Project 2

ESE 545, Data Mining: Learning from Massive Datasets

October 12, 2018

Due at 11:59PM on **November 5, 2018**

These related problems consist of several parts. You are required to solve them using Python and turn in your code. You are allowed to work in groups of at most two members. You should submit your code with a brief report containing responses to each part. **Each group member should turn in a different report, but you may share your code**. Upload a zipped file containing the report and the code on Canvas.

Problem 1. Using Python 3, install the `tensorflow`, `sklearn` and `keras` packages. We will be using the RCV1 dataset, which is an archive of around 800,000 manually-categorized news stories. Import the RCV1 dataset using

```
from sklearn.datasets import fetch_rcv1
rcv1 = fetch_rcv1()
```

The data is presented as rows of articles with features as columns in `rcv1['data']`. The topics of the articles are presented in `rcv1['target']`, where rows are articles and columns are topics. We will be training classifiers to determine whether articles belong to the `CCAT` topic (column 33 in `rcv1['target']`), which corresponds to corporate and industrial news. Roughly 380,000 articles relate to this topic.

(a) Make a new label vector, where each article has a 1 if it has been classified as `CCAT` and a -1 otherwise. **10 points**

(b) The first 100,000 articles will be used for training and the remaining articles will be used to test. Split the data and labels into training and test sets. **5 points**

Problem 2. Using PEGASOS, train an SVM on the training articles. Make a plot of training error vs. number of iterations and include it in your Report. Make sure to justify the selection of any parameters (with evidence) in your Report. You may not use any libraries for this task. **20 points**

Problem 3. Using AdaGrad, train a classifier on the training articles. Plot the training error vs. number of iterations in the same plot as above. Make sure to justify the selection of any parameters (with evidence) in your Report. You may not use any libraries for this task. **20 points**

Problem 4. You will now train a neural net on the training articles. We strongly suggest using `keras` for this task.

(a) Train neural nets over 5 epochs with 1, 2 and 3 hidden layers, each with 100 hidden units. Plot the training error and include it in your report. **10 points**

(b) You will now try to design the best neural net you can for this task. Train several NNs, each with no more than hidden 6 layers. Create a table of the training error for different numbers of hidden units in your hidden layers and include it in your Report. You should try several different values of these parameters (at least 5 combinations), and justify your final selection (with evidence) in your Report. Your NN must take less than 10 minutes to train for any of your choices of parameters. **25 points**

Problem 5. Evaluate your best SVM, AdaGrad classifier and NN from Problems 3-5 on the test articles and report the test error. You should use the parameters you selected previously. Which classifier did best? Why do you think that is? **10 points**