CS-E4650 Methods of Data Mining

Exercise 1 / Autumn 2025

1.1 Cows with numerical and categorical features

Learning goals: How to use distance/similarity measures when there are both numerical and categorical features in data; similarity graphs.

Look at the cow data in Table 1. The task is to evaluate distances between cows. Note that field 'name' is the cow identifier and not used in any distance calculations. You can calculate distances manually or make scripts, but **implement the distance measures yourself** (do not use ready-made library functions). You can use library functions for min, max, mean, and standard deviation, if you want. For standard deviation, use the corrected sample standard deviation (with $n-1$ in the denominator). Be ready show others your script or how you performed the calculations.

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

a) In this part, use only numerical features. Scale the features with the min-max scaling described in the book (Aggarwal section 2.3.3) and calculate pairwise **Euclidean distances** (L_2 norm) between cows (i.e., 15 distances). Evaluate also standard deviation of pairwise distances (needed in c).

b) In this part, use only categorical features. First, define Goodall distance measure d_G from the Goodall similarity measure G with $d_G = 1 - G$. The Goodall similarity measure is presented in Aggarwal sec. 3.2.2 and the slides of lecture 2 (use that version, since there are many alternative Goodall measures). Then calculate pairwise **Goodall distances**. Evaluate also standard deviation of pairwise distances (needed in c).

c) In this part, use both numerical and categorical features. Create a distance measure that **combines the previous distance measures** (L_2 and d_G) using Equation 3.9 in the book (Aggarwal sec. 3.2.3). (Note that Aggarwal gives similarity measure, but you can combine distance measures in the same manner.) Set λ as the proportion of numerical features. Calculate pairwise distances with the combined measure.

d) Present the results of c) as a **nearest neighbour graph** (as described in Lecture 1 and Aggarwal Sec. 2.2.2.9). Select the threshold ϵ as large as possible such that there are two connected components in your graph. Try to interpret the graph: Are there clear clusters or outliers?

1.2 Metrics or not?

Learning goals: How to study if a distance measure is a metric. Familiarize yourself with some useful distance measures.

a) Show that the **fractional quasinorms** (lecture 2), i.e., L_p with $p \in]0, 1[$, are not metrics. Hint: It suffices to find one counter-example. It is easiest when you have small dimensional vectors and $p = 0.5$.

b) Show that the **string edit distance** (lecture 2) is a metric when insertion and deletion operations have cost 1 and substitution cost 2.

c) Show that the **shortest path distance** (Aggarwal 3.2.1.7) is a metric, when the edge weights are Euclidean distances between the corresponding data points. Here we assume the underlying nearest neighbour graph is undirected.

1.3 Principal component projection

Learning goal: How to use PCA to project data to lower dimensions.

Note: you can use code to do the computations, but you must describe all steps and report intermediate results.

Consider the following two-variable data set, where each row corresponds to a point in a two-dimensional space:

$$\begin{pmatrix} 0 & 1 \\ -1/2 & 3/2 \\ 3/2 & 5/2 \\ 1 & 3 \end{pmatrix}.$$

1. Carry out the principal component analysis of these data, that is, compute the eigenvalue decomposition of the corresponding sample covariance matrix.

2. Consider the resulting decomposition:

   (a) Use it to transform the original 2-dimensional data set into a 1-dimensional representation (a 4×1 matrix) such that the variance of the resulting data is equal to the largest eigenvalue.

   (b) Next, use it to transform the original data set into a 2-dimensional representation, such that the variance of one of the columns is equal to the smallest eigenvalue.

3. Given two points in a k-dimensional Euclidean space,

$$\mathbf{x} = (x_1, x_2, \ldots, x_k)^T,$$
$$\mathbf{y} = (y_1, y_2, \ldots, y_k)^T,$$

   the *Euclidean distance* between them is computed as follows:

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^{k} (x_i - y_i)^2}.$$

   (a) Calculate the Euclidean distances between all pairs of the points in the original data set.

   (b) Calculate the Euclidean distances between all pairs of the points in the 1-dimensional representation obtained above in subtask 2(a).

   (c) Calculate the Euclidean distances between all pairs of the points in the 2-dimensional representation obtained above in subtask 2(b).

   (d) What is the effect of the previous principal component transformations on these distances?

