CS-E4650 Methods of Data mining

Exercise 1 / Autumn 2022

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. Remember to report intermediate steps and prepare to show your code and its outputs.

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	10	calm	rock
Sunny	Ayrshire	2	15	lively	country
Rose	Holstein	5	20	calm	classical
Daisy	Ayrshire	4	25	kind	rock
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. Present the results as a nearest neighbour graph (as described in Lecture 1 and Aggarwal Sec. 2.2.2.9). Select the threshold ϵ as small as possible still keeping the graph connected.

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 and

present the results as a nearest neighbour graph, once again selecting minimal ϵ such that the graph remains connected.

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. It is recommended to use the unbiased estimate of standard deviation. Create now a nearest neighbour graph using the combined measure and select minimal ϵ that keeps the graph connected.

d) Compare the results. Is the combined measure graph (c) more similar to the numerical (a) or categorical (b) measure graph? Can you explain why?

1.2 Similarity in social media profiles

Learning goal: To study distance functions and metrics for set form 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. Depending on your measure, this can be tricky, but study at least the easy properties!

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

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

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 the meaning of the curse of dimensionality.

In this task, the idea is to study experimentally how distances with different L_p measures behave when data dimensionality increases. The experiment is somewhat similar than described in the textbook (Aggarwal Ch. 3.2). Here is the testing procedure:

1. Test dimensions $k = 2, 3, 4, 5, 10, 20, 30, \ldots, 100$.

2. In each test, generate q random data sets of $n = 100$ points, where each feature has a uniform distribution in the range $[0, 1[$[1] However, you are not allowed to generate a zero vector $\mathbf{0}$, since the origin will be used as our query point. For stable results, you should use relatively large q (e.g., $q = 100$ or more – tell q in the report).

3. For all q random data sets of dimensionality k, evaluate distances to all points from the origin $\mathbf{0}$ using L_p measures (L_p norms or the L_p fractional quasinorm) with a) $p = 0.5$, b) $p = 1$, c) $p = 2$, d) $p = 5$, e) $p = \infty$.

   With each measure, calculate the following five statistics: **minimum** distance (D_{min}), **maximum** distance (D_{max}), **mean** distance, **variance** of the distance, and **relative contrast** that is defined as

   $$Ctr = \frac{D_{max} - D_{min}}{D_{min}}.$$

4. For each k, calculate mean of the five statistics (i.e., average minimum distance, average maximum distance, average mean distance, average variance of distance, and average relative contrast) over all q randomizations. Let these statistics be notated $Min(L_p)$, $Max(L_p)$, $Avg(L_p)$, $Var(L_p)$, $Ctr(L_p)$.

5. Prepare plots of results as a function of k (x-axis):

   a) All relative contrasts $Ctr(L_p)$ in one plot (5 curves $L_{0.5}, \ldots, L_\infty$). Here the y-axis is Ctr.

   b) One plot for each L_p showing average minimum, maximum and mean distances ($Min(L_p)$, $Max(L_p)$, $Avg(L_p)$). Here the y-axis is distance. For better interpretation, it is recommended to use the

[1] If your library generates random numbers from $[0, 1]$ or $]0, 1]$, that's equally fine, just tell the interval in your report.

same scale for plots of L_2, L_5 and L_∞, but don't mix them in one plot.

c) One plot for each L_p showing average variance of the distance ($Var(L_p)$). For better interpretation, it is recommended to use the same scale for plots of L_2, L_5 and L_∞, or you can even present these three curves in the same plot.

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 "Methods": Describe *very briefly* (one paragraph) your methods: what language/tools you used, randomization interval (if not $[0, 1[$), possible libraries for the L_p measures (own implementations encouraged!), value of q. If you made more experiments than asked, tell it here (e.g, tested $k > 100$ or other L_p measures) and report the results in section "Extra experiments".

3. Section "Relative contrast": Present the relative contrast plot and discuss the results of contrast evaluation. What is happening when k increases? When Ctr drops below 1? What is the effect of p in different L_p measures? If a curve seems to be converging, tell also the value that it is approaching.

4. Section "Minimum, maximum and mean distances": Present the plots of minimum, maximum and mean distances for each L_p and discuss the results. How the curves are behaving when k increases? Can you characterize the form of curves? What is the effect of p? If a curve seems to be converging, tell also the value that it is approaching.

5. Section "Variance of distance": Present the variance plots for each L_p and discuss the results. How the curves are behaving when k increases? Can you characterize the form of curves? What is the effect of p? If a curve seems to be converging, tell also the value that it is approaching.

6. Section "Extra experiments": This is optional, but if you made any extra experiments, report them (the new insights) briefly here.

7. Section "Conclusions": What are your final conclusions? How do you interpret the results with respect to so called "curse of dimensionality"? Write briefly, just one or at most two paragraphs.

8. 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.

Hint: You can characterize the behaviour of the curves by common functions that they resemble, like linear ($y = ax$), constant ($y = a$), quadratic ($y = ax^2$), logarithmic ($y = a\log(x)$), squareroot ($y = \sqrt{x}$), multiplicative inverse ($y = (ax)^{-1}$) etc.