CPSC-540 Machine Learning 2007

Homework # 2

Due Monday Feb 22 in class.

NAME:________________________________

Signature:______________________________

STD. NUM: _____________________________

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. 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. **Conditioning and marginalisation:**

A band called Radiohead is inspired by an old band called The Beatles. 50% of music critics think the beatles was a great (G) band, 40% that it was moderate (M) and 10% that it was awful (A). These critics have also compiled the following table:

$$\begin{array}{cc|ccc} & & & B_2 & \\ & & G & M & A \\ & G & 0.8 & 0.1 & 0.1 \\ B_1 & M & 0.1 & 0.9 & 0 \\ & A & 0.2 & 0.3 & 0.5 \end{array}$$

The table says that the probability of a new band (B_2) being great given that the inspiring band (B_1) was great is $P(B_2 = G|B_1 = G) = 0.8$. Similarly, $P(B_2 = G|B_1 = M) = 0.1$, $P(B_2 = M|B_1 = A) = 0.3$, and so on. Note that the numbers in the rows add up to 1, so the table is a probability transition matrix.

(i) Given what the critics think of the Beatles and the fact that the Beatles inspired Radiohead, what is the probability that Radiohead is a great band?

(ii) What is $P(B_1 = G|B_2 = G)$?

2. **Eigenvalues and Gaussians:**

The purpose of this exercise is to revise eigenvalues and eigenvectors and to familiarize yourself with multivariate Gaussians and Matlab.

Consider the bivariate normal distribution $\mathcal{N}(\mu, \Sigma)$, where

$$\mu = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \quad \text{and} \quad \Sigma = \begin{pmatrix} 4 & 3 \\ 3 & 4 \end{pmatrix}$$

(i) Find the probability density at the point $x_0 = (1, 2)'$.

(ii) Compute the 2 eigenvalues and normalised eigenvectors (magnitude 1) of Σ. Form the diagonal matrix of eigenvalues Λ and a the matrix of corresponding eigenvectors Φ. Compute the whitening transform $A = \Phi\Lambda^{-1/2}$. Compute the distribution of $y = A'(x - \mu)$.

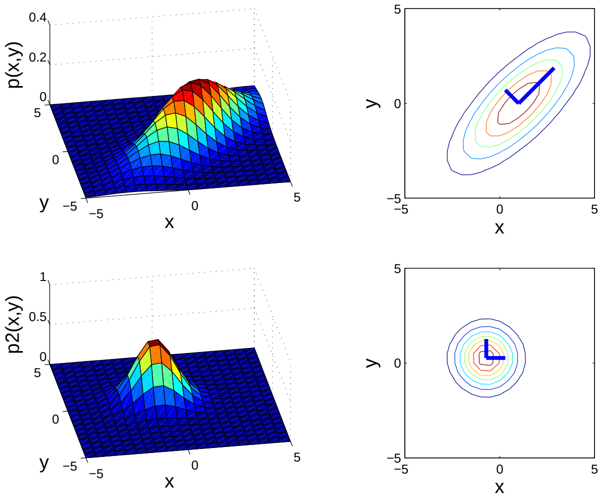
(iii) Do (i) and (ii) by hand and using Matlab. In Matlab, write a function `multigaussian(`x_0`,`μ`,`Σ`)` that returns the value of the Gaussian density at the point x_0. Also plot the contour plots and three-dimensional densities of $p(x)$ and $p(y)$. Plot the eigenvectors (centred at the means and of scale proportional to the square root of the eigenvalues) on top of the contour plots. What do you notice?

Some commands that will come handy are: `help`, `eig` (eigenvalues and eigenvectors), `meshgrid`, `mesh`, `surfc`, `contour`, `subplot`. To get familiar with plotting, type:

```
[x,y,z]= peaks(30);

surfc(x,y,z) % surf plot with contour plot.

xlabel('x-axis'), ylabel('y-axis','fontsize',15), zlabel('z-axis')

rotate3d on;
```

The last line allows you to rotate the plot using the mouse. To plot the graph, one can generate a postscript file using the command `print -dps` *filename*.