4. Now consider the following data set:

$$\begin{pmatrix} \sqrt{1/2} & \sqrt{1/2} \\ \sqrt{1/2} & 2\sqrt{1/2} \\ 4\sqrt{1/2} & \sqrt{1/2} \\ 4\sqrt{1/2} & 2\sqrt{1/2} \end{pmatrix}.$$

   Repeat subtasks 1, 2, and 3 on these data. What are the similarities and differences between the results on this data set and the first one? Can you give a geometric explanation for the observed similarities and differences? Hint: Plot the two data sets!

1.4 Homework: Clustering tendency

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. Remember that it is **forbidden to use any AI tools to solve the tasks, input course material (including tasks and their solutions) to AI tools or otherwise distribute them to other services**.

Learning goal: To familiarize yourself to the concept of clustering tendency and entropy-based measures for it.

In this task, the idea is to experimentally study how clustering tendency can be observed both visually and with entropy-based measures.

1. Download the data file *clustering-tendency.csv* from MyCourses. Check that it contains 1000 vectors that are three-dimensional with all values in the range $[0, 1]$.

2. Plot the *histograms* of the values of each component of the data. Use histograms with 64 bins and include the plots in your report. Identify clearly which components (indices 0, 1, and 2) they depict.

3. Create *scatter plots* of each of the three pairs of two components. Ensure that the aspect ratio of the plot axes is *equal*. Include the plots in your report and identify clearly which component each axis depicts.

4. Consider the potential clustering tendency of the data points in each of the single dimensions and in each two-dimensional projection.

5. Discretize the data with $m = 64$ uniform three-dimensional grid ranges in $[0, 1]^3$ and calculate the *probability-based entropy* E as defined in equation (6.2) in Aggarwal's book. Use the natural logarithm and notice that $0 \log 0 = 0$ because $\lim_{x \to 0^+} x \log x = 0$.

6. Repeat the above discretization with $m = 64$ uniform two-dimensional grid ranges of each pair of the data dimensions in $[0, 1]^2$. Calculate all three resulting entropies.

7. Finally, discretize each data dimension individually with $m = 64$ uniform grid ranges in $[0, 1]$ and calculate all three resulting entropies.

8. Analyze the seven entropy values together and discuss what one can guess about each data component's potential contribution to clustering tendency.

9. With the readily available entropy values, simulate *greedy backward selection* as a search strategy for a potentially optimal set of data dimensions with respect to clustering.

10. Similarly, simulate *greedy forward selection* and find a potentially optimal set of data dimensions.

11. Analyze the outcomes of the two greedy search procedures compared to your expectations based on the visual inspection and the knowledge on all seven calculated entropy values.

12. Consider, why $m = 64$ was a handy choice for the number of grid ranges in all entropy calculations.

Parts of the report:

1. Cover page/beginning: title (course name and assignment number), names and student ids of all participants of the team.

2. Section "Methods": Describe *very briefly* (one paragraph) your methods: what language, libraries and tools you used. If you made more experiments than asked, tell it here and report the results in section "Extra experiments".

3. Section "Histograms": Include the three component-wise histograms and discuss their differences and potential contribution to clustering tendency.

4. Section "Scatter plots": Present the three two-dimensional scatter plots and discuss what can be learned about the dependencies and independencies between the data components and what might be their consequences to the clustering tendency.

5. Section "Entropies": Include a table that specifies the different combinations of data dimensions used in calculating the probability-based entropies and the obtained entropy values. Discuss the range and variation of the entropies and what they might indicate about clustering tendencies with different dimension combinations.

6. Section "Greedy backward selection": Show the stages of the backward selection process and explain the choices made.

7. Section "Greedy forward selection": Show the stages of the forward selection process and explain the choices made.

8. Section "Analysis": Analyse and compare your findings from the visual inspections, entropy calculations, and the use of the two greedy search strategies.

9. Section "Why 64 bins?": Why was $m = 64$ a good choice? What other values of m could be similarly motivated with the used data? Discuss the selection of the m value with higher-dimensional data sets.

10. Section "Conclusions": What are your final conclusions? How do you interpret the results with respect to potential clustering tendency? Write briefly, just one or at most two paragraphs.

11. Section "Appendix – Code": Include here the code you used to produce your results.

Produce a PDF report containing all parts as described above and submit it in MyCourses before the deadline. You can find a LaTeX template for the report in MyCourses, under section "Exercises". One submission per group!