CS-E4650 Methods of Data mining
Home assignments 1

Deadline Sun 3.10.2021 23:30
Submit your solutions early, since MyCourses can get stuck, if many people submit simultaneously. If your solutions are in before 24:00, there is no penalty, but after that **-10% penalty**. The ultimate deadline (with -10% penalty) is Tue 5.10. 23:30.
Maximum 60 normal points + 10 extra points.

1. **(max 20p)** In this task, you should analyze correlations in a version of the Rat data (Exercise 1) and study what happens if you ignore the data types, all special codes and clear outliers. Load data corrtestdata.csv and its description from MyCourses.

   You can use any tool you like to calculate the Pearson correlation coefficient r or implement it yourself (see Appendix A), just tell what tool you used or return the program code. Concentrate on the strongest correlations, where the absolute value $|r|$ is at least 0.40 after rounding to two decimal places (these are now called "strong" correlations, although many of them are only moderate). Present correlation coefficients in your report as rounded to two decimal places.

   After each analysis step, you are asked to analyze changes to the previous step. It is sufficient to report only big changes, i.e., if some "strong" correlation became weak or changed a lot (at least 10 percentage points, i.e., 0.10), or a weak correlation became "strong".

   a) Calculate all pairwise correlations between features, excluding only the rat id. Report strongest correlations involving i) categorical features, ii) temporal features day and year, ii) other numerical features.

   b) Remove outlier rats, rat2, rat53, rat120, and rat434, and calculate correlations again. How did the correlations involving either liverind or heartind change? What is your explanation for big changes?

   c) Continue with the data from b), where the listed outliers are removed. Change the special codes for the freezer rats: day=0 and year=-1. How did the correlations involving either day or year change? What happens if you remove all freezer rats? What is your conclusion on the changing correlations involving day or year?

e) Continue with the data where all freezer rats are removed. Test changing special codes of categorical features femstate, kmethod and place. Can you generate any big changes in correlations involving these features?

f) What is your final conclusion, which correlations were reliable? Are there any strong correlations showing a linear trend? You can also make scatterplots to help in interpretation.

2. **Extra task, extra points, max 10p:** This task continues task 1. Here the goal is to study if there are real correlations involving the circular feature day. Use the data from task 1, but keep **only wild rats** (place 1–3) and **remove the listed outliers** (1b)).

   Appendix B explains how you can evaluate a correlation coefficient R^2 between a circular and a linear numerical feature. Implement R^2 and evaluate correlations between day and linear numerical features (at least gonfatind and batind). You can use a library function for the Pearson correlation coefficient.

   The initial hypothesis is that the time of the year could affect at least batind and gonfatind among wild rats. Do you find any evidence for the hypothesis? Since R^2 is squared, it can be easier to interpret after taking the square root (but the sign is unknown). Note that you need to present day first as an angle (see lecture 1 slides).

3. **(max 20p)** **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 2-dimensional space:

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

   (a) (5 points) Carry out the principal component analysis of these data, that is, compute the eigenvalue decomposition of the corresponding sample covariance matrix.

   (b) (5 points) Consider the resulting decomposition.

      i. 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.

ii. 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.

(c) (5 points) Given two points in d-dimensional Euclidean space,

$$x = (x_1, x_2, \ldots, x_d)^T,$$
$$y = (y_1, y_2, \ldots, y_d)^T,$$

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

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

i. Compute the Euclidean distance between all pairs of points in the original data set.

ii. Compute the Euclidean distance between all pairs of points in the 1-dimensional representation obtained in exercise 3b.

iii. Compute the Euclidean distance between all pairs of points in the 2-dimensional representation obtained in exercise 3b.

iv. What is the effect of the previous transformations on these distances?

(d) (5 points) 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 exercises 3a and 3b 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 similarities? Hint: plot the two data sets.

4. **(max 20p)** 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. In d) you can draw by hand, if the results are clearly readable. Remember to report intermediate steps and include possible scripts in your report.

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.

b) In this part, use only categorical features. First, define Goodall distance measure d_G from the Goodall similarity measure G with $d = 1 - G$. The Goodall similarity measure is presented in Aggarwal 3.2.2 and slides of lecture 3 (use that version, since there are many alternative Goodall measures). Then calculate pairwise Goodall distances.

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). Set λ as the proportion of numerical features. Calculate pairwise distances with the combined measure.

d) Compare results of a)–c):

   i) Present histograms (frequency distributions) of pairwise distances in all three cases. Test different numbers of bins to see, if you can find a bimodal (two peaked) distribution (it is sufficient to present your final choices). Finding two clearly separated peaks usually hints that there are clusters in data. What is your conclusion, which measure (from a)–c)) can best cluster the cows?

   ii) Try simple graph-based clustering of data using pairwise distances: First create a complete distance graph between cows (all cows are connected to each other and edge weights are the distances). Then remove edges with longest distances (largest weights) until the graph is broken into two connected components. These components are clusters. Present your final clusterings (graphs). What is your opinion, which measure (a–c) produced the best clustering and why?

Appendix A: Pearson correlation coefficient

Pearson correlation coefficient r between two numerical vectors $\mathbf{x}$ and $\mathbf{y}$ can be calculated as

$$r(\mathbf{x}, y) = \frac{\sum_{i=1}^{n}(x_i - \mu_x)(y_i - \mu_y)}{\sqrt{\sum_{i=1}^{n}(x_i - \mu_x)^2}\sqrt{\sum_{i=1}^{n}(y_i - \mu_y)^2}},$$

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

name	race	age	milk/d	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

where μ_x and μ_y are mean values of $\mathbf{x}$ and $\mathbf{y}$.

Appendix B: Coefficient R^2 between a circular and a linear feature

Let Θ be a circular feature (presented as an angle in $[0, 2\pi[$) and X some linear numerical feature. Correlation between corresponding data vectors $\theta = (\theta_1, \ldots, \theta_n)$ and $\mathbf{x} = (x_1, \ldots, x_n)$ can be evaluated by the following coefficient R^2:

$$R^2(\theta, x) = \frac{r_{xc}^2 + r_{xs}^2 - 2r_{xc}r_{xs}r_{cs}}{1 - r_{cs}^2},$$

where r_{xc} is Pearson correlation between X and $\cos(\Theta)$, r_{xs} between X and $\sin(\Theta)$, and r_{cs} between $\cos(\Theta)$ and $\sin(\Theta)$.