Prediction of Dialysis Length

Antje Schubotz, Adrian Loy

The human body is dependent on the functionality of many organs. Among them are the kidneys, whose main purpose is to filter poisonous substances from the blood, which then get excreted in the urine. If the kidneys malfunction, toxins have to be artificially removed via diffusion through a membrane. This treatment is called dialysis. Doctors have to determine the length of each dialysis session in advance by referring to their experience and medical guidelines. However, these guidelines only take few parameters into account. It is important to minimize the treatment‘s duration, because dialysis poses a high mortality risk and is very costly. We therefore research on predicting the length of dialysis from the patient‘s data.

OBJECTIVES

We apply regression algorithms to predict the dialysis length, because it is a continous variable, and selected Support Vector Machines (SVM) and Local Polynomial Regression (LPR) for this research. The former predicts values using a learned function where all datapoints of the training set lie inside a minimized margin around it. The latter derives values from the direct neighborhood by locally approximating the underlying function.

Moreover, we use various settings to compare the performance of different candidate models: The attributes of the algorithms (e.g., margin size, neighborhood size) are varied and different subsets of the data (e.g., no outliers) are analyzed.

We finally extract a gold standard of patients who responded well to their treatment. We train another model on this subset and apply it to the rest of the patients to see how the results vary.

METHODOLOGY

Data extraction and preparation:

We used the MIMIC-III database with data from an intensive care unit and extracted dialysis patients including several features such as the personal profile (e.g., age, sex, weight), laboratory values (e.g., creatinine, pH, urea concentration), medical condition (e.g., elixhauser score), and the duration of each dialysis session. Before the algorithms can work on this extracted dataset containing 2047 dialysis sessions, we had to impute missing values, remove any occuring dates and convert any nominal values to numerical ones.

Model development and validation:

To compare the different settings, we used ten-fold cross-validation with a local random seed to produce comparable results. The model development with various configurations included the use of different datasets as mentioned above, variation of the algorithm‘s attributes, and a separate prediction of feature weights with forward selection and backward elimination in order to minimize the number of features. Furthermore, we tuned the results manually by combining these methods.

Experimental setup:

The data extraction was done with an SQL query, whereas we performed the data preparation, model development, and validation with a visual editor called RapidMiner. The tool is directly connected to our database, so we can use the query to import the data.

PRELIMINARY RESULTS

We evaluated the algorithms on the whole dataset (n=2047) as well as on a small subset (n=79) with nearly no missing values. So far, the SVM has a poor accuracy on the larger dataset (correlation: 0.05), while it performs better on the smaller subset (correlation: 0.64). The LPR currently produces better results with a correlation of 0.877 on the whole dataset, but fails on the subset, because there are not enough datapoints in the neighborhood.

One of our next steps is to improve the accuracy of our prediction by reducing noise and outliers, splitting the data into subsets to train specialized models (e.g., male and female patients) and testing more parameter settings. Another challenge is to make the decision process transparent. This is essential so that doctors can rely on our predictions as they are responsible for treatment decisions. The SVM gives a list of all features and their corresponding weights, but LPR does not, potentially hindering its use in practice.