Projet Clustering

Afin de déterminer si une tumeur du sein est cancéreuse ou non, le médecin réalise une biopsie. Un prélèvement est effectué dans la tumeur et les cellules du prélèvement sont analysées au microscope. Dans ce projet, on souhaite déterminer quelles informations sur les cellules extraites permettent de reconnaitre une tumeur bénigne d'une tumeur maligne. Pour cela, on dispose d'une base de données composée de 569 exemples. Chaque exemple correspond à un prélèvement effectué sur une tumeur. Une image de la vue au microscope du prélèvement a été réalisée et analysée. L'analyse d'image permet de dénombrer et caractériser chaque cellule contenue dans le prélèvement. Chaque cellule est caractérisée par 10 grandeurs numériques :

- radius (mean of distances from center to points on the perimeter)
- texture (standard deviation of gray-scale values)
- perimeter
- area
- smoothness (local variation in radius lengths)
- compactness (perimeter^2 / area - 1.0)

- concavity (severity of concave portions of the contour)
- concave points (number of concave portions of the contour)
- symmetry
- fractal dimension ("coastline approximation" – 1)

Pour chaque prélèvement, on peut ensuite calculer pour chaque grandeur :

- la moyenne sur l'ensemble des cellules (mean)

- l'écart type (se)

- la plus grande valeur observée (worst)

L'ensemble des informations est stocké dans le fichier excel nommé breast_cancer_sans_diag qui donne pour chaque exemple les 3 groupes de 10 caractéristiques (mean, se, worst).

Parmi les 569 exemples de la base de données, on suppose qu'il existe des cas de tumeurs bégnines et de tumeurs malignes. A l'aide d'outils de clustering, déterminer quel groupe de caractéristiques parmi _mean, _se et _worst permet de mieux mettre en évidence deux groupes distincts.

Une fois le groupe de caractéristiques choisi, proposer une solution (autre que l'utilisation de la méthode .fit) pour associer une classe à chaque exemple.

Une fois chaque exemple associé à une classe, déterminer quelles caractéristiques parmi celles du groupe retenu discriminent le mieux deux classes. On utilisera pour cela les outils d'analyse vu en traitement de données (boxplot, ACP, ALD, courbes COR). Conclure sur ce qui discrimine les deux classes.

Pour finir, à l'aide du fichier breast_cancer_diag, déterminer si les classes formées correspondent à un type de tumeur.

Fonctions et liens utiles

https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html#sphx-glr-auto-examples-cluster-plot-kmeans-silhouette-analysis-py

from sklearn.discriminant_analysis import LinearDiscriminantAnalysis

from sklearn.metrics import silhouette_samples, silhouette_score

from sklearn.cluster import KMeans