CS-E4650 Methods of Data mining

Exercise 2 / Autumn 2023

2.1 Clustering tendency in the cow data

Learning goal: Evaluating clustering tendency

This task continues Exercise task 1.1. In file *cowdist.csv* (link on MyCourses), you can find pairwise Euclidean, Goodall and combined distances between the cows. Now you should evaluate the clustering tendency, if you were using only numerical, only categorical, or all features (assuming the above mentioned distance measures).

a) Evaluate the clustering tendency visually for all three cases: i) only numerical, ii) only categorical, or iii) all features. In each case, discretize the distances into 5 equi-width bins whose ranges are $[b_1, e_1[, \ldots, [b_5, e_5[$ for suitable b_i, e_i. Plot the histogram (i.e., frequencies of distances in each bin) and evaluate it visually, if it is suggesting a clustering structure.

b) Use the same discretization as in a), but this time evaluate the clustering tendency with the entropy of the distance distribution. (This is described in Aggarwal Sec. 6.2.1.3 and as Approach 2 in the slides of lecture 3.) Calculate entropies for the three cases (numerical, categorical, combined).

c) For comparison, calculate the entropy of uniform distance distribution, where distances are discretized into five bins. Then interpret your results in b). Which features are suggesting a clustering tendency (if any)?

2.2 Spectral clustering of the cow data

Learning goal: Idea of spectral embedding (and clustering)

In this task, you should perform 1-dimensional spectral embedding for the cow data, based on Goodall similarity. Since the file *cowdist.csv* gives the Goodall distance d_{GO}, you need to convert it to Goodall similarity by $sim = 1 - d_{GO}$.

a) Create a 2-nearest neighbour similarity graph, where the edge weight is the Goodall similarity. Present the corresponding weight matrix $\mathbf{W}$.

b) Calculate the corresponding (unnormalized) Laplacian matrix $\mathbf{L} = \mathbf{\Lambda} - \mathbf{W}$, where $\mathbf{\Lambda}$ is the degree matrix.

c) Calculate eigenvalues and eigenvectors of $\mathbf{L}$ and present the data in one dimension. Remember to skip the smallest eigenvalue $\lambda \approx 0$. What is the corresponding clustering of cows?

d) (Optional): Calculate the random-walk Laplacian $\mathbf{L}_{rw} = \mathbf{\Lambda}^{-1}\mathbf{L}$ and repeat step c).

2.3 Hierarchical clustering of bird species

Learning goal: Hierarchical clustering and evaluation of results

In file *birdspecies.csv* you can find data on 64 Finnish birds species that live near watersides. You can find a link to the data and its description in MyCourses. There are three numerical and two categorical features that will be used for clustering and a class variable that tells the biological group of the species.

a) Feature extraction: Create two new variables, *BMI* and *WSI*, from length, wspan and weight. Body-mass index $BMI = \text{weight}/\text{length}^2$ describes how thin the bird is and wing span index $WSI = \text{wspan}/\text{length}$ how long wings it has. (See description how to treat the range-valued features.)

b) Pairwise distances: i) Calculate pairwise **Euclidean distances** using only the new numerical features *BMI* and *WSI*. ii) Calculate pairwise **overlap distances** using only the categorical features (back and belly). The overlap similarity is described in slides of lecture 2, but now we will need distance, i.e., $d_{OL} = 1 - \frac{1}{2}\sum_{i=6}^{7} s(\mathbf{x}_i, \mathbf{y}_i)$, where

$$s(x_i, y_i) = \begin{cases} 1 & \text{if } x_i = y_i \\ 0 & \text{otherwise} \end{cases}$$

iii) Combine the distances as in Task 1.1, but experiment to find good weight λ (suggestion: try $\lambda > 0.5$, since otherwise the categorical distances will dominate too much). You can find a Python implementation of the distance calculation (*combdist.py*) in MyCourses.

c) Cluster the data with agglomerative hierarchical clustering using the combined pairwise distance. Try at least complete and average linkage and different values of clusters, $K = 5, \ldots, 20$. Choose the best clustering with the normalized mutual information *NMI* by Strehl and Ghosh (see slides of lecture 5).

d) What is your opinion, how well does the clustering match the biological grouping? Are there differences between the biological groups, how easily they can be detected by clusters? You can find more information on the hierarchy of the biological groups in the description. Plotting a dendrogram is optional but will help in the interpretation.

2.4 Homework: Balls and spirals

Learning goal: Understanding how internal clustering validation indices are biased towards certain objectives and what are the consequences.

In this task, you should study two internal clustering validation indices, **Silhouette index (SI) and Calinski-Harabasz index (CH)**, 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, CH, 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. Plotting the best clusterings can also help in the interpretation.

d) Plot the 2-dimensional spectral embedding of "spirals.txt". Use the Gaussian kernel, the same parameter settings and the same Laplacian that produced the best spectral clustering into 3 clusters. The partition into 3 clusters should be obvious from the plot. What is the SI value of the clustering in the embedded space? (This is very easy calculation, so if you have a need to evaluate pairwise distances etc, there is something wrong!) Why cannot you estimate the CH index of the clustering in the embedded space? What is your opinion, could we evaluate the internal indices for spectral clustering in the new embedded space and compare with the K-means clustering?

Parts of the report:

1. Cover page: title (course name and assignment number), names and student ids of all participants of the team.
2. Section 1 “Methods”: Describe with 1–2 sentences what programming language and libraries (or own implementations of clustering methods) you used to solve the task.
3. Section 2 “Clustering of the balls data” (subtask a)
4. Section 3 “Clustering of the spirals data” (subtask b)
5. Section 4 “Analysis of results” (subtask c)
6. Section 5 “2-D spectral embedding of the sprirals data” (subtask d)
7. Section “Appendix”: Include here the code you used to produce your results. You can exclude the plotting script, if you tell in “Methods” what tool you used for plotting.

Produce a pdf report including all parts and submit it in MyCourses before the deadline. Recall that this task is homework that is done in groups of 2–3 students – you cannot do the task alone or in a larger group. You can search collaborators in zulip, exercise sessions, or ask help from the TAs.