

Praktikum zu Moderne Methoden der Datenanalyse

Exercise 8: Data Mining Cup: Likelihood Ratio

In this exercise we continue to work on the Data Mining Cup task introduced in exercise 7. There we made a classification based on cuts on several variables. But if space of variables has a high dimension, this is usually not the best approach. Therefore a possible improvement is to combine information from different variables to one quantity, which is then used for the selection. One possibility to do this is to form a likelihood ratio.

- **Exercise 8.1:**

Take the variables, which you used for the cut based approach in exercise 7.2, and calculate the ratio of the probability density functions for good and bad customers $P_{good}(\vec{x})/P_{bad}(\vec{x})$.

To calculate this use the training sample to obtain $P_{good}(x_i)$ and $P_{bad}(x_i)$ for each individual variable x_i and assume all variables to be uncorrelated. As a first approximation take the histograms for the two classes of events and normalise them to obtain the corresponding probability density functions. How can numerical problems with small numbers be avoided?

Determine the likelihood ratio for each order and plot the likelihood ratio distribution separately for good and bad customers.

- **Exercise 8.2:**

Find a reasonable cut on the formed likelihood ratio to classify the events of the training sample. Calculate a score for this probability cut as you did in exercise 7.2 for the cut based approach.

- **Exercise 8.3:**

Some variables might be correlated. Try to improve the selection by taking some of the correlations into account. This is done for two correlated variables x_m and x_n by replacing $P(x_m) \cdot P(x_n)$ with $P(x_m, x_n)$ for both classes while the rest of

the probability function remains unchanged. Use multi dimensional histograms to describe such probability distributions.

Plot the likelihood ratios, choose a cut and calculate your score on the training sample.

- **Exercise 8.4:**

Try to improve your selection by using a parametrised function instead of a histogram for the probability density distribution of one or more variables. First identify distributions which could be reasonably parametrised. Then make up a parametrised function (TF1) and fit it to the histogram of that variable.

Again calculate the score on the training sample for a cut on the likelihood ratio constructed with a parametrised probability distribution.

- **Exercise 8.5:**

Classify the test dataset `class.root` as described in exercise 7.3 with the selection which you think is your best and give it to a tutor, who will calculate a score for you.