Midterm

Machine Learning COMS 4771, Spring 2017, Itsik Pe'er

Assigned: Wednesday March 22nd 2:30pm Due: Friday, March 24th 23:55pm

Submission: Your submission folder on Courseworks. Submit a zipped folder for `Midterm`

You are provided with a dataset of N datapoints in the pickle file `Data_x.pkl` along with their respective binary classification labels `Data_y.pkl` .These files respectively includes variables:
`X`: $N \times D$ numpy array of real values, interpreted as vectors that are *available training input* datapoints in D dimensions
`y`: $N \times 1$ numpy array of {1,2} values, interpreted as *available training output* labels in one dimension

Implement a python function `predict` that receives an input $M \times D$ numpy array of real values, interpreted as vectors that are input datapoints in D dimensions unavailable for training. `predict` returns an output $M \times 1$ numpy array of {1,2} values that are your respective predictions for the class labels of the M input datapoints.

You will be evaluated (50% of grade) on testing data drawn from the same distribution as the training data (classification accuracy). If `predict.py` uses any other `Python` code or libraries that you wrote, they should be included in your submission folder. `predict` is intended to use the results of learning, rather than involve compute-intensive training of a classifier. Specifically, it is supposed to be fast, running for under a minute per M=1000 datapoints on a standard machine.

Besides `predict.py` you are further required to submit an explanation of your solution (50% of the grade). This should be submitted electronically in typeset file `explain.pdf` , limited to two pages in 11pt font including any charts/visuals, not including bibliography listing.

You are free to use any method discussed in the course before spring recess. Cite any references you use. If your explanation relies on any computation, e.g. training, failed attempts, or statistics extracted from the data, include `python` code for all of these in a file called `train.py` in your submission folder.

Recall, that this is an exam. You are not allowed to discuss the exam with any person besides the course staff. You are not allowed to post any message about it on any forum, user community or social media outlet, with the exception of private messages to the course staff on Piazza. Course staff will be on call 9am-6pm and if possible beyond in order to respond, but will not offer office hours during the midterm.

The exam is open book in the sense that you may use class material from courseworks, and any reading material that was generated by others and you obtained legally before the start of the midterm. In order to timestamp such material, upload a list of any such printed material and a copy of any such electronic material to your course Google Cloud `Midterm` folder before the start of the midterm. `python` and `scikit-learn` documentation are exempt from the timestamp requirement, and you can use any `python scikit-learn` code that implements material covered in class before spring recess under the same conditions. You may search for additional electronic information at the course courseworks site or across scikit-learn documentation, but not beyond that.

Good luck!