IESE4 année 2024-2025

TP3 – TSD

Analyse Discriminante Linéaire (ADL)

Le but du TP2 est de comprendre l'ADL et de programmer la méthode en utilisant Python.
Attention à écrire un programme le plus générique possible afin que vous puissiez le réutiliser sur d'autres jeux de données ; commenter également bien votre programme car vous devrez l'utiliser en TIS5

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.

1/ Charger les données

```
from sklearn import datasets
iris = datasets.load_iris()
On peut ensuite se ramener à une DataFrame
Xdf = pd.DataFrame(iris.data,columns=iris.feature_names)
Xdf['Group']=iris.target
```

2/ Réaliser une analyse univariée des variables (boxplot de chaque variable en tenant compte des espèces ainsi que la moyenne et l variance de chaque variable en tenant compte des espèces)

```
Xdf.boxplot(by='Group')
```

```
Xdf.groupby('Group').mean()
```

3/ Réaliser une analyse bivariée ; on utilisera la fonction
pd.plotting.scatter_matrix ; en mettant l'espèce en colormap

4/ Réaliser l'analyse multivariée (approche statistiques descriptives).

```
from sklearn.decomposition import PCA
lda = LinearDiscriminantAnalysis()
coord_lda = lda.fit_transform(Xdf.iloc[:,0:4],Xdf['Group'])
```

- Vous représenterez les iris dans le plan des 2 vecteurs discriminants

- Vous calculerez la corrélation entre chaque composante discriminante et les variables de départ pour tracer le cercle des corrélations qui permet d'interpréter les nouveaux axes

Reprendre ce qui a été fait en ACP.

- Enfin, vous calculerez les centres de gravité de chaque classe dans l'espace de départ et, en utilisant la fonction d'analyse discriminante, vous calculerez les coordonnées des centres de gravité dans le plan des 2 vecteurs discriminants et vous les ajouterez à la figure précédente.

A réfléchir ! aucune fonction nouvelle n'est à utiliser

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). Pensez à bien reprendre les différentes étapes en vous inspirant de la Partie 1.

Attention : penser à supprimer les lignes qui correspondent à des individus inconnus ! Mettez ces individus de côté pour la fin de l'exercice

2/ Quelles conclusions pouvez-vous faire ?

3/ Projeter les individus « inconnus » dans l'espace des composantes discriminantes. Que pouvez-vous dire ? Quelle méthode proposeriez-vous pour donner la classe des individus inconnus ?