COMP40370 – Practical 7

Clustering Analysis

Prof. Tahar Kechadi
Academic year 2021-2022

Assignment Files

• ./practical07.pdf	assignment questions (this file)
• ./specs/marks_question1.csv	data file for Question 1
• ./specs/marks_question2.csv	data file for Question 2
• . /specs/marks_question3.csv	data file for Question 3
• ./test_practical7.py	test file to validate your python program.

Question 1: K-Means Clustering

The file specs/question_1.csv contains coordinates of 2-dimensional (x and y) points with their original cluster labels (org_cluster). Write a Python script that:

1. Visualise the original data points with different colours for their original cluster labels in a scatter plot. Save the plot into output/question_1_1.pdf.

2. Using x and y, use the k-means algorithm to cluster the dataset into x (from 1 to 10) number of clusters. If you are using sklearn, set a fixed random state to 0. Plot inertia (within cluster sum of squares) against the number of clusters. What is the best number of clusters for this data? Save the plot into output/question_1_2.pdf

3. Calculate the Rand Index as an extrinsic measure (i.e. when are know the original/groud-truth clusters), and Silhouette Score as an intrinsic measure (unsupervised) for the given dataset with 3 clusters.

4. Save the input data with an extra column that contains the labels generated by K-Means into a file called output/question_1.csv. The new column should be called cluster_kmeans.

5. Plots the clustering results (including the centroids) as in Q1.1 and save into output/question_1_5.pdf. Make sure that new clusters are marked with the same colours of the corresponding original clusters).

Question 2: K-Means Clustering

The file specs/question_2.csv contains data related to nutritional content of several cereal brands.

1. Discard the columns NAME, MANUF, TYPE, and RATING.
2. Run the k-means algorithm using 5 clusters as a target, 5 maximum runs, and 100 maximum optimization steps. Keep the random state to 0. Save the cluster labels in a new column called config1.
3. Run k-means again, but this time use 100 maximum runs and 100 maximum optimization steps. Again, use a random state of 0. Save the cluster labels in a new column called config2.
4. Are the clustering results obtained with the first configuration different from the results obtained with the second configuration? Explain your answer in your report.
5. Run the clustering algorithm again, but this time use only 3 clusters. Save the generated cluster labels in a new column called config3.
6. Which clustering solution is better? Discuss it in your report.
7. Save the input data with the newly generated columns into a file called output/question_2.csv.

Question 3: DBSCAN Clustering Algorithm

The file specs/question_3.csv contains coordinates of 2-dimensional points. Write a Python script to perform the following tasks.

1. Discard the ID column, the use the X and Y coordinates as data input to the K-Means algorithm to cluster it into 7 clusters. Perform 5 maximum runs, and 100 maximum optimization steps. Keep a random state to 0. Save the cluster labels into a new column called k-means. Discuss the results in your report.

2. Plot the generated clusters and save the plot in a file called ./output/question_3_1.pdf.

3. Normalize the X and Y columns in a range between 0 and 1, then use the DBSCAN algorithm to cluster the points again. Use a value of 0.4 for epsilon, and set the

minimum points equals to 4. Save the generated plot in a file, called ./output/question_3_2.pdf, and save the cluster labels into a new column called dbscan1.

4. Execute DBSCAN again, but this time use a value of 0.08 for epsilon. Plot the generated clusters in a file called ./output/question_3_3.pdf, and save the cluster labels into a new column called dbscan2.

5. Save the data with the cluster labels in a file called ./output/question_3.csv.

6. Discuss the different clustering solutions in your report. Which solution is the best? What is the reason behind the dereferences in the results?

Expected output and submission data

Your submission should be a single archive file (zip, tar, tgz, ...) containing one folder called output and the following files and directories:

- ./run.py main Python scrip
- ./report.pdf your PDF report (4 pages maximum)
- ./output/question_1.csv: cluster results for first question
- ./output/question_2.csv: cluster results for second question
- ./output/question_3.csv: cluster results for third question
- ./output/question_1_1.pdf: cluster plot for first (part 1) question
- ./output/question_1_2.pdf: cluster plot for first (part 2) question
- ./output/question_1_5.pdf: cluster plot for first (part 5) question
- ./output/question_3_1.pdf: cluster plot for third question (k-means)
- ./output/question_3_2.pdf: cluster plot for third question (DBSCAN, first configuration)
- ./output/question_3_2.pdf: cluster plot for third question (DBSCAN, second configuration)
- ./specs/ the original specs folder included in the assignment archive, containing the input data

The final deadline for the submission is **Thursday, 11th of November**, 2021, at **23:55**. You should submit your solution on Brightspace.

Grading

The grading for the assignment will be assigned as follows:

- Question 1: 25%
- Question 2: 20%
- Question 3: 25%
- Report quality and content, code quality, submission format: 30%

Programming requirements and tools

The assignment should be solved in Python, version 3.8 or above (3.9 is recommended). You shall use the following packages for this assignment:

- pandas 1.3+
- matplotlib 3.4+
- sklearn 0.24+ (earlier versions do not support plot_tree)

In particular, the following user guides are available for the required algorithms of the assignment:

- K-Means:
  https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html
- DBSCAN:
  https://scikit-learn.org/stable/modules/generated/sklearn.cluster.DBSCAN.html
- Cluster Evaluation Metrics:
  https://scikit-learn.org/stable/modules/classes.html#module-sklearn.metrics