CS-E4650 Methods of Data Mining

Exercise 2 / Autumn 2022

2.1 Clustering tendency of cows

Learning goal: To study clustering tendency based on pair-wise distances.

We start with the same set of cows we know already from the previous exercise. Three distance matrices between the cows are given below: **N** for the numerical only features, **C** for the categorical only features, and **M** for the combined numerical and categorical distances:

$$\mathbf{N} = \begin{bmatrix} 0.000 & 0.286 & 0.520 & 0.363 & 0.937 & 1.229 \\ 0.286 & 0.000 & 0.520 & 0.543 & 1.098 & 1.414 \\ 0.520 & 0.520 & 0.000 & 0.331 & 0.662 & 0.992 \\ 0.363 & 0.543 & 0.331 & 0.000 & 0.576 & 0.878 \\ 0.937 & 1.098 & 0.662 & 0.576 & 0.000 & 0.331 \\ 1.229 & 1.414 & 0.992 & 0.878 & 0.331 & 0.000 \end{bmatrix}$$

$$\mathbf{C} = \begin{bmatrix} 0.000 & 0.704 & 0.704 & 1.000 & 1.000 & 1.000 \\ 0.704 & 0.000 & 1.000 & 0.750 & 1.000 & 0.454 \\ 0.704 & 1.000 & 0.000 & 0.750 & 0.750 & 0.704 \\ 1.000 & 0.750 & 0.750 & 0.000 & 0.454 & 0.750 \\ 1.000 & 1.000 & 0.750 & 0.454 & 0.000 & 1.000 \\ 1.000 & 0.454 & 0.704 & 0.750 & 1.000 & 0.000 \end{bmatrix}$$

$$\mathbf{M} = \begin{bmatrix} 0.000 & 2.516 & 2.779 & 3.527 & 4.170 & 4.498 \\ 2.516 & 0.000 & 3.703 & 2.949 & 4.350 & 3.001 \\ 2.779 & 3.703 & 0.000 & 2.711 & 3.082 & 3.308 \\ 3.527 & 2.949 & 2.711 & 0.000 & 2.061 & 3.324 \\ 4.170 & 4.350 & 3.082 & 2.061 & 0.000 & 3.491 \\ 4.498 & 3.001 & 3.308 & 3.324 & 3.491 & 0.000 \end{bmatrix}$$

Form equi-width histograms (frequency distributions) of pairwise distances in all three cases. Try different numbers of bins to see if you can find a bimodal (two peaked) distribution, but it is sufficient to present just your final choices. Finding two clearly separated peaks usually hints that there are clusters in data. What is your conclusion, which distance matrix **N**, **C** or **M** might best cluster the cows? Which one is the worst distance matrix in this respect?

2.2 K-modes clustering

Learning goal: To apply K-modes clustering to categorical data.

In this task, you will study the use of K-modes algorithm to categorical data. Use the cows data in the table below, pick the categorical variables race, character, and music to form a data set for this task. After performing the clustering, study different initializations.

Complete the following steps:

a) Build the data set using the categorical variables race, character, and music.

b) Cluster the data using a K-modes algorithm with $K = 2$ clusters.

c) Study if different initializations produce different solutions.

Table 1: Cow data: name, race, age (years), daily milk yield (litres/day), character and music taste.

name	race	age	milk	character	music
Clover	Holstein	2	20	lively	rock
Sunny	Ayrshire	2	10	kind	rock
Rose	Holstein	5	15	calm	country
Daisy	Ayrshire	4	25	calm	classical
Strawberry	Finncattle	7	35	calm	classical
Molly	Ayrshire	8	45	kind	country

2.3 Hierachical clustering

Learning goal: To study two hierarchical clustering methods: complete linkage and single linkage.

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: {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) Use 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 ordering of the data points. Why does this happen?

You can present the clustering process by updating the distance matrix or, alternatively, draw the corresponding dendrogram and provide the required inter-cluster distances. Explain in each step, why certain clusters are merged.

c) Repeat part b) with the single linkage metric. Are the results now dependent on the ordering of the data points? Why?

2.4 Homework: Cluster validation indices

Learning goal: To study different cluster validation indices on different datasets and different clusterings.

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 the slides of lecture 5).

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

a) Cluster "balls.txt" with i) K-means and ii) hierarchical single linkage clustering, both using the Euclidean distance measure.

   Use values $K = 2, \ldots, 5$ in K-means and similarly cut the dendrogram in $2, \ldots, 5$ clusters. Plot the data points with different colors to visualize all your clustering results.

   Determine the optimal number of clusters for both methods using all three indices SI, DB and NMI. Report the results as a table.

   Which clustering method and K value seem to be the best for the data i) based on the validation indices and ii) by visual observation?

b) Repeat all steps of a) for "spirals.txt".

c) Explain and analyze your observations. Which index captured the performance of the clustering algorithm most accurately? Why some indices might have failed to reflect good performance?

Produce a PDF file where you include your plots, tables, discussions and the code you used to produce your results. On the cover page, list the names and student ids of all the participants of your team. Submit the PDF in MyCourses before the deadline!