ECE 595 Homework 4 Due: 8 PM, March 20

Using the attached data file, ex2data1.txt, perform a multivariate logistic regression with the two features, Exam 1 score and Exam 2 score for the admitted and not admitted classes. The decision boundary for the two-feature case is a straight line. (Features do not need to be normalized for this case.)

Note that the data file was created in MATLAB; hence, you can open it using

data = load('ex2data1.txt');

X = data(:, [1, 2]); y = data(:, 3); % y has values of 0 and 1.

Initialize the parameters using

theta = zeros(n + 1, 1); % The size of theta (n) must have been defined.

Write m-functions for plotting, computing the cost (J), gradient and updating theta (all three in one function, or more for ease in debugging). These functions, in turn, may call other functions such as for calculating the sigmoid and others.

Gradient calculation, hypothesis and parameter updating need to be carried out in vectorized form.

You need to turn in the following.

1. Scatter plot of the two features with y = 1 and y = 0 distinguished (similar to Fig. 1).

Fig. 1

1. Cost at initial theta of all zeros.
2. Parameter values for theta.
3. Cost after 400 iterations of logistic regression algorithm with .
4. Decision boundary for logistic regression superimposed on the scatter plot.
5. Training accuracy – No. of correct decisions/Total No. of students, in percentage.
6. Probability of getting admitted for a student with Exam 1 score of 45 and Exam 2 score of 85.

For the next part, you are given a training set of two-class data (MATLAB file: ex2data2.txt)

that is not linearly separable. The data comes from microchip testing for acceptable and rejected chips based on two features, Test 1 and Test2. For this case, you need to use 6th order polynomial features,  (27 elements) for X. Notice that the features now have widely different ranges; hence, each must be normalized using mean and standard deviation. With  , a complex decision boundary can be obtained using logistic regression. To avoid overfitting, you need to use regularization of all the parameters (except ) while minimizing the cost function in logistic regression. Develop the code for regularized logistic regression with  and run for 500 iterations with . If the error J does not reach a minimum, you may increase the number of iterations or change .

Fig. 2 shows an example of decision boundary.



Fig. 2

The decision boundary plot may be obtained using two arrays defined as

u = linspace(-1, 1.5, 50);

v = linspace(-1, 1.5, 50);

Treating u and v as input, obtain the polynomial features, normalize them and get the hypothesis values z for each pair of u and v. Now use contour(u, v, z, [0, 0]) to plot.

Try running your code for different values of  to see the effect of regularization. Turn in the following.

1. Scatter plot of the two features with y = 1 and y = 0 distinguished.
2. Cost at initial theta of all zeros.
3. Parameter values for theta.
4. Cost after 400 iterations of regularized logistic regression algorithm.
5. Training accuracy.
6. Plots of data along with decision boundary for at least three cases of .