[F22] MACHINE LEARNING ASSIGNMENT 1

Due Date: 5 October 2022 (14:10)
Submission Format: GitHub repository link and report (PDF).
Data: [Regression], [Classification]

CLOUD GAMING

Cloud gaming is a new way of online gaming, which renders the game data on the high-speed cloud server instead of the end user's system and is forwarded via a high-speed network. However, users do not always have ideal or stable internet for high streaming quality which leads to loss of data packets during data transfer. In this assignment you will use machine learning to detect unstable network which can be used to take steps towards adapting the data streaming and minimizing data packet loss.

Expected outcomes:

- Data preprocessing and visualization (5 points)
- Feature selection and engineering: Selection of best data features (15 points)
- Prediction the bitrate to send data packets in a stream using three or more appropriate machine learning models (i.e Linear regression, polynomial regression, etc..) whereby at least 1 machine learning model has regularization (Lasso or Ridge) - **Regression task** (15 points)
- Detection of stream quality (good or bad) - **Classification task** (10 points)
- Comparison of the selected machine learning models' performance using the appropriate evaluation metrics. (10 points)
- Describe which model is better based on the test and training set performance. Does the model overfit? Underfit? (5 points)
- Removal of outliers (5 points)
- For classification task : Balancing of data to improve model performance (10 points)
- Report and source code (25 points)

DATASET

The Datasets comes from Innopolis University partner company providing a cloud gaming service. Each entry of the dataset file is represented by features derived from the five variables. Description of the features can be found in the README file for each dataset. The concrete features in the dataset are derived using statistics such as mean, standard deviation and maximum. The

dataset contains numerical and categorical attributes. This format is not very friendly for learning algorithms. Further, we are going to discuss how to preprocess the data before passing it to the ML algorithm.

DATA PREPROCESSING

Data preprocessing is a crucial step in data analysis. The simplest way to convert the string representation into the machine-readable format is to substitute the characters with a unique integer identifier. This can be easily achieved by using Label encoder from sklearn. You are free to apply other ways for handling categorical data. You can explore more encoders from library called **Category Encoders**. Some models such as neural networks work best with scaled data, therefore to scale the data as part of the preprocessing stage sklearn provides different methods for scaling the data (i.e Standard scaler, Min-Max scaler)

FEATURE SELECTION

It is desirable to reduce the number of input variables to both reduce the computational cost of modeling and, in some cases, to improve the performance of the model. Methods such as Lasso have feature selection embedded to their algorithm. However there exist other methods which result in improved model prediction performance (i.e statistical based). For feature selection in this assignment you will have to use one of the methods introduced in the labs or Lecture.

DATA VISUALIZATION

For data visualization on 2D plane dimension reduction methods such as Principal component analysis (PCA) can be used. The simplest way is to select one meaningful feature and plot it against the target variable. We suggest plotting the independent variables against the dependent variables. To better understand the data and create the data profile you can use pandas-profiling, which will generate report one line of code.

MACHINE LEARNING MODELS

For the regression task (prediction of bitrate), you will need to select the appropriate machine learning algorithm. From the Course of machine learning, you have studied several machine learning algorithms (i.e linear regression, logistic regression, polynomial regression, etc.). If you decide to use an algorithm taking one variable as input, RTT or FPS related feature should be used as an independent (predictor) variable. A minimum of 3 machine learning algorithms should be used for the prediction of bitrate. One of the algorithms should have regularization.

For the classification task (detection of stream quality), you will need to select the appropriate machine learning algorithm. From the Course of machine learning, you have studied several

machine learning algorithms (i.e Logistic regression). Logistic regression is the minimum machine learning algorithm to be used for the detection of stream quality. Logistic regression should be used with L1 or L2 regularization.

PERFORMANCE MEASUREMENT

To measure the performance of the selected models there, exist a number of metrics studied in the course of machine learning (i.e. MSE, precision, recall, RMSE, F1-score, and weighted F1-score). Compare the trained machine learning models in regression and classification task using the appropriate metrics (use minimum of two metrics).

DATA BALANCING AND OUTLIER DETECTION

The classification task dataset is imbalanced and contains outliers. This possess threat to the learning and performance of the ML model. There exist different ways to balance the dataset and remove outliers. The main task is to select one data balancing approach from literature, remove outliers using an approach from literature, balance the data and evaluate the impact of balancing the data on one the model used in the classification task. For a start you can explore **imbalanced-learn** and [1, 2, 3, 4]

REPORT AND SOURCE CODE

After performing the comparison of the machine learning models, the results should be presented in a form of a report. The implementation should be in python. The implementation repository should be available in GitHub or GitLab.

Your repository should contain

- main script and well documented Jupyter Notebook
- Readme file (how to run the main script)
- Documentation (code documentation and Readme)

Your report should contain

- **Motivation** : explanation (written in your own words) of the importance of the two tasks that you solved from the perspective of cloud gaming.
- **Data** : Brief description of the both Regression and Classification data (features and predictor)
- **Exploratory data analysis** : What are the insights from data exploration and did these insights help in feature selection or model design?
- **Task**: The definition of the learning tasks in terms of machine learning, i.e. estimating function

- **Input Format**: If you used an alternative data input format, explain it.
- **Comparison of selected ML models**: Describe which model is better in each task based on the cross-validation performance. Does the model overfit? underfit? How did you avoid both? Use graphs and tables to document the results of your experiments.
- **Outlier detection**: Does removal of outliers improve the performance of the selected ML model?
- **Data imbalance**: How did you solve class imbalance and did it help?

The report should be submitted in PDF format. The report should not be more than 2 pages using Association for Computing Machinery (ACM) - SIGPLAN Proceedings template. (if specified report format is not used, 5 points will be deducted from the overall grade) **Template Link**