The answer should look something like this:



3. **Laws of Large Numbers**:

Let $X_{1:n}$ be a sequence of i.i.d. random variables with $\mathbb{E}(X_i) = \mu$ and $var(X_i) = \sigma^2$. Let also

$$\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

(i) Markov's inequality: For a random variable X and any positive measurable function $f(X)$ and positive scalar ε, show that

$$P(f(X) \geq \varepsilon) \leq \frac{\mathbb{E}(f(X))}{\varepsilon}$$

Hint: express $f(X)$ as $f(X) \geq \varepsilon \mathbb{I}_{f(X) \geq \varepsilon}$ and apply the expectation operator.

(ii) Chebyshev's inequality: Choose an appropriate $f(\overline{X}_n)$ in (i) to show:

$$P(|\overline{X}_n - \mu| \geq \varepsilon) \leq \frac{var(\overline{X}_n)}{\varepsilon^2}$$

(iii): Show that the asymptotic estimator of the mean is unbiased; $\mathbb{E}(\overline{X}_n) = \mu$.

(iv) Using the fact that for independent X_i's we have $var(\sum X_i) = \sum var(X_i)$, show that

$$var(\overline{X}_n) = \frac{\sigma^2}{n}$$

(v) Show the following weak law of large numbers:

$$P(|\overline{X}_n - \mu| \geq \varepsilon) \underset{n \to \infty}{\longrightarrow} 0$$

This statement says that the sample mean $\overline{X}_n$ converges to the true mean *in probability* as n goes to infinity. There is another mode of convergence, called *strong convergence* or almost sure convergence, which asserts a bit more. $\overline{X}_n$ is said to converge *almost surely* to μ if for every $\varepsilon > 0$, $|\overline{X}_n - \mu| \geq \varepsilon$ happens only a finite number of times *with probability 1*. Finally, to study the speed of convergence, one introduces *central limit theorems.*

4. **Least squares prediction:**

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:
% ====================================
???


% 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’);
```

5. **Ridge prediction:**

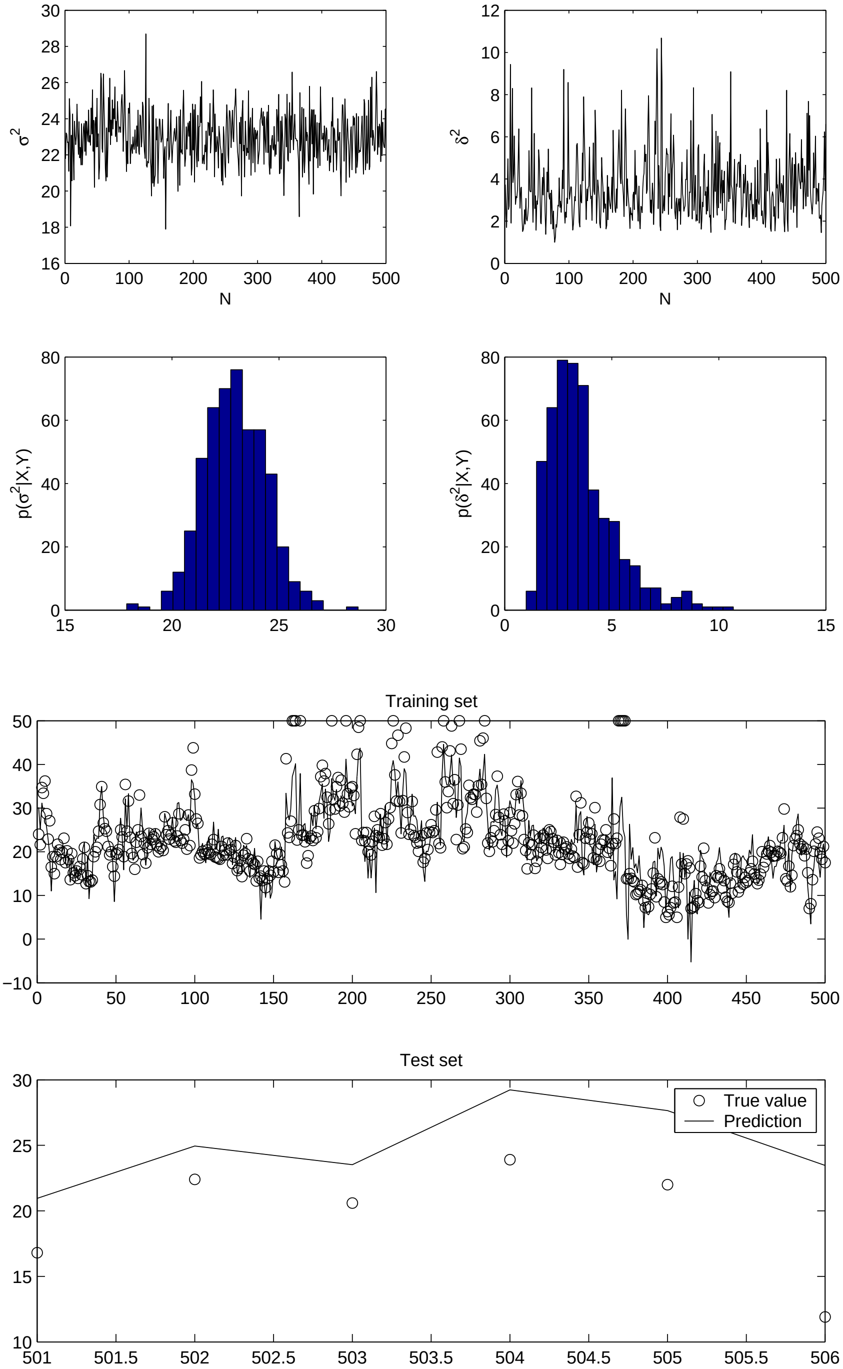
   Repeat the experiment of the previous question, but this time instead of least squares, you should use ridge regression. You should repeat the experiment for different values of the regulariser δ^2. Your code should generate a plot of the entries of the vector θ against the value of the regulariser δ^2. You should also use cross-validation to choose a good value of δ^2. Hand in the plots of training and test set errors, your code, the θ against δ^2 figure, and the values of the regularisers that provide the best results in a min-max sense.

6. **Gibbs Sampling for Linear Regression**:

Implement the Gibbs sampler for linear regression from page 93 of the lecture notes. Apply it to the problem of predicting house prices in Boston. The code should start with the following lines:

