Data Science with Python

Day 06 - Activity

1. To answer below questions, use the ‘breast cancer’ dataset available in scikit-learn.

    a. Load the breast cancer dataset using scikit-learn and Preprocess the dataset using scikit-learn's StandardScaler.
    b. Split the preprocessed dataset into training and testing sets, with 80% for training and 20% for testing and train an SVM classifier with a linear kernel using scikit-learn's SVC.
    c. Evaluate the accuracy of the trained SVM classifier on the testing set
    d. Try using different values of the C parameter in the SVM classifier and observe how it affects the accuracy. (You can iterate over different C values and evaluate the accuracy for each)
    e. Compare their performances of different kernel functions, by using different values for the 'kernel' parameter in the SVC constructor.
    f. Compute the precision, recall, and F1-score of the SVM classifier on the testing set
    g. Perform cross-validation on the SVM classifier to assess its generalization performance using scikit-learn's cross_val_score.

2. To answer below questions use the Mall Customer Segmentation Dataset which contains the basic information (ID, age, gender, income, spending score) about the customers

    a. Load the Mall_Customers.csv file using the Pandas library and preprocess the data by using the StandardScaler from Scikit Learn
    b. Plot a scatter plot of the annual income vs. the spending score. Do you see any clear clusters in the data?
    c. Perform K Means Clustering on the dataset with k=4. What are the cluster centers?
    d. Plot a scatter plot of the annual income vs. the spending score, colored by the cluster assignments. Which clusters are the most distinct and why?
    e. What is the within-cluster sum of squares (WCSS) for the K Means Clustering with k=4?
    f. Try different values of k (2-8) and plot the corresponding WCSS. At what value of k does the WCSS start to level off?
    g. Perform K Means Clustering on the dataset with the k value you get for question (f). What are the cluster centers? Visualize it using a scatter plot.
    h. Calculate the silhouette score for the K Means Clustering with the k value for question (f). What does the silhouette score indicate about the quality of the clustering?

3. Use scipy in-built wine dataset to answer the following questions.
    a. Load the in-built wine dataset and Print the names of the attributes (features) in the loaded wine dataset.

b. Apply the standardization to the wine dataset and print the first few rows of standardized_data.
c. Use the linkage function from scipy to perform hierarchical clustering on the standardized_data using the 'ward' method and Plot the dendrogram using matplotlib.pyplot to visualize the hierarchical clustering.
d. Calculate the silhouette scores for different numbers of clusters (ranging from 2 to 10) in the hierarchical clustering and determine the optimal number of clusters with the highest silhouette score and print the result.
e. Perform cutting on the dendrogram using the fcluster function from scipy to obtain the cluster assignments. Print the cluster assignments for each data point.
f. Apply K-Means clustering with the optimal number of clusters to the standardized_data using sklearn.cluster.KMeans and print the K-Means cluster assignments for each data point.
g. Create a scatter plot to visualize the data points and their cluster assignments from hierarchical clustering. Color the data points based on their cluster assignments and label the plot accordingly.
h. Create a dendrogram with a red dashed line indicating the height corresponding to the maximum number of clusters.