**20CS6037 Fall2017**

**Assignment #2**

**Assigned 9/26/2017**

**Dye on BB 10/3/2017 11:59PM**

The data set is in hm2data.mat, which is a 97 x 4 data set: the first three columns contain the observations, the last column is the output (y).

**L2 – Regularization, Model Identification, Cross Validation**

Before proceeding, you should normalize the data, that is, if X is the data set (observations only), normalize it by subtracting from each column its mean and dividing it by the standard deviation. Do this, before you extend the data set by adding the extra column of 1s.

Example of standardizing the data set X.

Let the original data be

X =

75 23 83 8

45 91 54 44

8 15 100 11

**muX=mean(X)**

muX = 42.6667 43.0000 79.0000 21.0000

**stdX=std(X)**

stdX = 33.5609 41.7612 23.2594 19.9750

Now, create two matrices of the same size as X with repeated mu and stdX, respectively. Use repmat Matlab command:

**repstd=repmat(stdX,3,1)**

repstd =

33.5609 41.7612 23.2594 19.9750

33.5609 41.7612 23.2594 19.9750

33.5609 41.7612 23.2594 19.9750

**repmu=repmat(mu, 3, 1)**

repmu =

42.6667 43.0000 79.0000 21.0000

42.6667 43.0000 79.0000 21.0000

42.6667 43.0000 79.0000 21.0000

**standardizedX = (x-repmu)./repstd**

standardizedX =

0.9634 -0.4789 0.1720 -0.6508

0.0695 1.1494 -1.0748 1.1514

-1.0329 -0.6705 0.9029 -0.5006

Then the column means and column standard deviations will be 0 and 1, respectively.

**[mean(standardizedX) ; std(standardizedX)]**

ans =

0.0000 -0.0000 0 0

1.0000 1.0000 1.0000 1.0000

Divide your data randomly, into three subsets as follows:

**L2-regularization**

For this step divide your data set into two subsets of approximately equal sizes, as follows:

Train: approx. 50%

Test: remaining subset

For L2 regularization, you may modify the gradient descent approach used in the 1st assignment, or you may use the (normal) equation, which provides directly the solution. Recall that you will need to come up with a way to select the parameter 

Eliminate the attribute (column in X) corresponding to the smallest component of the weight vector.

**Cross-validation**

Now, with the reduced set of attributes, repeat training your model, this time without regularization, but using cross validation, to decide if the degree of the polynomial in X which relates the input data to the output y.

Do a k-fold cross-validation, where k = 3, as follows: divide the dataset X into three subsets, say X1, X2, X3, of approximately equal sizes. Alternatively use Xi, i=1, 2, 3 as for training, and the remaining Xj,Xl, j,l ≠ i, for testing. Repeat this for degrees d=1, and d=2. Be careful to use the same sets X1, X2, X3. Based on the average error on the validation set, select the degree of the polynomial.

Finally, divide the data sets into training and test subsets of approximately equal, and train again, using the attributes selected by the L2-regularization step, and the polynomial of degree d selected by the cross validation step.

Repeat this step 100 times. For each iteration, extract the modeling and generalization errors. Plot them (iteration, modelingError, generalizationError).

Also collect the min, max and average over 100 iterations, and report them.

For just one iteration provide the figures you provided in Assignment 1: the regression line for each of the attributes selected, and the cost function contours.