```
echo off; clear;

% LOAD THE DATA AND EDIT IT:
% ========================

data = load('housing.data');      % Load the data.
x = data(:,[1:3 5:13]);            % Input data.
[nn,dold] = size(x);
x = [ones(nn,1) x];               % Add 1 for bias term.
[nn,d] = size(x);                 % d is the dimension of theta
                                  % nn is the total number of data
y = data(:,14);                   % Output data.

% PRIOR PARAMETERS
% ================
a = 0; b = 0; alpha = 2; beta = 10;


% CREATE THE TRAIN AND TEST SETS:
% ==============================
n = 500;                % Number of training examples.
xTrain = x(1:n,:);      % Training input data.
yTrain = y(1:n,:);      % Training ouput data.
xTest = x(n+1:nn,:);    % Test input data.
yTest = y(n+1:nn,:);    % Test output data.
[n2,tmp] = size(xTest);

% INITIALIZATION:
% ===============
XY = xTrain'*yTrain; XX = xTrain'*xTrain; X = xTrain; Y = yTrain;
N = 500;                % Number of samples.
theta = zeros(d,N); mu = zeros(d,N); sigma2 = zeros(N,1); delta2 =
zeros(N,1); delta2(1) = inv(gengamma(alpha,beta));

for i=1:N,

???
```

Your code should be able to produce the following plots



Compare your training and test errors with the least squares solution.

7. **Multivariate Gaussian-Wishart Model**

   Give the following likelihood:

$$p(Y|X,\theta,\Sigma) = |2\pi\Sigma|^{-\frac{1}{2}}\, e^{-\frac{1}{2}(Y-X\theta)^T\Sigma^{-1}(Y-X\theta)}$$

   and the Wishart prior:

$$p(\Sigma|\alpha,\Sigma^*) \propto |\Sigma|^{-(\alpha+d+1)/2} e^{-\frac{1}{2}trace(\alpha\Sigma^*\Sigma^{-1})},$$

   derive an expression for the posterior distribution of Σ.