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 2, Mar 16nd, 2015
Two-Label Classification

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. You can make at most 200 submissions on the validation dataset.

1 Introduction

In this project, you are going to classify biomedical images of human tissue into two categories — one with seven types and one with three. Starting with a set of images, several features have been automatically extracted using image processing techniques. Based on these features, you are asked to solve a multi-label multi-class classification problem, by predicting to which two types each sample belongs.

2 Input and output specification

You are given the following four files.

- `train.csv` — The features of the training data.
- `train_y.csv` — The labels 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

<A>,<B>,<C>,<D>,<E>,<F>,<G>,<H>,<I>,<K+>,<L+>

where the fields <A>–<I> contain numbers representing the parameters about the geometrical and texture-related features. The field <K+> consists of four binary features representing a four-valued categorical feature in one-of-k format[1]. Similarly, the field <L+> consists of 40 binary columns in one-of-k format.

The file `train_y.csv` contains two categorical outputs per line in the following format

<Y>,<Z>

where <Y> takes values in $\{1,2,3,4,5,6,7\}$ and <Z> in $\{0,1,2\}$. There is one such line for each corresponding feature vector in `train.csv`, that is, the files `train.csv` and `train_y.csv` have the same number of lines. The solutions you submit should be in the same format—*two* numbers per line, separated by a comma.

[1] This means that one and only one of the k fields is equal to 1 and the rest are 0.

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 *two* comma-separated numbers per line, which are the predictions for the corresponding row (same format as `train_y.csv`). Let us denote the true labels of the two output fields as $\mathbf{y}$ and $\mathbf{z}$. If your predictions for these two labels are $\mathbf{y}'$ and $\mathbf{z}'$ respectively, the submissions are evaluated using the classification error across both output labels, that is[2],

$$\ell(\mathbf{y}, \mathbf{z}\,;\, \mathbf{y}', \mathbf{z}') = \frac{1}{2n}\sum_{i=1}^{n}[y_i \neq y_i'] + \frac{1}{2n}\sum_{i=1}^{n}[z_i \neq z_i'].$$

We will compare the score 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 scores of the easy baseline and the hard baseline on the *test set* be BE and BH respectively. If we denote the score that you reach on the *test set* as E, then you will obtain a score of

  $$\text{Perc}_B = \left(1 - \frac{\text{E} - \text{BH}}{\text{BE} - \text{BH}}\right) \times 50\% + 50\%.$$

  If you perform better than the hard baseline, you will receive $\text{Perc}_B = 100\%$.

3.1 Report

You are requested to upload a ZIP archive containing the team report *and* the code. We have 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 in `report.tex`. Reports are uploaded on the same page as the test set submissions.

3.2 Deadline

The submission system will be open until **Sunday, 29.03.2015, 23:59:59**.

[2] The notation $[x]$ means the expression evaluates to 1 if x is true and 0 otherwise.