DATA ANALYSIS AND KNOWLEDGE DISCOVERY

Exercise 4.
Unsupervised learning (Fail / Pass with +0 / +1 / +2)

Return your exercise as .ipynb-file via Moodle.

The goal of this exercise is to get some practical experience of using clustering tools. Note that in the following exercises PCA is used only for visualization, all the cluster analyses are done on the original data before dimensionality reduction (10 dimensions for data.txt, 4 dimensions for Iris).

Part 1.

Download the file data.txt from the Moodle web page (right click and “save link as” or similar). Use principal component analysis to map the data to 2 dimensions, visualize the data as scatter plot (you can re-use code for exercise 2). How many clusters can you identify from the data just by looking at the visualization?

Part 2.

Run K-means clustering on the data for different values of K (use the original 10-dimensional data as input to K-means, not the PCA projection). Select the number K for which the clustering has the maximal Silhouette Score. Color the scatter plot of the PCA projection so that members of each cluster are colored differently.

Did you end up with the same clustering of the data as you did based on visual inspection of the PCA plot? If no, do you known which clustering captures the true structure of the data better?

Part 3.

Load the Iris data set used in Exercise 2. Visualize the data with scatter plot of 2-dimensional PCA projection, color each species separately (same picture as in Exercise 2). Remove the true species column, and pretend from now on that you do not know it.

Part 1: Assume that you are told in advance, that there are three different species you should try to find from the data. Cluster the original 4-dimensional Iris data into 3 clusters with K-means method. Create another PCA scatter plot, where you visualize these three clusters. How well does the clustering found by K-means agree with the true class labels?

Part 2: Assume that you have no prior information about the number of clusters in the data. Select the number K for which the clustering of Iris has the maximal Silhouette Score. Visualize the K-means clustering for this value of K. How well does the clustering found by K-means agree with the true class labels?

Bonus question (optional, parts 1-3 give together +0 or +1 points, bonus question gives additional +1 point if done well)

In this bonus exercise you will design and implement an experiment for comparing different clustering algorithms. The aim is to find out, how well they can for a given multi-class classification data set discover the true class structure. Depending on the data you choose it might happen that none of the methods can find the class structure well, this can commonly happen on real world data sets. How well the methods work does not affect whether you get the bonus point or not, only that you compare them as fairly as you can, and report the results.

Download a real-world multi-class classification data set other than Iris data for analysis (see e.g. https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/multiclass.html or https://archive.ics.uci.edu/ml/datasets.html).

Compare three clustering methods chosen from here on the data set:

https://scikit-learn.org/stable/modules/clustering.html

You can compare how well the clustering found by each method matches the true class labels using the adjusted Rand index (see https://en.wikipedia.org/wiki/Rand_index and https://scikit-learn.org/stable/modules/generated/sklearn.metrics.adjusted_rand_score.html).

Questions:

Assume that you know in advance the real number of clusters in the data set. How good Rand index can you get for your clustering, does one method work clearly better on the data than another?

Select the number of clusters automatically based on either the Silhouette Score, or some other selection criterion suitable for the methods you are using (different methods can use different criteria). Does the number of clusters automatically selected match the number of classes in the data?

Are the clustering methods sensitive to initialization, do they give very different clusterings when initialized with different random seed numbers?