CS-E4650 Methods of Data mining
Home assignments 2

Deadline Sun 24.10.2021 23:30. Maximum 80 points.
Submit your solutions early, since MyCourses can get stuck, if many people submit simultaneously. If your solutions are in before 24:00, there is no penalty, but after that **-10% penalty**. The ultimate deadline (with -10% penalty) is Tue 26.10. 23:30.

For each task, write a **pdf report** where you present everything **clearly and compactly**. The idea is that your solution is understandable based on the report alone, including all middle steps. **Note that notebooks are not accepted as reports.** By default, the source code is not opened or executed to see if you had actually done something correctly, if it is missing or incorrect in the report. Put possible source code into a **zip** package that is attached (you can include the pdf also in zip, it doesn't harm, but **remember to submit it also separately**). The main reason for submitting the source code is checking that you have done the work independently and using allowed functions (if asked to implement something on your own, without library functions). Include your name and student number in the report and code header.

We will add an optional latex template for the report, including examples, to the MyCourses page, section Assignments.

1. **(max 20p)** This tasks continues Exercise session 2, clustering the rat data. Use the same data set ratdataNormChecked.csv. Use only features liverind, heartind, appind, batind, tailind, ADWBind, gonind, BMI. The distance function is Euclidean.

   a) Cluster the data with agglomerative hierarchical clustering and compare four linkage metrics:
   - single link
   - complete link
   - ~~average link~~
   - ~~Ward's method~~

   Test values of $K = 2, \ldots, 8$ and choose the best clustering with **Silhouette index (SI)**. Report the best clustering with each linkage metric (its K and SI value).

   b) Perform PCA and present the data using only the first principal component. Repeat the same tests as in a) and report the best clusterings.

c) Compare briefly the best clusterings with each linkage metric (for each, choose the better from a) or b)). Did the methods find the same number of clusters? Are cluster sizes similar? If any of the methods found outlier clusters (containing only one or at most a few rats) check them and try to find reasons (like extreme values in individual features).

2. **(max 20p)** In this task you will study hierarchical clustering of set type data and the effect of data order on the clustering results.

   Consider the following market basket data of 8 transactions $(t_1, \ldots, t_8)$:
   t_1: {coffee, milk, sugar, eggs, bread}
   t_2: {bread, coffee, butter, milk, eggs}
   t_3: {sugar, cheese, cream, ham, salt}
   t_4: {eggs, cheese, apples, bread, butter}
   t_5: {apples, bread, eggs, butter, tea}
   t_6: {cheese, bread, coffee, milk, tea}
   t_7: {apples, salt, butter, ham, coffee}
   t_8: {salt, butter, bread, ham, apples}

   a) Calculate pairwise Jaccard distances for each pair of transactions using the following equation. Jaccard distance between sets S_1 and S_2 is defined as

   $$d_J = 1 - \frac{|S_1 \cap S_2|}{|S_1 \cup S_2|}$$

   (i.e., one minus Jaccard similarity, given in Aggarwal Eq. 4.9).

   b) Simulate the agglomerative hierarchical clustering algorithm (Aggarwal Figure 6.7) with the complete linkage metric until transactions are divided into two clusters. The distance function is Jaccard distance. Show that it is possible to yield two different clusterings (into two clusters) depending on the data order.

   You can present the simulation by updating the distance matrix or, alternatively, draw the corresponding dendrogram, if you provide the required inter-cluster distances. Explain the steps (why certain clusters are merged).

   c) Repeat part b) with the single linkage metric. Are the results dependent on the data order?

   You can hand-write/draw the simulations and scan them, if the results are sufficiently clear and readable.

3. **(max 20p)** In this task, you should study two internal clustering validation indices, **Silhouette index (SI) and Davies-Bouldin index (DB)**, and one external index, **Normalized Mutual Information (NMI)**, the version by Strehl and Ghosh, 2003 (see lecture 5 slides).

   Load two data sets, "balls.txt" and "spirals.txt". Both are two-dimensional data, where the third feature ("class") contains the ground-truth labels. Remember to discard the label before running the clustering algorithms.

   It is recommended to plot the data sets for better interpretation.

   a) (Warm-up) Cluster "balls.txt" with i) K-means and ii) spectral clustering using a Gaussian kernel and a Laplacian matrix of your choice. You can try different values of the kernel parameter, to see if it has any effect. Note: if you are using a software package, try to figure out which Laplacian matrix it uses. The distance measure is Euclidean.

      Test values $K = 2, \ldots, 5$ and determine the optimal number of clusters for both methods using all three indices (SI, DB, NMI). Report the results as a table. Which method and K are best for the data?

   b) Repeat a) for "spirals.txt".

   c) Explain and analyze your observations. Which index captured the performance of the algorithm most accurately? Why some indices failed to reflect good performance? Can you use internal indices to determine optimal K for spectral clustering? It is recommended to look at the definitions of indices to better understand their objectives.

4. (20 p) In this task, you should study an alternative internal validation index, τ. Use **only the spiral.txt data** from the previous task.

   Let s be Gaussian similarity function, $s(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(\frac{-\|x_\mathbf{i} - \mathbf{x_j}\|^\mathbf{2}}{2\sigma^2}\right)$, where $\mathbf{x}_i$ is the i-th data point. Given a clustering of the data, define

$$c_{ij} = \begin{cases} 1 & \text{if } \mathbf{x}_i \text{ and } \mathbf{x}_j \text{ are in the same cluster} \\ 0 & \text{otherwise} \end{cases}$$

   For a given clustering, we define

$$\tau = \frac{1}{n} \sum_{i=1}^{n} \frac{\sum_{j \neq i} c_{ij} s(\mathbf{x}_i, \mathbf{x}_j)}{\sum_{j \neq i} s(\mathbf{x}_i, \mathbf{x}_j)},$$

where n is the number of data points.

(a) Compute τ for the previously obtained clusterings by K-means and spectral clustering. Which clustering is now the best one?

(b) Compare the results with the ones indicated by Silhouette and Davies-Bouldin indices. Discuss the results. Is τ index a better choice in this case?

(c) Propose an alternative validation index that allows non-convex clusters and makes better justice to spectral clustering than SI and DB. Here you can use creativity or search literature on existing indices. One option is to combine the idea of τ with other similarity matrices, such as the k-nearest-neighbour graph adjacency matrix. Discuss the motivation and possible disadvantages of your index.