COMP40370 – Practical 3-B

Clustering Analysis

Prof. Tahar Kechadi

Academic year 2023-2024

Assignment Files

- ./practical03-B.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

Expected output files

- ./Prcatical-03_B.ipynb Python notebook programs.
- ./Prcatical-03_B.html Notebook in HTML showing the outputs.

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

pandas 1.3+, sklearn 0.24+, seaborn, matplotlib

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

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):

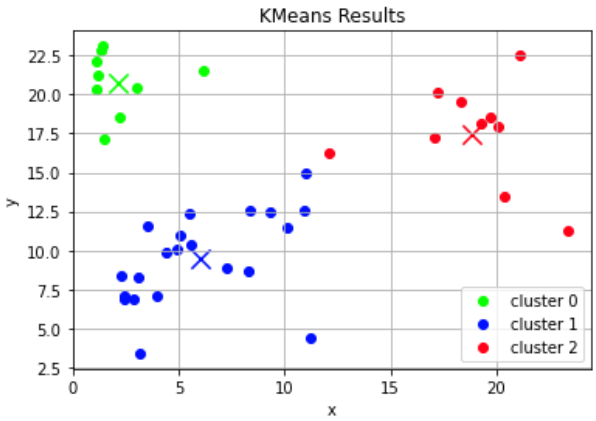
1. Visualise the original data points with different colours for their original cluster labels in a scatter plot.

2. Using only x and y (exclude 'org_cluster' column), 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 and why?

3. Use the k-means algorithm to cluster the dataset into 3 clusters. Then:

    1. Calculate Rand index as an extrinsic measure (between the result of your clustering, and the original cluster labels (org_cluster),
    2. Calculate Silhouette Score as an intrinsic measure (unsupervised) (between the result of your clustering and the data (x and y)).

4. Plots the clustering results, and highlight the centroids. E.g:



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. Are the clustering results obtained with the first configuration different from the results obtained with the second configuration?

4. Run the clustering algorithm again, but this time use only 3 clusters. Save the generated cluster labels in a new column called config3. Which clustering solution is better (among config1, config2 and config3), and why?

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, then 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.

2. Plot the generated clusters. Do you think this clustering is suitable and why?

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, and set the metric to 'Euclidean'. Plot the result, 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 result, and save the cluster labels into a new column called dbscan2.

5. Compare the three clustering results (k-means , dbscan1 and dbscan2). Which solution is the best? And why?

The final deadline for the submission of Practical 01 (Part A and B) is Wednesday, 29th of November at 23:00. Submissions should be in a single file with FirstName_LastName-P3.zip (or tar.gz) format. All submissions must be done in Brightspace.