CS-E4650 Methods of Data Mining

Exercise 1 / Autumn 2022

1.1 Cows with numerical and categorical features

Learning goal: To use distance 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 and possible groupings 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. Remember to report intermediate steps and prepare to show your code and its outputs.

a) In this part, use only the numerical features. Scale the features with the min-max scaling described in the book (Aggarwal sec. 2.3.3) and calculate pairwise Euclidean distances (L_2 norm) between the cows. Present the results as a *similarity graph* as described in sec. 2.2.2.9 of the book for a *neighborhood graph*. Select the threshold ϵ as small as possible still keeping the graph connected.

b) In this part, use only the categorical features and calculate pairwise similarities using the Goodall similarity measure. The Goodall similarity measure is presented in Aggarwal sec. 3.2.2 and the slides of lecture L02 (use that version, since there are many alternative Goodall measures). Similarly to the previous case, present the results as the smallest possible connected similarity graph.

c) In this part, use both the numerical and categorical features. Create a distance or similarity measure that combines the previous distance measures 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. Create now a similarity graph using the combined measure.

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

1.2 Similarity in social media profiles

Learning goal: To study distance functions and metrics in set data.

Consider a social network where each user is associated with a set of labels that best describe a set of properties of the user. We define the profile of the user to be the set of associated labels, i.e., given a set $P = \{p_1, \ldots, p_n\}$ of user profiles and a universe of labels $L = \{l_1, \ldots, l_m\}$, each profile $p_i \in P$ is a set of labels $L_i \subseteq L$. The task is to design functions to measure the distance between two labels and similarity between two user profiles.

a) Propose a distance measure between labels, more precisely, given any two labels $l_1, l_2 \in L$, present a distance function d such that $d(l_1, l_2)$ returns a distance measure between labels l_1 and l_2. The distance function should be (i) intuitive and (ii) satisfy the metric properties (see next parts).

b) Discuss the intuition, strengths, and limitations of your measure.

c) Prove that your distance function is a metric.

d) Now we want to compare the similarity of two user profiles. Propose an appropriate function $s(p_1, p_2)$ to compute the similarity of any two profiles $p_1, p_2 \in P$ and discuss its intuition.

e) Be prepared to show code that implements d and s and demonstrate its behavior with a small set of toy data.

1.3 Lower bounding a distance

Learning goal: To consider effective implementations of nearest neighbor search.

Consider the *nearest neighbor search problem*: Given a dataset of n objects $X = \{x_1, \ldots, x_n\}$ and a query object q, we want to find the object $x^* \in X$ that minimizes the distance $d(q, x)$, that is,

$$d(q, x^*) \leq d(q, x) \;\; \forall x \in X \; . \tag{1}$$

Assume that computing the distance function d is *very expensive*. Assume now that we are able to define another distance function d_{LB}, which is a *lower bound* of distance d. This means that for all pairs of objects x and y it should be

$$d_{\text{LB}}(x, y) \leq d(x, y) \; . \tag{2}$$

Furthermore, assume that computing d_{LB} is *significantly more efficient* than computing d.

a) Write pseudocode of an algorithm for the nearest neighbor search using distance d.

b) Explain how to use the lower-bound distance d_{LB} to speed up the search algorithm of the previous part. Write pseudocode for the modified search algorithm.

c) What is a desirable property for the lower-bound distance d_{LB} to be as effective as possible for the modified algorithm? Explain why.

1.4 Homework: Curse of dimensionality

Learning goal: To understand that data may behave differently when its dimensionality increases.

In this task the idea is to study the behaviour of *extreme distances* and *relative contrast* with different L_p norm based distance measures when the dimensionality of data increases. The relative contrast for data vector $\mathbf{x}_i$ is defined as

$$C_i = \frac{D_i^{max} - D_i^{min}}{D_i^{min}} , \tag{3}$$

where D_i^{max} and D_i^{min} are the maximum and minimum distances from vector $\mathbf{x}_i$ to any other vector in the data, calculated with a given distance function.

In the experiment, vary the dimensionality as $d = 1, 2, 3, 4, 5, 10, 20, \ldots, 100$. For each d, simulate random data with $n = 100$ data points $\mathbf{x}_i \in \mathbb{R}^d$, where each component has uniform distribution in the range $[0, 1]$. For each $\mathbf{x}_i \in \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$, find the minimum, average and maximum distances from it to all other vectors. Then calculate the difference between the maximum and minimum distances and the relative contrast value. Finally compute the averages of all these five measures over all $\mathbf{x}_i$. Plot the average values and their logarithms as functions of increasing d. (Place all distance values in one plot and the contrast values in another plot. Similarly, do not mix the original and the logarithmic values in the same plot. You should thus have in total four plots.) Repeat the experiment four times with the following L_p measures as the distance function: $L_{0.5}$, L_1, L_2, and L_3.

a) What happens to each one of the distance values and the difference between the maximum and minimum distances when d increases? Explain, how you use the original and logarithmic plots to draw these conclusions.

b) What happens to the relative contrast measure when d increases? How can this finding be explained based on the distance plots?

c) What kind of an effect does the p value have on the bahavior of the relative contrast measure as a function of d?

d) How do you interpret the results with respect to so called *"curse of dimensionality"*?

Produce a PDF file where you include your plots, 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!