Université de Montpellier Année 2018-2019
M2 MIND - Discrimination et scoring

Examen partiel - 19 octobre 2018
Durée 1h30 - Documents interdits

Exercice 1 (4 pts) On se place dans le contexte de la classification supervisée binaire.

1) (2 pts) Expliquer ce qu'est une courbe ROC ? Que compare-t-on avec cet objet géométrique ? Comment choisir un classifieur avec cet outils ?

2) (2 pts) Soit $(x_1, \ldots, x_n)$ un n-échantillon de loi inconnue F.
Proposer un estimateur $\hat{f}$ de la densité de f de F.

Exercice 2 (10 pts) Soit Y un facteur binaire à expliquer, $Y \in \{0, 1\}$ et $X \in [-2, 3]$ une variable explicative réelle. Nous supposons que $\mathbb{P}(Y = 1) = 1/3$ et

$$X|Y = 0 \sim \mathcal{U}_{[-2,1]}$$

$$X|Y = 1 \sim \mathcal{U}_{[0,3]}\,.$$

1) (2 pts) Donner la loi de $Y|X = x$.

2) (2 pts) Pour la fonction de coût élémentaire

$$h(y, d) = \begin{cases} 0 & \text{si} \quad y = d \\ 2 & \text{si} \quad y = 1 \quad \text{et} \quad d = 0 \\ 1 & \text{si} \quad y = 0 \quad \text{et} \quad d = 1 \end{cases}$$

donner le classifieur optimal g^*.

3) (2 pts) Calculer le taux d'erreur moyen associé à g^*.

4) (2 pts) Nous supposons que nous observons n réalisations indépendantes $(x_1, y_1), \ldots, (x_n, y_n)$ du couple (X, Y). Expliciter $g^{\#}$ le classifieur correspondant à l'analyse discriminante linéaire et le comparer à g^*.

5) (2 pts) Les données de la question précédente sont stockées dans un `data.frame` R nommé `data.train`, objet contenant `y` le vecteur binaire à expliquer et `x` la valeur variable explicative quantitative.
Donner le code R qui
a) met en oeuvre la technique de l'analyse discriminante linéaire
b) évalue ses performances en terme de courbe ROC.

Exercice 3 (6 pts) Expliquer en détails le résultat produit par le code `R` donné ci-dessous.

```
> voice <- read.csv("voice.csv")
> index <- createDataPartition(voice$label, p = 0.75, list = FALSE)
> test <- voice[-index, ]
> train <- voice[index, ]

> control <- trainControl(method="repeatedcv", number=10,  repeats=100)
> metric <- "Accuracy"

> model1 <- train(label~., data=train, method="lda", metric=metric, trControl=control)
> prediction1 <- predict(model1, test)
> confusionMatrix(prediction1, test$label)

Confusion Matrix and Statistics

Reference
Prediction female male
female    371    4
male       25  392

Accuracy : 0.9634
95% CI : (0.9478, 0.9753)
No Information Rate : 0.5
P-Value [Acc > NIR] : < 2.2e-16

Kappa : 0.9268
Mcnemar's Test P-Value : 0.0002041

            Sensitivity : 0.9369
            Specificity : 0.9899
         Pos Pred Value : 0.9893
         Neg Pred Value : 0.9400
             Prevalence : 0.5000
         Detection Rate : 0.4684
   Detection Prevalence : 0.4735
      Balanced Accuracy : 0.9634

       'Positive' Class : female


> model2 <- train(label~., data=train, method="knn", metric=metric, trControl=control,
+ tuneGrid=data.frame(k=21))
> prediction2 <- predict(model2, test)
> confusionMatrix(prediction2, test$label)
Confusion Matrix and Statistics

          Reference
Prediction female male
    female    257  111
    male      139  285
```

```
               Accuracy : 0.6843
                 95% CI : (0.6507, 0.7166)
    No Information Rate : 0.5
    P-Value [Acc > NIR] : < 2e-16

                  Kappa : 0.3687
 Mcnemar's Test P-Value : 0.08771

            Sensitivity : 0.6490
            Specificity : 0.7197
         Pos Pred Value : 0.6984
         Neg Pred Value : 0.6722
             Prevalence : 0.5000
         Detection Rate : 0.3245
   Detection Prevalence : 0.4646
      Balanced Accuracy : 0.6843

       'Positive' Class : female

> res <- resamples(model1=model1, model2=model2)
> summary(res)

Call:
summary.resamples(object = res)

Models: model1, model2
Number of resamples: 1000

Accuracy
            Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
model1 0.9285714 0.9621849 0.9704641 0.9690826 0.9747899 1.0000000
model2 0.5991561 0.6733638 0.6919831 0.6918967 0.7100840 0.7754237
       NA's
model1    0
model2    0

Kappa
            Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
model1 0.8571429 0.9243697 0.9409335 0.9381650 0.9495798 1.0000000
model2 0.1982694 0.3467277 0.3837358 0.3837888 0.4201681 0.5508475
       NA's
model1    0
model2    0
```

Correction examen partiel HMMA303 19/10/2018

Exercice 1

1] Voir cours

2] $\hat{f}_n(x) = \frac{1}{2\varepsilon n} \sum_{k=1}^{n} \mathbb{1}_{[x-\varepsilon, x+\varepsilon]}(x_k)$

$\hat{f}_n(x)$ est l'estimateur de $f(x)$ par une méthode à noyau avec noyau uniforme associé à une taille de fenêtre $[2\varepsilon]$.

