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: This will be added later

This task is homework that is done in groups of 2–3 students. Note that you cannot do the task alone or in a larger group, so it is recommended to search a group now. The TAs can help to find collaborators.