Project
Learning and Intelligent Systems
SS 2015

LAS Group, Machine Learning Institute
Dept. of Computer Science, ETH Zürich
Prof. Dr. Andreas Krause
Web: http://las.ethz.ch/courses/lis-s15/
Email questions to:
lis2015@lists.inf.ethz.ch

Project 1, Mar 2nd, 2015
Regression

You should not use any other data other than those that we provide you. You are also not allowed to hand-label the given data.

1 Introduction

The rail authority has been measuring the number of passengers that travel on some specific route with the goal of optimizing both custom satisfaction and financial costs. They are specifically interested in predicting this number given some measurable parameters so that they can send a suitable number of train cars. Your task in this project is to predict the number of passengers that will board the train given the time and day of departure, and some parameters about the weather.

2 Input and output specification

You are given the following four files.

- `train.csv` — The features of the training data.
- `train_y.csv` — The target variable for the training data.
- `validate.csv` — The features of the validation data.
- `test.csv` — The features of the testing data.

The files containing the features have one example per line and the features are delimited with commas. Each line has the following format

`YYYY-MM-DD HH:MM:SS,<A>,<B>,<C>,<D>,<E>,<F>`

where the fields `<A>`,`<B>`,`<C>`,`<D>`,`<E>` and `<F>` contain numbers (the parameters about the weather). The first field is the timestamp indicating the time of departure (`YYYY` is the year, `MM` is the month and `DD` is the day). The third field, denoted by `<B>`, is categorical taking values in $\{0, 1, 2, 3\}$. All other fields are real and non-negative. As an example, consider the following lines.

```
2013-11-09 18:41:00,0.595,2,0.26,13.531,0.006,0.476
2013-02-12 00:31:00,0.106,3,0.268,13.633,0,0.885
2014-09-14 09:27:00,0.637,2,0.796,21.902,0.368,0.526
2014-01-25 01:40:00,0.199,1,0,19.419,0.17,0.736
2014-12-15 20:06:00,0.523,2,0.592,22.015,0.269,0.906
2014-05-22 04:10:00,0.556,2,0.528,21.133,0.169,0.691
```

The file `train_y.csv` has one non-negative number per line, one for each corresponding training data point (the files `train.csv` and `train_y.csv` have the same number of lines). The solutions that you will submit should be in the same format — one prediction per line.

3 Evaluation and Grading

You have to provide two files of predictions — one for the validation dataset, and one for the testing dataset. You should produce files that contain one number per line, which is the prediction for the corresponding row (same format as `train_y.csv`). If the true measurements are $\mathbf{y}$ and your predictions are $\mathbf{y}'$, you will be evaluated using the following loss function

$$\ell(\mathbf{y}, \mathbf{y}') = \sqrt{\frac{1}{n} \sum_i \left[\log(1 + y_i) - \log(1 + \max\{0, y_i'\}) \right]^2}.$$

We will compare the loss of your submission to two baseline solutions: a weak one (called "baseline easy") and a strong one (called "baseline hard"). The grade is computed as the *maximum* of the following two percentages.

- Perc_A — Equal to 50% if you are performing at least as good as the easy baseline on the *validation set* and 0% otherwise. Hence, by looking at the ranking you can immediately know if you will receive at least 50% of the grade.
- Perc_B — Let the losses of the easy baseline and the hard baseline on the *test set* be BE and BH respectively. If we denote the loss that you reach on the *test set* as E, you will obtain

$$\text{Perc}_B = \begin{cases} 0\% & \text{if } E > BE \\ 100\% & \text{if } E \leq BH \\ \left(1 - \frac{\text{E}-\text{BH}}{\text{BE}-\text{BH}}\right) \times 50\% + 50\% & \text{otherwise} \end{cases}.$$

3.1 Report

You are requested to upload a ZIP archive containing the team report *and* the code. We included a template for LaTeX in the file `report.tex`. Please keep the reports brief (under 2 pages). If you do not want to use LaTeX, please use the same sections as shown in `report.pdf`. The reports are uploaded on the same page as the test set submissions.

3.2 Deadline

The submission system will be open from **Monday, 02.03.2015, 18:00** until **Sunday, 15.03.2015, 23:59:59**.