Université de Montpellier Année 2020-2021
HMMA303

Examen de contrôle continu
Durée 1h30 - Sans document - 09 octobre 2020

Exercice 1 (3 pts) On se place dans le contexte de la classification supervisée binaire. Qu'est qu'une courbe ROC ? Que permet-elle de faire ?

Exercice 2 (3 pts) On se place dans le contexte de la classification supervisée binaire. On suppose que l'on dispose d'une seule variable explicative réelle X. La variance empirique de X est égale 1, la moyenne des X pour $Y = 0$ est égale à 0 et la moyenne des X pour $Y = 1$ est égale à 2. Par ailleurs, pour les données observées, les deux classes sont équilibrées. Donner l'expression explicite du classifieur LDA.

Exercice 3 (8 pts) Soit Y un facteur binaire à expliquer, $Y \in \{0, 1\}$ et X une variable explicative réelle. Nous supposons que $\mathbb{P}(Y = 1) = 1/3$ et

$$X|Y = 1 \sim f_1(x) = 2\exp(-2x)\mathbb{I}_{\mathbb{R}_+}(x)$$

$$X|Y = 0 \sim f_0(x) = \frac{1}{2}\exp(-x/2)\mathbb{I}_{\mathbb{R}_+}(x)\,.$$

1) (0.5 pt) Donner la loi marginale de X.

2) (0.5 pt) Donner la loi de Y sachant $X = x$.

3) (2 pts) Donner le classifieur optimal g^* lorsque

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

4) (2 pts) Calculer le taux d'erreur moyen associé à g^*.

5) (3 pts) Donner deux classifieurs aléatoires différents et calculer leurs taux d'erreur moyens. Comparer tous les taux d'erreur.

Exercice 3 (6 pts) Expliquer en détails les résultats produits par le code `R` donné ci-dessous.

```
> cancer.train <- read.csv("~/Desktop/data-cancer-train.csv")
> cancer.train <- cancer.train[,-1]
> cancer.test <- read.csv("~/Desktop/data-cancer-test.csv")
> cancer.test <- cancer.test[,-1]
> cancer.train$diagnosis <- as.factor(cancer.train$diagnosis)
> cancer.test$diagnosis <- as.factor(cancer.test$diagnosis)
```

```
> library(MASS)
> model1 <- qda(diagnosis~.,data=cancer.train)
> library(caret)
> confusionMatrix(predict(model1)$class,cancer.train$diagnosis,positive="M")

Confusion Matrix and Statistics

          Reference
Prediction   B   M
         B 318   5
         M   3 141

               Accuracy : 0.9829
                 95% CI : (0.9665, 0.9926)
    No Information Rate : 0.6874
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.96

 Mcnemar's Test P-Value : 0.7237

            Sensitivity : 0.9658
            Specificity : 0.9907
         Pos Pred Value : 0.9792
         Neg Pred Value : 0.9845
             Prevalence : 0.3126
         Detection Rate : 0.3019
   Detection Prevalence : 0.3084
      Balanced Accuracy : 0.9782

       'Positive' Class : M

> model2 <-train(diagnosis~.,data=cancer.train,method="qda",metric="Accuracy",
    trControl=trainControl(method="repeatedcv",number=10,repeats=100))
> confusionMatrix(model2)

Cross-Validated (10 fold, repeated 100 times) Confusion Matrix

(entries are percentual average cell counts across resamples)

          Reference
Prediction    B    M
         B 66.6  1.5
         M  2.2 29.7

 Accuracy (average) : 0.963
```

```
> dim(cancer.test)
[1] 102  31

> sum(predict(model1,cancer.test)$class==
    predict(model2,cancer.test,type="raw"))
[1] 102

> confusionMatrix(predict(model1,cancer.test)$class,
    cancer.test$diagnosis,positive="M")

Confusion Matrix and Statistics

          Reference
Prediction  B  M
         B 35  5
         M  1 61

               Accuracy : 0.9412
                 95% CI : (0.8764, 0.9781)
    No Information Rate : 0.6471
    P-Value [Acc > NIR] : 2.076e-12

                  Kappa : 0.8744

 Mcnemar's Test P-Value : 0.2207

            Sensitivity : 0.9242
            Specificity : 0.9722
         Pos Pred Value : 0.9839
         Neg Pred Value : 0.8750
             Prevalence : 0.6471
         Detection Rate : 0.5980
   Detection Prevalence : 0.6078
      Balanced Accuracy : 0.9482

       'Positive' Class : M
```

Correction examen de contrôle continu

HMMA 303

09/10/2020

Exercice 1

Voir cours

Exercice 2

$$(Y_1, X_1), \dots, (Y_m, X_m)$$

$$M_0 = \sum_{i=1}^{m} \mathbb{1}\{Y_i = 0\}$$

$$M_1 = \sum_{i=1}^{m} \mathbb{1}\{Y_i = 1\}$$

$$\hat{\mu}_0 = \frac{1}{M_0} \sum_{i=1}^{m} \mathbb{1}\{y_i = 0\} x_i$$

$$\hat{\mu}_1 = \frac{1}{M_1} \sum_{i=1}^{m} \mathbb{1}\{y_i = 1\} x_i$$

$\hat{\mu}_0 = 0$, $\hat{\mu}_1 = 2$, $M_0 = M_1$

$\hat{\pi}_0 = \hat{\pi}_1 = 1/2$

Enfin, $\hat{\sigma}^2 = \left[\frac{1}{m-2}\right] \sum_{i=1}^{m} \mathbb{1}\{y_i = j\} (x_i - \hat{\mu}_j)^2$

