Homework 3

CSE 802 - Pattern Recognition and Analysis
Instructor: Dr. Arun Ross
Points: 100 (+10 bonus points)
Due Date: April 13, 2016

Please read the following instructions carefully:

1. You are permitted to discuss the following questions with others in the class. However, you must write up your own solutions to these questions. Any indication to the contrary will be considered an act of academic dishonesty.

2. A hard-copy of this assignment must be submitted by April 13, 3:00pm. Late submissions will not be graded. In this copy, please include the names of individuals you discussed this homework with and the list of external resources (e.g., websites, other books, articles, etc.) that you used to complete the assignment (if any).

3. Code developed as part of this assignment should be placed in a zip file and sent to rossarun at cse.msu.edu with the subject line “CSE 802: Homework 3”.

1. [10 points] Consider a set of \( n \) i.i.d. samples (one-dimensional training patterns), \( D = \{x_1, x_2, \ldots, x_n\} \), that are drawn from the following distribution:

\[
p(x|\lambda) = \frac{e^{-\lambda}\lambda^x}{x!}
\]

(a) Derive the maximum likelihood estimate of \( \lambda \), i.e., \( \hat{\lambda}_{mle} \).

(b) Consider a set of 1000 training patterns that can be accessed here. Plot the distribution after computing \( \hat{\lambda}_{mle} \) using these training patterns.

2. [10 points] Let \( x \) have a uniform density

\[
p(x|\theta) \sim U(0, \theta) = \begin{cases} 
1/\theta, & 0 \leq x \leq \theta \\
0, & \text{otherwise}.
\end{cases}
\]

(a) Suppose that \( n \) samples \( D = \{x_1, \ldots, x_n\} \) are drawn independently according to \( p(x|\theta) \). Show that the MLE for \( \theta \) is \( \max[D] \), i.e., the value of the maximum element in \( D \).

(b) Suppose that \( n = 5 \) points are drawn from the distribution and the maximum value of which happens to be 0.6. Plot the likelihood \( p(D|\theta) \) in the range \( 0 \leq \theta \leq 1 \). Explain in words why you do not need to know the values of the other 4 points.
3. [10 points] Let \( x = (x_1, \ldots, x_d)^t \) be a \( d \)-dimensional binary (0 or 1) vector with a multivariate Bernoulli distribution

\[
P(x|\theta) = \prod_{i=1}^{d} \theta_i^{x_i} (1 - \theta_i)^{1-x_i},
\]

where \( \theta = (\theta_1, \ldots, \theta_d)^t \) is an unknown parameter vector, \( \theta_i \) being the probability that \( x_i = 1 \). Let \( D = \{x_1, \ldots, x_n\} \) be a set of \( n \) i.i.d. training samples. Show that the maximum likelihood estimate for \( \theta \) is

\[
\hat{\theta} = \frac{1}{n} \sum_{k=1}^{n} x_k.
\]

Explain in words what this means.

4. [30 points] Consider a two-category (\( \omega_1 \) and \( \omega_2 \)) classification problem with equal priors. Each feature is a two-dimensional vector \( x = (x_1, x_2)^t \). The true class-conditional densities are:

\[
p(x|\omega_1) \sim N(\mu_1 = [5, 5]^t, \Sigma_1 = I),
p(x|\omega_2) \sim N(\mu_2 = [10, 10]^t, \Sigma_2 = I).
\]

Generate 50 bivariate random training samples from each of the two densities.

(a) Write a program to find the values for the maximum likelihood estimates of \( \mu_1, \mu_2, \Sigma_1, \) and \( \Sigma_2 \) using these training samples (see page 89, use equations (18) and (19)).

(b) Compute the Bayes decision boundary using the estimated parameters and plot it along with the training samples. What is the empirical error rate on the training samples?

(c) Compute the Bayes decision boundary using the true parameters and plot it on the same graph. What is the empirical error rate on the training samples?

(d) Repeat (a) - (c) after generating 1,000 and 10,000 random training samples from each of the two densities. How do the estimated parameters and the empirical error rate change in (b) when the number of representative training samples increases?

5. [20 points] The iris (flower) dataset consists of 150 4-dimensional patterns (i.e., feature vectors) belonging to three classes (setosa=1, versicolor=2, and virginica=3). There are 50 patterns per class. The 4 features correspond to sepal length in cm \((x_1)\), sepal width in cm \((x_2)\), petal length in cm \((x_3)\), and petal width in cm \((x_4)\). Note that the class labels are indicated at the end of every pattern.

Assume that each class can be modeled by a multivariate Gaussian density, i.e., \( p(x|\omega_i) \sim N(\mu_i, \Sigma_i) \), \( i = 1, 2, 3 \). Design a Bayes classifier and test it by following the steps below:

(a) Train the classifier: Using the first 25 patterns of each class (training data), compute \( \mu_i \) and \( \Sigma_i \), \( i = 1, 2, 3 \). Report these values.

(b) Design the Bayes classifier: Assuming that the three classes are equally probable and a 0-1 loss function, write a program that inputs a 4-dimensional pattern \( x \) and assigns it to one of the three classes based on the maximum posterior rule, i.e., assign \( x \) to \( \omega_j \) if,

\[
j = \arg \max_{i=1,2,3} \{P(\omega_i|x)\}.
\]
(c) Test the classifier: Classify the remaining 25 patterns of each class (test data) using the Bayes classifier constructed above and report the confusion matrix for this three-class problem. What is the empirical error rate on the test set?

6. [20 points] The IMOX dataset consists of 192 8-dimensional patterns pertaining to four classes (digital characters ‘I’, ‘M’, ‘O’ and ‘X’). There are 48 patterns per class. The 8 features correspond to the distance of a character to the (a) upper left boundary, (b) lower right boundary, (c) upper right boundary, (d) lower left boundary, (e) middle left boundary, (f) middle right boundary, (g) middle upper boundary, and (h) middle lower boundary. Note that the class labels (1, 2, 3 or 4) are indicated at the end of every pattern.

(a) Write a program to project these 8-dimensional points onto a two dimensional plane using PCA (the top 2 eigenvectors). Report the two projection vectors estimated by the technique. Plot the entire dataset in two dimensions using these projection vectors. Use different markers to distinguish the patterns belonging to different classes.

(b) Write a program to project these 8-dimensional points onto a two dimensional plane using MDA (the top 2 eigenvectors). Report the two projection vectors estimated by the technique. Plot the entire dataset in two dimensions using these projection vectors. Use different markers to distinguish the patterns belonging to different classes. Discuss the differences between the PCA and MDA projection vectors.

**Bonus Points Question**

1. [10 points] Let \( \theta = \text{“Probability of heads of a fair coin”}. Suppose the coin has been tossed \( n \) times and the sequence \( x_1, x_2, \ldots, x_n \) is observed. Here, \( x_i = 0 \) when the outcome is a “tail”, and \( x_i = 1 \) when the outcome is a “head”. Let \( n_H \) be the number of times the coin turned up heads, i.e., \( n_H = \sum_{i=1}^{n} x_i \).

   (a) Show that the maximum likelihood estimate of \( \theta \) is \( \hat{\theta}_{MLE} = \frac{n_H}{n} \).

   (b) Consider the following prior probability for \( \theta \): \( p(\theta) = 2\theta, \; 0 \leq \theta \leq 1 \). Derive the Bayesian estimate of \( \theta \) under the squared-error loss function (Hint: First compute the posterior probability of \( \theta \), and use this to find the expected value of \( \theta \)).

   (Note: \( \int_0^1 \theta^m(1-\theta)^n d\theta = \frac{m!n!}{(m+n+1)!} \))