CSE 417A: Homework 4

Due: October 9, 2014

Notes:

- Please keep in mind the collaboration policy as specified in the course syllabus. If you discuss questions with others you **must** write their names on your submission, and if you use any outside resources you **must** reference them. **Do not look at each others' writeups, including code.**
- Instructions for how to get files from the SVN repository are available on the course website and on Piazza.
- Homework (in hardcopy) is due **at the beginning of lecture.** In addition, your code submission for Problem 2 must also be timestamped before lecture begins.
- Please comment your code properly.
- There are 4 problems on 2 pages in this homework.
- **Keep in mind that problems and exercises are distinct in LFD.**

Problems:

1. (10 points) Repeat the problem of applying gradient descent to minimize E_{in} on the training dataset (`cleveland.train`) from Homework 3, but this time scale the features by subtracting the mean and dividing by the standard deviation for each of the features in advance of calling the learning algorithm (you may find the matlab function `zscore` useful). Experiment with the learning rate η (you may want to start by trying different orders of magnitude), this time using a tolerance (how close to zero you need each element of the gradient to be in order to terminate) of 10^{-6}. Report the results in terms of number of iterations until the algorithm terminates, and also the final E_{in}. How does this compare to the E_{in} of `glmfit`? You do not need to submit any code for this problem.

2. (60 points) For this problem, you will be doing LFD Problem 4.4 parts (a) through (d) with some changes / help / instructions / requirements. First, you can find headers for all the code you need to implement in your SVN repository for the class. There is also a matlab script called `run_expts.m` which you can use as an example for how to run your code to return the results we want. Second, read Problem 4.3 carefully. You can (and will need to) use the recurrence defined there as well as the formula in 4.3(e).

   (a) In addition to answering the question about why we need to normalize f, also prove that the term to normalize by is $\sqrt{\sum_{q=0}^{Q} \frac{1}{2q+1}}$ (hint: use the formula in 4.3(e)).

(b) Answer the question. For your implementation, we suggest you use `glmfit` with the additional options `'normal','constant','off'`.

(c) Answer the question (hint: use the formula in 4.3(e)).

(d) Implement the framework and answer the questions, with the modification that you only need to look at $Q_f \in \{5, 10, 15, 20\}, N \in \{40, 80, 120\}, \sigma^2 \in \{0, 0.5, 1.0, 1.5, 2.0\}$. Compute both the median and the mean of the overfit measure applied to many (at least 500) different datasets for each choice of parameters, and report how these measures vary as a function of the complexity of the true hypothesis, the number of training examples, and the level of stochastic noise (use line graphs). Explain your observations, and also comment on the differences you observe between the mean and median measures.

3. (15 points) LFD Exercise 4.4

4. (15 points) LFD Problem 4.8