Université de Montpellier Année 2018-2019
HMMA303 - Discrimination et scoring

Examen final - 7 novembre 2018
Sans documents - Durée 2h00

Exercice 1 (8 pts) Nous nous plaçons dans le contexte de la classification binaire. Nous disposons des p variables explicatives quantitatives $x^1, \ldots, x^p$ et d'une variable à expliquer y.

1) (2 pts) Proposer un modèle de perceptron à un couche cachée contenant q neurones et décrire ses composantes.

2) (2 pts) Décrire la construction d'un arbre de classification non élagué.

3) (2 pts) De quoi parle-t-on lorsque l'on évoque l'astuce du noyau dans le cadre de la mise en oeuvre d'un classifieur SVM ?

4) (2 pts) Nous notons $\pi_0 = \mathbb{P}(Y = 0)$, rappeler la définition du classifieur aléatoire **pondéré** et donner son taux d'erreur en fonction π_0.

Exercice 2 (6 pts) Soient Y un facteur binaire à expliquer et X une variable explicative quantitative. Nous avons $X|Y = 0 \sim \mathcal{N}(0, 1)$, $X|Y = 1 \sim \mathcal{N}(0, 2)$, $\mathbb{P}(Y = 0) = 1/2$.

1) (2 pts) Donner l'expression du classifieur optimal g_1 pour la fonction de coût élémentaire $h_1(y, d) = \mathbf{1}_{y \neq d}$.

2) (2 pts) Donner le taux d'erreur moyen de g_1 en fonction de $F_{\mathcal{N}(0,1)}(\cdot)$ fonction de répartition de la loi normale centrée réduite.

3) (2 pts) Donner l'expression du classifieur optimal g_2 pour la fonction de coût élémentaire

$$h_2(y, d) = \begin{cases} 0 & y = d \\ 1 & y = 0 \quad d = 1 \\ 10 & y = 1 \quad d = 0 \end{cases} .$$

Exercice 3 (3 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 svm et forêts aléatoires.

Exercice 4 (3 pts) Commenter les résultats produits par le code `Python` ci-dessous.

```
import pandas as pd
data = pd.read_csv("voice.csv")
data.head()
#    meanfreq        sd    median  ...     dfrange   modindx  label
# 0  0.059781  0.064241  0.032027  ...    0.000000  0.000000   male
# 1  0.066009  0.067310  0.040229  ...    0.046875  0.052632   male
# 2  0.077316  0.083829  0.036718  ...    0.007812  0.046512   male
# 3  0.151228  0.072111  0.158011  ...    0.554688  0.247119   male
# 4  0.135120  0.079146  0.124656  ...    5.476562  0.208274   male

data.label = [1 if each == "male" else 0 for each in data.label]
data.head()
#    meanfreq        sd    median  ...     dfrange   modindx  label
# 0  0.059781  0.064241  0.032027  ...    0.000000  0.000000      1
# 1  0.066009  0.067310  0.040229  ...    0.046875  0.052632      1
# 2  0.077316  0.083829  0.036718  ...    0.007812  0.046512      1
# 3  0.151228  0.072111  0.158011  ...    0.554688  0.247119      1
# 4  0.135120  0.079146  0.124656  ...    5.476562  0.208274      1

gender = data.label.values
features = data.drop(["label"], axis = 1)
features = (features-features.min())/(features.max()-features.min())

gender.mean()
#  0.5

features.head()
#    meanfreq        sd    median    ...       maxdom   dfrange   modindx
# 0  0.096419  0.473409  0.084125    ...     0.000000  0.000000  0.000000
# 1  0.125828  0.505075  0.116900    ...     0.002144  0.002146  0.056449
# 2  0.179222  0.675536  0.102873    ...     0.000357  0.000358  0.049885
# 3  0.528261  0.554611  0.587559    ...     0.025375  0.025393  0.265043
# 4  0.452195  0.627209  0.454272    ...     0.250536  0.250715  0.223380

from sklearn.model_selection import train_test_split
x_train, x_test, y_train, y_test = train_test_split(features, gender,
        test_size = 0.2, random_state = 42)
from sklearn.metrics import confusion_matrix

from sklearn.naive_bayes import GaussianNB
nb = GaussianNB()
nb.fit(x_train, y_train)
y_pred_NB = nb.predict(x_test)
NB_cm = confusion_matrix(y_test, y_pred_NB)
NB_cm.view()
# array([[270,  27],
#        [ 31, 306]])

from sklearn.linear_model import LogisticRegression
lr = LogisticRegression()
lr.fit(x_train, y_train)
y_pred_LR = lr.predict(x_test)
LR_cm = confusion_matrix(y_test, y_pred_LR)
LR_cm.view()
# array([[290,   7],
#        [  5, 332]])
```

