ANALYSIS OF MACHINE LEARNING ALGORITHMS

HODOSSY, SZABOLCS

1. Excerpt

My thesis was born due to my interest and lack of knowledge of machine learning algorithms. My aim was to understand them better, how they are used and in what circumstances are they best utilized. I chose binary classification as my main topic as it is widely used and most often these techniques can be used for regression as well. I chose Logistic Regression, Random Forest, Gradient Boosting Machine (using decision trees), k-Nearest Neighbours and Support Vector Machine to study and apply to a some datasets. They are the most commonly used, basic algorithms of the field, and good understanding of these is essential in data science. The first step was to understand the modelling process: what is needed to be taken care of before model fitting can start. So I started my thesis with describing common problems in the raw data, how they can be tackled and how can we measure the effectiveness of the pedictor we build. I chose the Gini value as measurement of goodness, as it shows how much better the models is than a coin flip. After I continued with understanding the theory behind these algorithms: how are they derived and what calculations are needed in order to have a predicting model from the training dataset. With that knowledge I was finally able to look for datasets, and I selected five from different areas of science or life to demonstrate how agile these methods are. I managed to find such data sets that are frequently used by the community to test their new algorithms or methodology,

Date: May 25, 2018.

therefore I did not have to work too much on the data preparation, so I could concentrate on the models. I inspected how the models behave in three different aspects: parametrization, data availability and target class density. I performed an exhaustive serch over a predefined parameter grid in case of all sets and I used the best performing combinations for later investigation. Then I fitted the model on smaller and smaller part of the data and used the rest to calculate the Gini value. Then I prepared new datasets, where I removed some of the target class records and fitted the models on them using the standard 70/30 train/test splitting. I learned that decision trees are usually powerfull on most datasets, and that k-Nearest Neighbours and SVC are expensive to use. I also learned that medical uses are surprisingly powerfull, and that it is not worth to try to predict totally random distributions.

Future plans are examining the negative Gini value cases and the ones with extremely rare positive observations (which is extremely needed in case of predictive maintenance in factories) and exploring other models as well.