frac{1}{10}$$
$$\widehat{\mathbb{P}}((X_1,X_2)=(a,d)|Y=0) = \frac{5}{50} = \frac{1}{10}$$
$$\widehat{\mathbb{P}}((X_1,X_2)=(b,d)|Y=0) = \frac{25}{50} = \frac{1}{2}$$

De même,

$$\widehat{\mathbb{P}}((X_1,X_2)=(i,j)|Y=1) = \begin{cases} 0 & (i,j)=(a,c) \\ \frac{3}{5} & (i,j)=(b,c) \\ \frac{1}{5} & (i,j)=(a,d) \\ \frac{1}{5} & (i,j)=(b,d) \end{cases}$$

Comme, par ailleurs,

$$\widehat{\mathbb{P}}(Y=1) = \widehat{\mathbb{P}}(Y=0) = \frac{1}{2}$$

mieux obtenons

$$\widehat{\mathbb{P}}(Y=0 \mid (X_1,X_2)=(i,j)) = \begin{cases} 1 & (i,j)=(1,C) \\ \frac{1}{7} & (i,j)=(1,C) \\ \frac{1}{3} & (i,j)=(1,1) \\ \frac{5}{7} & (i,j)=(1,1) \end{cases}$$

On en déduit la décision

$$g_2(x_1,x_2) = \begin{cases} 0 & (x_1,x_2)=(0,C) \\ 1 & (x_1,x_2)=(1,C) \\ 1 & (x_1,x_2)=(0,1) \\ 0 & (x_1,x_2)=(1,1) \end{cases}$$

3] $C(g_1) = \mathbb{E}[h(Y, g_1(X))] \quad X=(X_1,X_2)$

$$C(g_1) = \mathbb{P}(Y \neq g_1(X))$$

$$C(g_1) = \mathbb{P}(\{Y=0\} \cap \{g_1(X)=1\}) + \mathbb{P}(\{Y=1\} \cap \{g_1(X)=0\})$$

$$C(g_1) = \mathbb{P}\left(\{Y=0\} \cap \{ (X_1,X_2)=(r,c) \} \right)$$

$$+\mathbb{P}\left(\{Y=1\} \cap \{ (X_1,X_2)=(w,c) \cup (X_1,X_2)=(w,d) \cup (X_1,X_2)=(r,d) \} \right)$$

$$C(g_1) = \mathbb{P}((X_1,X_2)=(r,c) \mid Y=0)\, \mathbb{P}(Y=0)$$

$$+\mathbb{P}((X_1,X_2)=(w,c) \mid Y=1)\, \mathbb{P}(Y=1)$$

$$+\mathbb{P}((X_1,X_2)=(w,d) \mid Y=1)\, \mathbb{P}(Y=1)$$

$$+\mathbb{P}((X_1,X_2)=(r,d) \mid Y=1)\, \mathbb{P}(Y=1)$$

$$= \frac{2}{9} \times \frac{1}{2} + \frac{1}{2}\left[0 + \frac{1}{6} + \frac{1}{6} \right]$$

$$= \frac{1}{2}\left[\frac{2}{9} + \frac{2}{6} \right] = \left(\frac{15}{54} \right)$$

De même

$$C(g_2) = \mathbb{P}\left(\{Y=0\} \cap \{ (X_1,X_2)=(r,c) \cup (X_1,X_2)=(w,d) \} \right)$$

$$+\mathbb{P}\left(\{Y=1\} \cap \{ (X_1,X_2)=(w,c) \cup (X_1,X_2)=(r,r) \} \right)$$

$$C(g_2) = \frac{1}{2}\left[\frac{2}{9} + \frac{1}{4}\right] + \frac{1}{2}\left[0 + \frac{1}{6}\right]$$

$$C(g_2) = \left[\frac{1}{9} + \frac{1}{8} + \frac{1}{12}\right]$$

$$= \left[\frac{8}{72} + \frac{9}{72} + \frac{6}{72}\right] = \frac{23}{72}$$

Au final, $C(g_1) < C(g_2)$

$\#\ 0{,}28 \qquad \#\ 0.32$

Exercice 3

Dans une première partie, les données sont pré-traitées avec notamment la création de la variable "title" qui contient le titre des passagers. Cette information est extraite de leurs noms.

Dans un second temps, les données manquantes imputées. Sur l'âge, il est utilisé l'âge médian correspondant au titre, ie calculé sur la base de données fusionnée (apprentissage et test) pour le chef de titre de l'individu pour lequel l'âge est manquant.

Dans une dernière partie, un modèle de classification par vote, dont la première étape est fixé par validation croisée est mis en œuvre. Enfin, un modèle de forêt aléatoire est mis en œuvre.