Examen final
HMMA 303
Correction
7/11/2018

Exercice 1

1] Voir cours

2] Voir cours

3] On s'aperçoit le fait que les données sont changées dans un espace de très grande dimension et qu'avec la moyenne on calcule des produits scalaires dans cet espace virtuel ; les produits scalaires sont les seules quantités nécessaires pour construire la classification SVM.

4/ Soit $g(\underline{x})$ le classifieur optimal pour $\underline{x} = (x^1, \dots, x^p)$

$$g(\underline{x}) = \begin{cases} 0 & \text{avec proba. } \pi_0 \\ 1 & \text{avec proba. } \pi_1 \end{cases}$$

$\pi_1 = 1 - \pi_0$

Son terme d'erreur est

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

$$\Leftrightarrow \mathbb{P}(Y \neq g(\underline{X})) = \mathbb{P}(Y=1)\,\mathbb{P}(g(\underline{X})=0 \mid Y=1) + \mathbb{P}(Y=0)\,\mathbb{P}(g(\underline{X})=1 \mid Y=0)$$

$$\Leftrightarrow \mathbb{P}(Y \neq g(\underline{X})) = \pi_1 \pi_0 + \pi_0 \pi_1$$

$$\Leftrightarrow \mathbb{P}(Y \neq g(\underline{X})) = (1-\pi_0)\pi_0 + \pi_0(1-\pi_0)$$

$$= 2\,\pi_0(1-\pi_0)$$

Exercice 2

1) $g_1(x) = \begin{cases} 1 & \text{si } \mathbb{P}(Y=1 \mid X=x) \geq \frac{1}{2} \\ 0 & \text{si } \mathbb{P}(Y=1 \mid X=x) < \frac{1}{2} \end{cases}$

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

$$= \left[\frac{\exp\left(-\frac{x^2}{4}\right)\frac{1}{\sqrt{2}}}{\exp\left(-\frac{x^2}{2}\right) + \frac{1}{\sqrt{2}}\exp\left(-\frac{x^2}{4}\right)} \right]$$

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

$$\Leftrightarrow \exp\left(-\frac{x^2}{4}\right) \geq \sqrt{2}\exp\left(-\frac{x^2}{2}\right)$$

$$\Leftrightarrow x^2 \leq 2x^2 - 2\ln(2)$$

$$\Leftrightarrow x^2 \geq 2\ln(2) \; (= a)$$

$$\Leftrightarrow x \geq \sqrt{a} \text{ ou } x \leq -\sqrt{a}$$

$$g_1(x) = \begin{cases} 1 & \text{si } x \notin [-\sqrt{a}, \sqrt{a}] \\ 0 & \text{si } x \in [-\sqrt{a}, \sqrt{a}] \end{cases}$$

2] $\mathbb{P}(Y \neq g_1(X))$

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

$= \mathbb{P}(Y=0)\,\mathbb{P}(g_1(X)=1 \mid Y=0) + \mathbb{P}(Y=1)\,\mathbb{P}(g_1(X)=0 \mid Y=1)$

