Université de Montpellier Année 2015-2016
HMMA303 - Discrimination et scoring

Examen de rattrapage - 13 janvier 2016
Durée 1h30 - Documents interdits

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

1) (2 pts) Décrire le modèle de la régression logistique.

2) (2 pts) Donner trois critères de sélection entre deux modèles de régression logistique.

3) (2 pts) Donner deux stratégies de sélection de variables basées sur les critères précédents.

Exercice 2 (6 pts) Nous considérons un jeu de données contenu dans le data.frame `spam` de la bibliothèque `kernlab`.

```
A data frame with 4601 observations and 58 variables.

The first 48 variables contain the frequency of the variable name
(e.g., business) in the e-mail. If the variable name starts with num
(e.g., num650) the it indicates the frequency of the corresponding
number (e.g., 650). The variables 49-54 indicate the frequency of
the characters ';', '(', '[', '!', '\$', and '\#'. The variables 55-57
contain the average, longest and total run-length of capital letters.
Variable 58 indicates the type of the mail and is either ``nonspam''
or ``spam'',  i.e. unsolicited commercial e-mail.

The data set contains 2788 e-mails classified as ``nonspam'' and
1813 classified as ``spam''.
```

Donner le code `R` comparant les classifieurs produits par les méthodes suivantes :
- classification par regression,
- analyse discriminante quadratique,
- forêts aléatoires.

Exercice 3 (8 pts) Commenter en détails le fichier R Markdown fourni en annexes.

HMMA 303
Discrimination et scoring - Correction examen mathématique 13/01/2016

Exercice 1

1) $Y \in \{0,1\}$ et $x \in \mathbb{R}^{d}$

$$\mathbb{P}(Y=1 \mid X=x) = \frac{e^{x^T\beta}}{1+e^{x^T\beta}}$$

où $\beta \in \mathbb{R}^{d}$ est un paramètre inconnu.

2] Le modèle de régression est (1)
utilisé dans le contexte de
la classification supervisée
binaire. On peut donc
utiliser le critère de
l'erreur de classification,
erreur que l'on pourra
estimer par une technique
de validation croisée.
Deux autres critères sont
basés sur des versions pénalisées
de la fonction de vraisemblance,
il s'agit des critères AIC et
BIC.

3] On peut utiliser une méthode de sélection ascendante ou une stratégie descendante. (3)

Exercice 2

```
library(kernlab) ; library(MASS)
data(spam) ; library(randomForest)
model.rf <- randomForest(type ~., data=spam)
model.rf $ err.rate [500,1]
K <- 10 ; N <- 4600
g <- N/K ; res.loo <- (g,N)
res.reg <- rep(0,N) ; index <- sample(N)
for(i in 1:K)
{
}
```

```
ent <- indi[((K-1)*g+1):(K*g)]
model.lda <- lda(type~., data = spam[-indi,])
res.lda[indi] <- predict(model.lda, spam[indi,])$posterior[,2]
model.reg <- lm(as.numeric(type)-1~., data = spam[-indi,])
res.reg[indi] <- predict(model.reg, spam[indi,])
}

mean(round(res.lda) != (as.numeric(type)-1))
mean(as.numeric(res.reg >= 1/2) != (as.numeric(type)-1))
```

Dans le code présenté ci-dessus, nous avons calculé une seule fois l'erreur par validation croisée à K=10 ensembles.

Exercice 3

5

On enlève 500 observations de notre échantillon pour les utiliser comme données de validation. Avec ces données nous pouvons évaluer et comparer différentes méthodes de classification. On le fait pour LDA, les forêts aléatoires et la classification bayésienne naïf. Les forêts aléatoires donnent les meilleurs résultats.

Puis, on évalue par validation croisée à 10 - ensemble le nombre de variables optimal pour la méthode

des k-plus-proches voisins.

On trouve $k = 1$.