SM4 Statistical Machine Learning - assessed practical HT 2018

This practical has an associated Kaggle-in-class challenge:

https://www.kaggle.com/c/sm4-assessed-practical

The data is a set of features extracted from images of women and men. This is a classification problem with a binary response, i.e. whether an image x represents a woman (y=1|x) or a man (y=0|x). You are asked to provide an estimate of P(y=1|x), and prediction performance will be evaluated using the log-loss. You are free to use any machine learning method and model discussed in the course (and beyond), as long as you describe clearly in the report all the steps and choices you have made. Note that the data you have been given is a shuffled and transformed version of the original data, so there would be no use in looking for the data set online.

While getting a good prediction performance of your method will be important, remember that you will be assessed based on the quality of your report, so explaining your steps and choices clearly and discussing all the issues you have faced in this challenge will be essential. Besides explaining your final solution, please briefly describe some of the other techniques you have tried and include a brief description of the more computational aspects of your work. For instance, did you try improving the training and/or prediction speed?

The report has a limit of 2,500 words. Please be as concise as you can. You should work in teams of 4 participants (to account for class size, one group may be composed of only 3 participants). Remember to place your team name, which consists of the collated anonymous IDs of all group members, on the cover page of the report. Please name the pdf file of your submitted report using the list of anonymous IDs of all team members, separated by underscores, e.g. P001-P002-P003-P004.pdf. Please include the code you used to get your final score. Please make sure the code is readable (i.e. it contains comments explaining what you are doing).

Data

The images have been preprocessed for you. You cannot see the original images (it would be very easy to predict the label yourself). You will be working with 128 features that compactly represent the image data. The total number of training images is 150,000, stored (one per line) in the train.data.csv file. The file contains a header and one line per image. The training labels are stored in the train.labels.csv file. This file also contains a header, and a one line per label. Note that there is a one-to-one ordered correspondence between the training data file and the training labels file.

The testing data is provided to you in the file test.data.csv. This file has the same format as the training data file, i.e. there is a header, followed by one line per image. You are not given labels for these images, and your task is to provide an estimate of P(y=1|x), i.e. the probability that the image contains a female.

Submission and evaluation

Each team should use a single anonymous Kaggle account. You can make up to 3 submissions every day.

Format: Submission files should contain two columns: Id and ProbFemale (a number between 0 and 1). Id refers to the indices of the testing file (the line number, excluding the header).

The file should contain a header and have the following format:

Id,ProbFemale

1,0

2,0

3,0

4,0

...

You can download the sample.csv file to see an example (the labels were randomly generated in this file).

Evaluation Metric: Accuracy will be assessed using the log-loss, so being confident about an incorrect prediction will be penalized. When submitting the predictions to Kaggle, you will instantly see the accuracy of your method on approximately 30% of the test set. At the end of the competition, the results on the other 70% will be available.

Deadline

The submission deadline is Wednesday March 14th at 12:00pm.

Good luck!