$\hat{\sigma}^2 = 1$

$$\hat{P}(Y=1 \mid X=x) = \frac{e^{-\frac{1}{2}(x-2)^2}}{e^{-\frac{1}{2}x^2} + e^{-\frac{1}{2}(x-2)^2}}$$

Le classifieur LDA, en l'absence de fonction de perte est tel que

$$\hat{g}^{LDA}(x) = \begin{cases} 1 \text{ si } \hat{\mathbb{P}}(Y=1/X=x) \geq \frac{1}{2} \\ 0 \text{ sinon} \end{cases}$$ (3)

$$\hat{\mathbb{P}}(Y=1/X=x) \geq \frac{1}{2}$$

$$\Leftrightarrow e^{-\frac{1}{2}x^2} \leq e^{-\frac{1}{2}(x-2)^2}$$

$$\Leftrightarrow x^2 \geq (x-2)^2$$

$$\Leftrightarrow 0 \geq -4x + 4$$

$$\Leftrightarrow 4x \geq 4$$

$$\Leftrightarrow x \geq 1$$

$$\hat{g}^{LDA}(x) = \begin{cases} 1 \text{ si } x \geq 1 \\ 0 \text{ si } x < 1 \end{cases}$$

Exercice 3

1) $f_X(x) = \left\{ \frac{2}{3}\left(\frac{1}{2}e^{-x/2}\right) + \frac{1}{3}\left(2e^{-2x}\right) \right\} \mathbb{1}_{\{x \geq 0\}}$

2) $\mathbb{P}(Y=1 \mid X=x) = \left[\frac{\frac{2}{3}e^{-2x}}{\frac{1}{3}e^{-x/2} + \frac{2}{3}e^{-2x}} \right]$

si $x \geq 0$

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

$$\mathbb{P}(Y=1 \mid X=x) \geq 1/3$$

$$\Leftrightarrow \frac{4}{9}e^{-2x} \geq \frac{1}{9}e^{-x/2}$$

$$\Leftrightarrow 4e^{-2x} \geq e^{-x/2}$$

$$\Leftrightarrow -x/2 \leq -2x + \log(4)$$

$$\Leftrightarrow \quad \frac{3k}{6} \leq \log(4)$$

$$\Leftrightarrow \quad k \leq \frac{2\log(4)}{3}$$

$$g^*(11) = \begin{cases} 1 & \text{si } k \leq \frac{2\log(4)}{3} \\ 0 & \text{sinon} \end{cases}$$

4) Terme d'erreur moyen
≠ Coût moyen

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

$$= \mathbb{P}(\{Y=0\} \cap \{g^*(X)=1\})$$
$$+ \mathbb{P}(\{Y=1\} \cap \{g^*(X)=0\})$$

$$= \frac{2}{3}\,\mathbb{P}(g^*(X)=1 \mid Y=0)$$
$$+ \frac{1}{3}\,\mathbb{P}(g^*(X)=0 \mid Y=1)$$

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

$$= \frac{2}{3}\,\mathbb{P}\left(X \leq \frac{2\log(4)}{3} \mid Y = 0\right)$$

$$+ \frac{1}{3}\,\mathbb{P}\left(X > \frac{2\log(4)}{3} \mid Y = 1\right)$$

$$= \frac{2}{3}\int_0^{2\log(4)/3} \frac{1}{2} e^{-x/2}\,dx$$

$$+ \frac{1}{3}\int_{\frac{2\log(4)}{3}}^{+\infty} 2e^{-2x}\,dx$$

$$= \frac{2}{3}\left[1 - e^{-\log(4)/3}\right]$$

$$+ \frac{1}{3}\left[e^{-4\frac{\log(4)}{3}}\right]$$

$$= \frac{2}{3} - \frac{2(4)^{-1/3}}{3} + \frac{(4)^{-4/3}}{3} \approx 0.299$$

5]

$$g^1(x) = \begin{cases} 1 & u \le 1/2 \\ 0 & \end{cases}$$

$$g^2(x) = \begin{cases} 1 & u \le 1/3 \\ 0 & \end{cases}$$

$$P(Y \neq g^1(x))$$

$$= \frac{2}{3} \times \frac{1}{2} + \frac{1}{3} \times \frac{1}{2}$$

$$= \frac{1}{2}$$

$$P(Y \neq g^2(x)) = \left(\frac{2}{3}\right) P\left(U \le \frac{1}{3} \mid Y=0\right) + \left(\frac{1}{3}\right) P\left(U \ge \frac{1}{3} \mid Y=1\right)$$

$$= \left(\frac{2}{3}\right)\left(\frac{1}{3}\right) + \left(\frac{1}{3}\right)\left(\frac{2}{3}\right)$$

$$= \frac{4}{9} \# 0.444$$

Exercice 3

(3)

Mise en œuvre d'une méthode LDA sur les données d'training

Comparaison de trois matrices de confusion :

- La première sur les données d'training Accuracy : 97,29 %
- La deuxième estimée par validation croisée à 10 ensembles répétée 100 fois Accuracy : 96,3 %
- La troisième sur les données de test Accuracy : 94,12 %

Discussion des intervalles de confiance.