DAT200: Compulsory assignment 3

Titanic: Machine Learning from Disaster

Deadline: April 27th at 14:00.

Overview

The Titanic survival data is one of the most popular data sets used for introduction to machine learning. You will be participating in the open Kaggle competition (https://www.kaggle.com/c/titanic), competing against more than 11000 other contestants and their close to 60000 entries.

Data

The Titanic training data contains both numeric and text based predictors. Some of these can be used more or less directly, e.g. the ticket class, gender, family relations and fare. The age column contains many missing values. Some of these may be possible to impute/estimate if you combine the cabin and family relation information. DataCamp gives a suggestion on imputing with medians of other predictor combinations. It should be noted that missingness is not completely random in this data set (more data are missing in cheaper classes). The passenger names contain information on title (Mr., Mrs, …), which can be useful, and the ticket numbers have codes that may be of interest if you have extra time.

Discussions and tutorials

There is a lot to learn from other people's trials and mistakes. For instance, discussions like this: https://www.kaggle.com/c/titanic/discussion/14919 go through different strategies and recodings that lead to a maximum Kaggle score close to 80%. A point to take from many of these is to also include the test data when exploring relations and patterns between variables. Tutorials and blogs with bragging (including code) are quite common.

Analysis

The primary goal of the compulsory assignment is to use a handful of tools from DAT200 to create an entry to the Kaggle Titanic competition and an accompanying Jupyter Notebook with PDF export to submit to Canvas (including hyper-parameter tuning etc.). You are expected to perform:

- imputation: anything from the simplest tricks in scikit-learn to exploiting the structure of the data to create more tailored solutions,
- pre-processing: typically scaling and OneHots, possibly combining features,
- modelling and prediction: use tools from Chapter 6 to select classifier(s) and tune parameters, include tricks from Chapter 7 if you like to, or find external classifiers that can be included in the same framework, e.g. XGBoost.

You may co-operate on the compulsory assignment, but Canvas uploads are personal. Please, also add your Kaggle username and close co-operators in your Canvas uploads.