C-M-005: Machine Learning I Lab. ’20-21

Final Project – Server Load Prediction

Due : November 23, 11:59PM

1 Problem

The final project for the course involves the problem of Server Load Prediction. You are being provided a dataset containing details of various aspects of a web server running a variety of different software applications.

The dataset consists of a set of variables that were measured over about a one month period. The data is not sequential and each data point can be considered a snapshot of the server state. Given this snapshot, you need to predict whether the server is on low, medium or high load. Measurements were taken in one minute intervals and on each server. They are usually the average or sum over that one minute interval. For instance, the number of packets received, the average number of IO operations, etc.

More information and the dataset can be obtained from the Kaggle contest page : https://www.kaggle.com/c/plaksha21-cm005-project

For an introduction to using Kaggle, follow this tutorial : https://www.kaggle.com/alexisbcook/titanic-tutorial

2 Data Fields

m_id	the ID of the server the data was sampled at.
appxxxx	data about specific application.
pagexxx	data on memory usage of the server.
syst_xxx	data on page fault rate, number of processes, etc.
state_xxx	data on the state the system is in.
io_xxx	data about general IO usage, (file IO, direct IO).
tcp_xxx	data on incoming and outgoing TCP traffic.
llxxx, ewxxx	data on incoming and outgoing network traffic.
cpu_load	the output class : one of low, medium or high

3 Submission

You will be provided with a test set without the true labels. Using the model built from the training data, obtain predictions on the test set and upload it to the Kaggle interface for obtaining the performance of your model. Each day, Kaggle allows a maximum of 10 submissions.

Till the deadline, the performance will be based on 30% of the test set and visible on the Public Leaderboard on Kaggle. After the deadline, all the submissions will be evaluated on the entire test set to get the final Private Leaderboard.

Apart from submission on the Kaggle Website, you need to submit all your code on moodle. The code must be able to reproduce the results that are obtained on the Kaggle submission. Please document the code well with details on the exploratory analysis, modelling choices etc.

3 Grading

Grading will be done based on the combination of following 2 criteria :

1. Correctness of the final model on the test set
2. Correctness of the code submitted

The top N performing fellows will be awarded with prizes.

4 Rules

1. The submitted code must be able to reproduce results obtained from the Kaggle submission in a reasonable amount of time.
2. To ensure that everyone uses the same set of tools to solve the problem, there is a restriction in the libraries that can be used for the contest. The following is the list of libraries allowed :
    a. Numpy
    b. Scipy
    c. Pandas
    d. Matplotlib
    e. Scikit Learn
    f. Keras and TensorFlow

   You are free to use any functionality available in these libraries.
3. Updates, if any, will be posted on the Kaggle page.