POLYTECH E2I5 année 2021-2022

TD3 –TrD

Traitement de données – Classification

Partie 1 : Les iris de Fisher

Nous allons ici reprendre le jeu de données concernant les iris de Fisher. Ce jeu de données comprend 50 iris de chacune des trois espèces d'iris (*Iris setosa*, *Iris virginica* et *Iris versicolor*). Ainsi sur les 150 iris, 4 variables ont été mesurées : la longueur des pétales, la largeur des pétales, la longueur des sépales et la largeur des sépales. Toutes les variables sont données en centimètres.

Voici les lignes de code qui permettent de charger les libraires utiles ainsi que les « datasets » fournis avec la librairie sklearn.

```
import matplotlib.pyplot as plt
import numpy as np
import seaborn as sns
import pandas as pd

from sklearn import datasets
from sklearn.discriminant_analysis import LinearDiscriminantAnalysis

# On charge les données
iris = datasets.load_iris()
```

On pourra ensuite travailler soit avec des arrays

```
# On travaille avec des arrays
X = iris.data
y = iris.target
target_names = iris.target_names
```

ou avec des DataFrames

```
# On cree un DataFrame

Xdf = pd.DataFrame(iris.data,columns=iris.feature_names)
Xdf['Group']=iris.target
Xdf.boxplot(by='Group')
```

1/ Construire le classifier basé sur une analyse discriminante linéaire

2/ Evaluer ce modèle ; on se posera ici la question de la spécificité du modèle à la base d'apprentissage…

Vous aurez besoin des fonctions :
from sklearn.model_selection import train_test_split
lda=LinearDiscriminantAnalysis()
lda.fit_transform
lda.predict
confusion_matrix
…

Partie 2 : Les données INFRACTUS de Saporta

Étude des données mises à disposition par Gilbert Saporta:
Il s'agit de victimes d'infarctus du myocarde, qui ont été observés à leur admission aux urgences, avec :
- la fréquence cardiaque (FRCAR),
- un index cardiaque(INCAR),
- un index systolique (INSYS),
- la pression diastolique (PRDIA),
- la pression artérielle pulmonaire (PAPUL),
- la pression ventriculaire (PVENT),
- la résistance vasculaire pulmonaire (REPUL).

1/ Vous réaliserez une analyse discriminante sur le jeu de données (en ayant au préalable pris le temps de faire les analyses univariées et bivariées).

2/ Vous étudierez chaque variable séparément et réaliserez les courbes ROC

3/ Vous comparerez les performances d'un classifier basé uniquement sur une variable et celui basé sur l'analyse discriminante linéaire (pour répondre à cette question vous apprendrez et testerez sur les mêmes données)

Partie 3 : toujours sur les données INFRACTUS de Saporta

1/ Vous testerez l'analyse discriminante quadratique et la classification de Bayes naïve
from sklearn.discriminant_analysis import QuadraticDiscriminantAnalysis
from sklearn.naîve_bayes import GaussianNB

2/ Enfin vous testerez la classification basée sur les k plus proches voisins
from sklearn.neighbors import KNeighborsClassifier