CSE 417T: Homework 3
Due: Wednesday, October 23rd, 2019 by 9 PM

Notes:

• Please check the submission instructions for Gradescope provided on the course website. You must follow those instructions exactly.

• Please download these files (two Matlab stub files and two data sets) for Question 1:
  - https://classes.cec.wustl.edu/~SEAS-SVC-CSE417T/find_test_error.m
  - https://classes.cec.wustl.edu/~SEAS-SVC-CSE417T/logistic_reg.m
  - https://classes.cec.wustl.edu/~SEAS-SVC-CSE417T/cleveland_train.csv
  - https://classes.cec.wustl.edu/~SEAS-SVC-CSE417T/cleveland_test.csv

• Your score on coding questions (Question 1) will be based on the exclusively on your written report. The code you submit is only used for checking correctness (i.e. resolving discrepancies in your writeup) and for plagiarism checking. Results included in the code submission will not be graded.

• Homework is due by 9 PM on the due date. Remember that you may not use more than 2 late days on any one homework and you only have a budget of 5 in total.

• 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.

• There are 5 problems and 1 bonus problem on 2 pages in this homework.

• Keep in mind that problems and exercises are distinct in LFD.

Problems:

1. (60 points) Complete the two stub files find_test_error.m and logistic_reg.m. Do not change the function definitions. Remember to submit your code to Gradescope. You are allowed to write supplementary scripts to help you complete the experiments below; you do not need to submit any additional code you write.

After completing the code, read more about the “Cleveland” dataset we’ll be using: https://archive.ics.uci.edu/ml/datasets/Heart+Disease

Learn a logistic regression model on the data in cleveland_train.csv; note that the labels in the data set are 0/1 so you will need to convert those to -1/+1 in order to ensure that everything we’ve done in class is still valid). Use a learning rate of \( \eta_0 = 10^{-5} \) and automatically terminate if the magnitude of every element of the gradient is less than \( 10^{-3} \). Initialize the weight vector to a vector of all zeros. Train the model three times with a different bound on the maximum number of iterations each time: \( 10^4 \), \( 10^5 \) and \( 10^6 \) (note that the termination condition based on the magnitude of the gradient still applies). Use each model to classify the data (using a cutoff probability of 0.5) in cleveland_test.csv.
(a) In your writeup, report $E_{in}$ (the cross-entropy error), the binary classification error on both the training and test data sets and how long the training process took (in seconds) for all three models. What can you say about the generalization properties of the logistic regression model? How does increasing the maximum number of iterations affect the model’s performance?

Now scale each feature by subtracting the mean and dividing by the standard deviation for each of the features in advance of calling the learning algorithm (hint: use the zscore function). Experiment with the learning rate $\eta_0$ by eliminating the iterations-based termination criteria and only terminating when the magnitude of every element of the gradient is less than $10^{-6}$.

(b) In your writeup, include the final $E_{in}$, the binary classification error on the test data set, the number of iterations required to terminate and the time required to train the model for $\eta_0 = 0.01, 0.1, 1, 4, 5, 6, 7, 7.5, 7.6$ and $7.7$. Did normalizing the data affect the performance of the model? What is the effect of increasing $\eta_0$?

2. (8 points) LFD Exercise 4.5
3. (6 points) LFD Problem 4.8
4. (10 points) LFD Problem 4.25 parts (a)-(b) only.
5. (16 points) LFD Problem 5.4
6. (Bonus 5 points) Write a multiple choice question related to the content of this homework (LFD Sections 3.3, 3.4, Chapters 4 and 5). Be sure to indicate the correct answer! See Lecture 2, Slides 3 and 4 for the rubric and guidelines to writing a good multiple choice question. If you write a great question, there’s a chance it will be included on the next exam!