Exercice 2

$Y \in \{0,1\}$, $\mathbb{P}(Y=1) = \frac{1}{3}$

$X | Y=0 \sim \mathcal{U}_{[-2,1]}$, $X | Y=1 \sim \mathcal{U}_{[0,3]}$

1] $\mathbb{P}(Y=1|X=x) = \dfrac{\mathbb{P}(Y=1)\,\Pi_{[0,3]}(x)\left(\frac{1}{3}\right)}{\mathbb{P}(Y=0)\,\Pi_{[-2,1]}(x)\left(\frac{1}{3}\right) + \mathbb{P}(Y=1)\,\Pi_{[0,3]}(x)\left(\frac{1}{3}\right)}$

$\Leftrightarrow \mathbb{P}(Y=1|X=x) = \begin{cases} 0 & \text{si } x \in [-2,0] \\ \frac{1}{3} & \text{si } x \in [0,1] \\ 1 & \text{si } x \in [1,3] \end{cases}$

On en déduit que

$\mathbb{P}(Y=0|X=x) = \begin{cases} 0 & \text{si } x \in [1,3] \\ \frac{2}{3} & \text{si } x \in [0,1] \\ 1 & \text{si } x \in [-2,0] \end{cases}$

2] $h(y,u) = \begin{cases} 0 & \text{si } y = u \\ 2 & \text{si } y = 1 \text{ et } u = 0 \\ 1 & \text{si } y = 0 \text{ et } u = 1 \end{cases}$

Le classifieur optimal est celui qui minimise le coût moyen

$\mathbb{E}[h(Y, g(X))] = 2\,\mathbb{P}(Y=1 \cap g(X)=0) + \mathbb{P}(Y=0 \cap g(X)=1)$

Sur ailleurs,

$$\mathbb{E}\left[\ell(Y, g(X))\right] = \int \left[2\,\Pi_{\{g(x)=0\}}\, \mathbb{P}(Y=1 \mid X=x) + \Pi_{\{g(x)=1\}}\, \mathbb{P}(Y=0 \mid X=x) \right] p_X(x)\, \mu(dx)$$

$$= \int q(x)\, p_X(x)\, \mu(dx)$$

Ainsi si l'on minimise $q(x)$ pour tout x, on minimise $\mathbb{E}\left[\ell(Y, g(X))\right]$

$$q(x) = 2\,\Pi_{\{g(x)=0\}}\, \mathbb{P}(Y=1 \mid X=x) + \Pi_{\{g(x)=1\}}\, \mathbb{P}(Y=0 \mid X=x)$$

$$g^*(x) = \begin{cases} 0 & \text{si } 2\,\mathbb{P}(Y=1 \mid X=x) \leq \mathbb{P}(Y=0 \mid X=x) \\ 1 & \text{si } 2\,\mathbb{P}(Y=1 \mid X=x) \geq \mathbb{P}(Y=0 \mid X=x) \end{cases}$$

<u>Remarque</u> : cet optimum n'est pas unique, un autre minimum est tel que $g^*(x) = 1$ si $2\,\mathbb{P}(Y=1 \mid X=x) = \mathbb{P}(Y=0 \mid X=x)$ toutes choses égales par ailleurs.

$$g^*(x) = \begin{cases} 0 & \text{si } \mathbb{P}(Y=0 \mid X=x) \geq \frac{2}{3} \\ 1 & \text{si } \mathbb{P}(Y=0 \mid X=x) < \frac{2}{3} \end{cases}$$

Ainsi,

$$g^*(x) = \begin{cases} 0 & \text{si } x \in [-2,1] \\ 1 & \text{si } x \in [1,3] \end{cases}$$

3] On veut calculer

$$\mathbb{P}(Y \neq g^*(X))$$

$$= \mathbb{P}(Y=1 \cap X \in [-2,1]) + \mathbb{P}(Y=0 \cap X \in [1,3])$$

$$= \mathbb{P}(X \in [-2,1] \mid Y=1)\,\mathbb{P}(Y=1) + \mathbb{P}(X \in [1,3] \mid Y=0)\,\mathbb{P}(Y=0)$$

$$\Leftrightarrow \mathbb{P}(Y \neq g^*(X))$$

$$= \left(\frac{1}{3}\right) \mathbb{P}(X \in [-2,1] \mid Y=1)$$

$$+ \left(\frac{2}{3}\right) \mathbb{P}(X \in [1,3] \mid Y=0)$$

$$= \left(\frac{1}{3}\right) \int_0^1 \left(\frac{1}{3}\right) dx + \left(\frac{2}{3}\right) \times 0$$

$$= \left(\frac{1}{3}\right)\left(\frac{1}{3}\right) [x]_0^1 = \left(\frac{1}{9}\right)$$

