Problem Set 05
Data Mining and Machine Learning – Spring 2016
Due date: 2016-03-14 13:00

All assignments must be uploaded to the assignments tab in ClassesV2 (notice that this is **not** the dropbox) by the date and time specified. Make sure that you follow the instructions exactly as described. You may discuss problem sets with others, but must write up your own solutions. This means that you should have no need to look at other's final written solutions.

You need to turn in all of your solutions as a zip compressed file, named `netid_pset05.zip`, with your actual netid filled in in all lower case letters. This archive should contain just one file:

- `pset05.pdf`

The contents of this file is described below.

Exploring Neural Networks

For this problem set you will need to work with python, though no actual 'coding' is required (everything should be straightforward following the example script and the only thing we will grade is your set of responses). If you have not already download the Anaconda version of Python 3.5 from here:

`https://www.continuum.io/downloads`

Then, download a zip file of the problem set data and code from here:

`http://www.stat.yale.edu/~tba3/class_data/pset05.zip`

Unzip the file and then look at the script `starter_code.py` for example of how to fit neural network models using the code.

The goal of this assignment is to fit neural networks to predict categorical responses from two different datasets, MNIST and CIFAR-10. Descriptions of both are found here:

`http://yann.lecun.com/exdb/mnist/`
`https://www.cs.toronto.edu/~kriz/cifar.html`

The data is already nicely parsed and cleaned for you as part of the `pset05.zip` file. For CIFAR-10, the numerical categories correspond to the alphabetical ordering of categories on the website.

Once you have played around with the various tuning parameters for both datasets (note: it is normal for the algorithm to take a few minutes to run each time; I would suggest keeping the number of epochs between 2 and 20 as you test different configurations), the only thing you need to hand in is a pdf file in which you address the following questions:

- what combinations of parameters give you the best models for the two sets?
- what are the best validation accuracy rates you achieved?
- how much does the choice of cost function effect your results?
- how do the number of nodes or depth of the model change our results?
- how does halving and doubling the learning rate effect the models?
- how does regularization effect the classification rates?
- are certain categories more or less difficult to separate? What seems to be the hardest to distinguish?

Please answer these in a prose format; that is, complete sentences with a logical framework. You can organize them as you see fit, for example answering these separately for the two sets may be easier. Note that you will likely have poorer results for the CIFAR-10 datasets; we'll learn techniques for fitting this much better after the break.