Introduction to Data Science Assignment 5

Due time: 11:59 pm, December 5, 2023

This assignment covers the section of “Clustering”.

Every student is expected to complete the assignment independently. Discussion between students is encouraged, but no plagiarism will be tolerated.

Please submit your answers in a PDF file. If you would like to submit more files, please pack all files, including a readme file, into a compressed file.

In this assignment, you will practise conducting clustering analysis on the data set <https://archive.ics.uci.edu/ml/datasets/Wilt>. The link for downloading the data set is <https://archive.ics.uci.edu/ml/machine-learning-databases/00285/wilt.zip>. The data set is provided in Excel. Please feel free to transform it to any format that fits your programs. The data set contains two subsets.

- The training data set has 4,339 tuples. The first column (class) is the class label, either w or n.
- The test data set has 500 tuples. The first column (class) is the class label, either w or n.

You may use any data mining toolkits. We recommend scikit-learn <https://scikit-learn.org/>.

Run the k-means algorithm using the 5 numeric attributes on the training data set with the following setting.

- Set the maximum number of iterations to 100.
- Set the number of clusters k to 2, 5, 10, and 20, respectively.

Please note that an implementation of k-means, such as the one in scikit-learn, may have more parameters. You can set those parameters to the default settings, or any values you deem appropriate. When you analyze the experimental results, please be aware that those parameters may take effect. Please discuss those effects if necessary.

Question 1 (40 points)

Report the sum of squared errors (SSE) in each iteration. Plot a curve where the x-axis is the iteration number, from 1 to 100, and the y-axis is the sum of squared errors (SSE) with respect to a specific value of k. Your figure should have 4 curves corresponding to the 4 different values of k.

Question 2 (20 points)

Discuss the trend in this figure: what do you observe and why? You should discuss this from two perspectives. Firstly, how SSE changes with respect to more and more iterations? Secondly, how SSF changes with respect to the increase of number of clusters k?

Question 3 (20 points)

Use the test data set to test the quality of the clustering results. Report the corresponding sum of squared distances between all points in the test data set to the closest mean computed by the k-means on the training data set. Plot a curve where the x-axis is k (2, 5, 10, 20) and the y-axis is the sum of squared distance. Plot another curve where the x-axis is k (2, 5, 10, 20) and the y-axis is the total purity of the clustering.

Question 4 (20 points)

What is the best value of k in your study in Question 3? Discuss the trend in the figure of Question 3, that is, how the sum of squared distance and the total purity change with respect to k.