4] La méthode de l'analyse discriminante linéaire permet d'estimer $\mathbb{P}(Y=1 \mid X=x)$, par

$$\widehat{\mathbb{P}}(Y=1 \mid X=x)$$

$$= \frac{\widehat{\pi}_1 \, e^{-\frac{1}{2\hat{\sigma}^2}(x-\hat{\mu}_1)^2}}{\sum_{k=0}^{1} \widehat{\pi}_0 \, e^{-\frac{1}{2\hat{\sigma}^2}(x-\hat{\mu}_k)^2}}$$

on $\hat{\pi}_k = \frac{M_k}{m}$, $M_k = \sum_{i=1}^{m} \mathbb{1}\{y_i = k\}$

$$\hat{\mu}_k = \frac{1}{M_k} \sum_{i=1}^{m} \mathbb{1}\{y_i = k\} x_i$$

$$\hat{\sigma}^2 = \frac{1}{m-2} \sum_{i=1}^{m} (x_i - \hat{\mu}_{y_i})^2$$

En suivant la classification de Bayes, on a :

$$g^{\#}(x) = \begin{cases} 0 \text{ si } \hat{\mathbb{P}}(Y=1 \mid X=x) \leq \frac{1}{3} \\ 1 \text{ si } \hat{\mathbb{P}}(Y=1 \mid X=x) > \frac{1}{3} \end{cases}$$

$$\Rightarrow g^{\#}(x) = 0 \text{ si}$$

$$3\, \hat{\pi}_1 e^{-\frac{1}{2\hat{\sigma}^2}(x-\hat{\mu}_1)^2} \leq \sum_{k=0} \hat{\pi}_k e^{-\frac{1}{2\hat{\sigma}^2}(x-\hat{\mu}_k)^2}$$

$$\Rightarrow g^{\#}(x) = 0 \text{ si}$$

$$2\, \hat{\pi}_1 e^{-\frac{1}{2\hat{\sigma}^2}(x-\hat{\mu}_1)^2} \leq \hat{\pi}_0 e^{-\frac{1}{2\hat{\sigma}^2}(x-\hat{\mu}_0)^2}$$

$$\Rightarrow g^{\#}(x) = 0 \qquad x$$

$$-\frac{1}{2\hat{\sigma}^2}(x-\hat{\mu}_1)^2 \leq \log\left(\frac{n_0}{n_1}\right) - \log(2) - \frac{1}{2\hat{\sigma}^2}(x-\hat{\mu}_0)^2$$

$$\Rightarrow g^{\#}(x) = 0 \qquad x$$

$$x\left[\frac{-\hat{\mu}_0+\hat{\mu}_1}{\hat{\sigma}^2}\right] \leq \log\left(\frac{n_0}{n_1}\right) - \log(2) + \left[\frac{-\hat{\mu}_0^2+\hat{\mu}_1^2}{2\hat{\sigma}^2}\right]$$

$$\Rightarrow g^{\#}(x) = 0 \qquad x$$

$$x \leq \left(\frac{\hat{\sigma}^2}{-\hat{\mu}_0+\hat{\mu}_1}\right)\left[\log\left(\frac{n_0}{n_1}\right) - \log(2) + \left[\frac{-\hat{\mu}_0^2+\hat{\mu}_1^2}{2\hat{\sigma}^2}\right]\right]$$

$$\hat{\mu}_0 \xrightarrow[n\to\infty]{} \mu_0 = -\frac{1}{2} \qquad \hat{\mu}_1 \xrightarrow[n\to\infty]{} \mu_1 = \frac{3}{2}$$

$$\frac{n_0}{n_1} \xrightarrow[n\to\infty]{} \frac{\pi_0}{\pi_1} = 2 \qquad \hat{\sigma}^2 \xrightarrow[n\to\infty]{} \frac{9}{12} = \frac{3}{4}$$

Donc asymptotiquement, nous avons $g^{\#}(n) = 0$ si

$$n \leq \frac{\left(\frac{3}{4}\right)}{(2)} \left\{ \frac{\left(\frac{-1}{4} + \frac{9}{4}\right)}{2\left(\frac{3}{4}\right)} \right\}$$

$$\Rightarrow g^{\#}(n) = 0 \text{ si}$$

$$n \leq \frac{\left(\frac{8}{4}\right)}{(4)}$$

$$\Rightarrow g^{\#}(n) = 0 \text{ si}$$

$$n \leq \frac{1}{2}$$

5] Voir (1)

Exercice 3

On compare les performances des classifieurs LDA et des classifieurs k-plus-proches-voisins, avec $k = 21$ voisins, sur les données de reconnaissance de parole d'enregistrements vocaux.

Pour ce faire, on met en œuvre une stratégie de validation croisée à 10 ensembles répétée 100 fois.

À l'aide de la fonction `resample`, on analyse la variabilité des résultats, sur les $10 \times 100 = 1000$ pour-échantillons produits.