CatBoost applicability in regression problems

Project Proposal

Anna Berger, Ashwin Sadananda Bhat (AML 25)

December 17, 2017

1 Project Goal

The idea of our project is to explore the applicability of CatBoost: a recently released gradient boosting library with categorical features support out of the box. This machine learning method was developed by Yandex, which claims that CatBoost can achieve both superior quality when compared to other libraries and can make the process of dealing with categorical features a lot easier.
To investigate its suitability in real-life problems, we will use it in Sberbank Russian Housing Market challenge hosted on Kaggle in which one should predict the sale prices of different realty objects on Russian market. It is a regression problem and we are interested in seeing how well gradient boosting approach can perform in this area. We will also compare the performance of CatBoost with the performance of other gradient boosting trees implementations (such as Gradient Boosting from sklearn and XGBoost library).
Our **project goal** is to explore a new gradient boosting library CatBoost when applied to one of the regression problems and compare its performance with existing gradient boosting libraries.

2 Dataset

The dataset for this project is provided by Sberbank for the Kaggle competition Sberbank Russian Housing Market. The train set contains the information about the housing market and corresponding prices from August 2011 to June 2015, and the test set from July 2015 to May 2016. Over 250 properties of individual transactions are described in this dataset including supplementary information about the local area of each property. Additional data on Russian macroeconomy and financial sector during the the time of investigation is also given in a separate file.

3 Methodology

During this project, we will perform an exploratory data analysis on the given datasets with the goal of understanding the properties of the data and choosing the most discriminative features. After that, we will comprise the training set from these features, compare different implementations of gradient boosting on decision trees and find out which of them performs better in the particular regression problem. Selection of technologies:

- Python
- Jupyter notebooks
- CatBoost, sklearn, XGBoost

4 Evaluation of Results

During the challenge Root Mean Squared Logarithmic Error (RMSLE) was used as the primary evaluation metric, so we will also make use of this metric during our research. This metric is calculated as follows:

$$RMSLE = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(\log(p_i+1)-\log(a_i+1))^2},$$

where p_i are predicted values, a_i — actual values.