CSE 417A: Homework 3

Due: October 2, 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 submissions must also be timestamped before lecture begins.
- Please comment your code properly.
- There are 5 problems on 2 pages in this homework.
- **Keep in mind that problems and exercises are distinct in LFD.**

Problems:

1. (40 points) Read the instructions on the course website (or Piazza) for how to check out files from the SVN repository set up for this assignment. The files `logistic_reg.m` and `find_test_error.m` are just function headers that need to be filled in. `find_test_error` should encode a function that, given as inputs a weight vector w, a data matrix X and a vector of true labels y (in the formats defined in the header), returns the classification error of w on the data (assuming that the classifier applies a threshold at 0 to the dot product of w and a feature vector x (augmented with a 1 in the first position in the vector to allow for a constant or bias term). `logistic_reg` should encode a gradient descent algorithm for learning a logistic regression model. Given the data matrix X, the true labels y, and the maximum number of iterations to run for `max_its`, as inputs, it should return a weight vector w and the training set error E_{in} as defined in class. Use a learning rate $\eta = 10^{-5}$ and automatically terminate the algorithm if the magnitude of each term in the gradient is below 10^{-3} at any step.
    - Implement the functions in the two files.
    - Read more about the "Cleveland" dataset we'll be using here: `https://archive.ics.uci.edu/ml/datasets/Heart+Disease`

- Learn a logistic regression model on the data in `cleveland.train` (be careful about the fact that the classes are $0/1$ – you should convert them to $-1/+1$ so that everything we've done in class is still valid). Apply the model to classify the data (using a probability of 0.5 as the threshold) in `cleveland.test`. In your writeup, report E_{in} as well as the classification error on both the training and test data when using three different bounds on the maximum number of iterations: ten thousand, one hundred thousand, and one million. What can you say about the generalization properties of the model?
- Now train and test a logistic regression model using the inbuilt matlab function `glmfit`. Compare the results with the best ones you achieved and also compare the time taken to achieve the results.

2. (20 points) Download the handwritten digits dataset from `http://amlbook.com/support.html` and familiarize yourself with the data. The matlab code for plotting images is also a useful tool to get acquainted with. Now, we will be working on the problem of deciding whether or not an image is a "1", using both the raw data (`zip.train` and `zip.test`) and another version of the dataset that extracts only two features, symmetry and intensity (`features.train` and `features.test`). Write a matlab script (which you will submit along with your code for Question 1) called `question2.m` that does the following:

   - Loads the data in `zip.train` and trains a logistic regression model using matlab's `glmfit` function for predicting the probability that an image is a "1" or not on that data.
   - Applies the model to classify the data (using a probability of 0.5 as the threshold) in both `zip.train` and `zip.test` (for the same problem of classifying whether an image is a "1" or not) and reports the classification error on both the training and test data.
   - Repeats the above two steps for `features.train` and `features.test`

   In addition to submitting your code, in your writeup, report the training and test set classification errors for both datasets. Interpret your results in the context of the generalization error of the model and the dimensionality of the data. Which model generalizes better? Are you sufficiently convinced by this one experiment? If not, describe another one you could do to become more confident in your interpretation.

3. (15 points) LFD Exercise 3.12

4. (15 points) LFD Problem 3.4

5. (10 points) LFD Problem 3.16