CPSC-340 Machine Learning and Data Mining 2004

Homework # 3

Due Friday, Oct 15 at 4pm.

NAME:________________________________

Signature:______________________________

STD. NUM: ____________________________

Email: _______________________________

General guidelines for homeworks:

You are encouraged to discuss the problems with others in the class, but all write-ups are to be done on your own.
Homework grades will be based not only on getting the "correct answer," but also on good writing style and clear presentation of your solution. It is your responsibility to make sure that the graders can easily follow your line of reasoning.
Try every problem. Even if you can't solve the problem, you will receive partial credit for explaining why you got stuck on a promising line of attack. More importantly, you will get valuable feedback that will help you learn the material.
Please acknowledge the people with whom you discussed the problems and what sources you used to help you solve the problem (e.g. websites, books from the library). This won't affect your grade but is important as academic honesty.
When dealing with Matlab exercises, please attach a printout with all your code and show your results clearly.

1. **(Linear supervised learning)**

   Download the Boston housing dataset from the course website. This dataset was taken from the StatLib library which is maintained at Carnegie Mellon University. It concerns housing values in suburbs of Boston.

   There are 506 measurements, 13 inputs and one output "MEDV". The inputs are:

   (a) CRIM: per capita crime rate by town
   (b) ZN: proportion of residential land zoned for lots over 25,000 sq.ft.
   (c) INDUS: proportion of non-retail business acres per town
   (d) CHAS: Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
   (e) NOX : nitric oxides concentration (parts per 10 million)
   (f) RM: average number of rooms per dwelling
   (g) AGE: proportion of owner-occupied units built prior to 1940
   (h) DIS: weighted distances to five Boston employment centres
   (i) RAD: index of accessibility to radial highways
   (j) TAX: full-value property-tax rate per 10,000 dollars
   (k) PTRATIO: pupil-teacher ratio by town
   (l) B: $1000(Bk - 0.63)^2$ where Bk is the proportion of blacks by town
   (m) LSTAT: Percentage lower status of the population

   Edit the following m-file that implements the maximum likelihood (least squares) estimator. Hand in the fixed code and the plot.

```
echo off; clear;

% LOAD THE DATA AND EDIT IT:
% ========================
data = load('housing.data');      % Load the data.
x = data(:,[1:3 5:9]);            % Input data: (a) to (c) and (e) to (i).
[n,d] = size(x);
x = [ones(n,1) x];                % Add 1 for bias term.
y = data(:,14);                   % Output data.


% CREATE THE TRAIN AND TEST SETS:
% ==============================
trainSize = 400;               % Number of training examples.
xTrain = x(1:trainSize,:);     % Training input data.
yTrain = y(1:trainSize,:);
xTest = x(trainSize+1:n,:);    % Test input data.
yTest = y(trainSize+1:n,:);    % Test output data.

% COMPUTE LEAST SQUARES (ML) ESTIMATE:
% ===================================
theta_ls = ???
```

```
% TEST THE LINEAR MODEL:
% =====================
yPredTrain = xTrain * theta_ls;  % Generate prediction.
yPredTest  = ???   % Generate prediction.

% COMPUTE THE PREDICTION ERRORS:
% =============================
trainError = mean( (yTrain-yPredTrain).^2 ); % RMS train error.
testError = ???    % RMS test error.
disp(' ');
disp('Errors');
disp('------');
disp(' ');
disp(['Train  = ' num2str(trainError)]);
disp(['Test   = ' num2str(testError)]);
disp(' ');

% PLOT THE TRAIN ERROR AND THE TEST ERROR:
% ======================================
figure(1)
clf;
subplot(211)
plot(1:trainSize,yTrain,'ro',1:trainSize,yPredTrain,'b')
title('Training set');
zoom on;
subplot(212)
plot(???)
title('Test set');
legend('True value','Prediction');
```