Machine Learning Assignment 1 Report

Nikita Bogomazov
n.bogomazov@innopolis.ru

1 Motivation

Our main goals are: detecting the quality of a game session for possible improvements of session parameters and predicting the bitrate of a session for possible network optimization.

2 Data

We have 2 datasets at our disposal stream quality dataset and bitrate prediction dataset for the classification and regression tasks respectively.
Stream quality dataset contains additional categorical features that should be encoded
Both datasets contain the means and standard deviations for both the fps and round trip time parameters. We also have the information about dropped frames in each sessions.

3 Exploratory data analysis

By using the profiling report and pairwise plots of the present features we have concluded that data imbalance is present, as well as a number of outliers that had to be dealth with.

4 Task

As was stated earlier, we have to deal with two tasks: stream quality classification and bitrate prediction. For both of the tasks the necessary preprocessing, outlier detection and feature sellection was done accordingly.

4.1 Regression

For bitrate prediction task the selection of Linear Regression, Lasso, Ridge and Polynomial Regression (of degree 2) models was chosen.
We have started with a baseline model without preprocessing, slowly iterating over different features and outlier detection. When all necessary actions were done we proceeded with training of all discussed models (Linear Regression, Lasso, Ridge, Polynomial Regression) and compared their performance.

4.2 Classification

The same aproach was taken with the classification of stream quality task: baseline model (Logistic Regression) was created, the performance was quite poor so futher preprocessing was needed. After the data imbalance problem was found necessary down sampling of one of the classes was done which helped to achieve adequate performance on the test dataset

5 Results

For classification task the final performance of our Logistic Regression showed promising results both on train and test datasets Higher scores on the test dataset show that our model didn't overfit and could generalize all necessary features.

Table 1. Stream Quality Classification with Logistic Regression

Dataset	Acc.	Recall	Precision
Train	0.7362	0.7362	0.7767
Test	0.8953	0.8953	0.9221

For regression task all models showed similar performance with a slight drop in all parameters for Polynomial Regression of degree 2

Table 2. Bitrate Prediction on test dataset with different Regressions

Model	R2 sccore	MAE	RMSE
Linear Regression	0.8934	1077.314	1949.594
Lasso	0.8934	1076.988	1949.314
Ridge	0.8934	1077.313	1949.593
Polynomial(2)	0.8807	1063.009	2062.420

6 Data Imbalance

Data Imbalance was a major problem for the classification task. We have picked the data down sampling approach to counter it which resulted in higher performance.

7 Conclusion

We were able to successfully solve both task achieving 89 percent or higher performance on both test datasets.
Data preprocessing, outlier filtration and data imbalance countering (for classification task) plays a major role in data analysis which results in adequate performance of chosen models.
In this assignment we were able to practice and demonstrate our knowledge of both data preprocessing and selection of appropriate models for a given task.