CS-E4650 Methods of Data mining
Exercise session 1

All tasks are related the "Rat data" (ratdataRaw.csv) that we will process and analyze during the course. Load the data and its description from MyCourses exercise section. The data set is small, so you can use a spreadsheet tool in some of the tasks.

1. Data types: Read the description and specify data types of all variables.

2. Handling missing values: It is safest to remove missing values, but try to keep as many features as possible and as many rats with the stomach ulcer as possible. (Another solution is imputation, studied in the extra task.)

   a) Remove columns that have so many missing values that they can't be used.

   b) Remove rats that have missing values in the features that are defined for them. (Note that female rats have more features than male rats.)

   Note: You can keep feature "year" even if it is not defined for the freezer rats – it can be useful later. Just remember the problem in the modelling.

3. Simple feature extraction/transformation:

   a) Feature ADWBWind is the weight of the adrenal gland divided by the body weight. Do a similar normalization for other weights of organs and fat. Let the new features be "liverind", "heartind", "appendixind", "gonfatind" and "batind".

   b) Normalize tail length by dividing with the body length. Let the new feature be "tailind".

   c) Calculate a new feature, BMI (body mass index), as BMI = weight/blength^2.

   d) Create a new binary variable, "mother", that is 1 for all female rats who are either pregnant, nursing or both, and 0 for other rats.

4. Plotting and studying distributions:

a) Plot histograms of features blength, heart, BMI and liverind. (Set a sufficiently large number of bins.) Check extreme outliers if you can find errors. There should be at least four rats with clear errors (two are obvious, for the other two you need to look at multiple fields). Remove these rats.

b) Plot a histogram of feature "day". Identify visually two break points that divide days into the "winter" and "summer" seasons. You may need to try different numbers of bins to find good break points in the spring and autumn. Note that 0 is a special value and the new variable "season" will have three values: "freezer", "summer", "winter" (no need to create the feature now, just know how to do it).

5. Extra task on imputation: Before this task, remove rat183 that has an erroneous heart measurement. Then remove temporarily heart measurements (heart weight and and heartind) of rat3, rat5 and rat6. The original heart weights are 0.23, 0.23 and 0.26 and these will be compared to imputed values.

   i) What is the mean value of remaining heart weights? This value is used in the mean value imputation. Would it be sensible for these rat puppies?

   ii) Calculate the mean of remaining heartind values and multiply it with the body weight of a rat to get an estimate for its heart weight. Are the results any better?

   iii) Try K-nearest neighbour imputation with different values of K. Use Euclidean distance and calculate it over some or all numerical (non-circular) features[1]. How accurate are the imputations now?

[1] If you want to use other features, you should first define a suitable distance measure for mixed categorical and numerical features, including circular features.