$= \frac{1}{2}\left[\mathbb{P}(X \geq \sqrt{a} \mid Y=0) + \mathbb{P}(X \leq -\sqrt{a} \mid Y=0)\right]$

$+ \frac{1}{2}\,\mathbb{P}(-\sqrt{a} \leq X \leq \sqrt{a} \mid Y=1)$

$= \frac{1}{2}\left[1 - F_{N(0,1)}(\sqrt{a}) + F_{N(0,1)}(-\sqrt{a})\right]$

$+ \frac{1}{2}\left(F_{N(0,1)}\left(\frac{\sqrt{a}}{\sqrt{2}}\right) - F_{N(0,1)}\left(-\frac{\sqrt{a}}{\sqrt{2}}\right)\right)$

$= \frac{3}{2} + F_{N(0,1)}\left(\frac{\sqrt{a}}{\sqrt{2}}\right) - F_{N(0,1)}(\sqrt{a})$

3] $\mathbb{E}\left[h_2(Y, g_2(X)) \mid X = x\right]$

$= \mathbb{P}(Y=0 \cap g_2(X)=1 \mid X=x) + 10\,\mathbb{P}(Y=1 \cap g_2(X)=0 \mid X=x)$

$= \mathbb{1}\{g_2(x)=1\}\,\mathbb{P}(Y=0 \mid X=x) + 10\,\mathbb{1}\{g_2(x)=0\}$

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

$\mathbb{E}[h_2(Y, g_2(X)) \mid X = x]$ est minimum uniformément en x

si $g_2(x) = \begin{cases} 1 & \text{si } \mathbb{P}(Y=0 \mid X=x) \leq \frac{1}{10} \, \mathbb{P}(Y=1 \mid X=x) \\ 0 & \text{sinon} \end{cases}$

$\Leftrightarrow g_2(x) = \begin{cases} 1 & \text{si } \exp\left(-\frac{x^2}{2}\right) \leq \frac{1}{10\sqrt{2}} \exp\left(-\frac{x^2}{4}\right) \\ 0 & \text{sinon} \end{cases}$

$\Rightarrow g_2(x) = 1 \Leftrightarrow -\frac{x^2}{2} \leq \log\left(\frac{1}{10}\right) - \frac{x^2}{4}$

$\Leftrightarrow -\frac{x^2}{4} \leq \log\left(\frac{1}{10}\right) - \frac{1}{2}\log(2)$

$\Leftrightarrow x^2 \geq -4 \log\left(\frac{1}{10}\right) + 2\log(2)$

tient la tempy

vrai

Donc, $g_2(x) = 1$!

Esercizio 3

```
library(kernlab)
data(spam)
library(ranger)
model.rf <- ranger(type~., data=spam)
model.rf$prediction.error
```

OOB prediction error

```
library(e1071)
svm.cv <- tune(svm, type~., data=spam)
svm.cv$performances
```

10-fold cross validation error

```
model.svm <- svm.cv$best.model
```

Exercice 4

1) Dans un premier temps, on recode en 0-1 la variable label :
male = 1 et female = 0
Puis, on renormalise les variables prédictives de telle sorte qu'elles soient comprises entre 0 et 1
Par exemple, le vecteur des valeurs de $\underline{x_1} = (x_{11}, \dots, x_{1n})$ est remplacé par le vecteur des valeurs

$$\tilde{x}_{1i} = \left[\frac{x_{1i} - \min(\underline{x_1})}{\max(\underline{x_1}) - \min(\underline{x_1})} \right]$$

On observe que le jeu de données contient autant de "male" que de "female".

Enfin, on compare les performances des classifieurs moins bayes avec celle des classifieurs par régression logistique.

Pour cela nous avons un jeu de données de test de taille 20% du jeu de données total qui a été isolé.

On observe d'après les matrices de confusion que la régression logistique est beaucoup plus performante.