Fall 2017 CS 4340/5340 Projects 2, 3, and 4

Project 2: Points 100; Due Oct 31 midnight
Project 3: Points 100; Due Nov 7 midnight
Project 4 (only for the grad section): Points 50; Due Nov 9 midnight

Project 2:

(a) Implement linear regression to obtain the regression of Y on X on the following data:

X	Y
1	0
2	1
3	0
4	1
5	0
6	1
7	1
8	1

Note that our X here is one-dimensional. (In class, we considered linear regression of Y on X when X was a d-dimensional feature-vector; when X is multidimensional, the regression is described by some as "multiple linear regression.")

You can solve this problem entirely mathematically, without having to write any code. Or you can write a program to do it (please do not use any ready-made package). Do whichever is convenient.
Clearly indicate your final solution. Submit a rough, hand-drawn sketch (scanned) of the given data along with your result along (or draw your sketch using some software, but no extra credit for fancy looks, no penalty for hand-drawing, either).

(b) Can you use your solution to the regression problem in (a) above as a classifier? Yes/No.

(c) If not, justify with brief reason(s). If yes, state the equation of the classifier (the decision boundary), and indicate how you would classify the two values x = 2.4 and x = 5.5.

Project 3. This is an extension of Project 2.

(a) Use the data set given in Project 2; assume that in that data set X represented the time (in weeks) during which a student did not do any coursework for the course(s) he/she was enrolled in, and let Y = 1 represent the fact that the student failed (with Y = 0 denoting passing). Implement logistic regression to obtain a prediction for the probability of failure, given the number of weeks of inaction (you will need to implement gradient ascent or gradient descent as part of this problem). Clearly indicate your final solution. Using your solution, predict the chances of a student passing a course if she does not study for (i) 3 weeks, (ii) 5 weeks.

(b) Can logistic regression be used for classification? Briefly explain.

(c) If there were two or more rows in the given data set with the same x–value but different y-values (e.g., (x=3, y = 1) and (x=3, y = 0)), would we still be able to obtain a valid logistic regression of Y on X, or would logistic regression make no sense in that case?

Please do not use any package; write C/C++/Java/Python code to implement everything from scratch. Please submit a single pdf (or Word) file on MyGateway. Please write legibly if you are submitting any hand-written (and scanned) stuff.

Project 4. [For graduate students only.] This is an extension of Project 3.

Re-do Project 3 replacing gradient ascent (descent) with stochastic gradient ascent (descent). Indicate the final solution obtained. Write a brief note on (i) the relative speed of computation (with respect to that in Project 3), (ii) the relative quality of the final solution (w.r.t that in Project 3).