Homework 3

CSE 802 - Pattern Recognition and Analysis
Instructor: Dr. Arun Ross
Points: 150
Due Date: April 6, 2020

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 soft-copy of this assignment must be uploaded in D2L by April 6, 12:40 pm. 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). Late submissions will not be graded.**
3. **When solving equations or reducing expressions you must explicitly show every step in your computation and/or include the code that was used to perform the computation. Missing steps or code will lead to a deduction of points.**
4. **Code developed as part of this assignment must be included as an appendix to your submission or inline with your solution.**

---

1. [15 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 (Rayleigh distribution):

$$p(x|\theta) = 2\theta x e^{-\theta x^2}, \quad x \geq 0, \theta > 0.$$

   (a) Derive the maximum likelihood estimate (MLE) of θ, i.e., $\widehat{\theta}_{mle}$.

   (b) Consider a set of 1000 training patterns that can be accessed **here**. Plot the normalized histogram of the training patterns. In the same graph, plot the distribution, $p(x)$, after estimating $\widehat{\theta}_{mle}$ from these training patterns.

   (c) Using the same training patterns, determine the MLE estimates for the mean and variance of a Gaussian distribution. In this case, you can use the MLE formulae directly. Plot the resulting Gaussian distribution on the same graph as above.

   (d) Comment on which of the two distributions better “fit" the training data.

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 $D = \{x_1, \dots x_n\}$ are drawn independently according to $p(x|\theta)$. Show that the MLE for θ 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. [15 points] Let $\boldsymbol{x} = (x_1, \dots x_d)^t$ be a d-dimensional binary (0 or 1) vector with a multivariate Bernoulli distribution

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

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

$$\widehat{\boldsymbol{\theta}} = \frac{1}{n}\sum_{k=1}^{n} \boldsymbol{x_k}.$$

(Hint: Consider deriving the MLE for a specific component, θ_i, of vector $\boldsymbol{\theta}$.)

4. [30 points] Consider a two-category (ω_1 and ω_2) classification problem with equal priors. Each feature is a two-dimensional vector $\boldsymbol{x} = (x_1, x_2)^t$. The *true* class-conditional densities are:
$p(\boldsymbol{x}|\omega_1) \sim N(\boldsymbol{\mu_1} = [0,0]^t, \Sigma_1 = I)$,
$p(\boldsymbol{x}|\omega_2) \sim N(\boldsymbol{\mu_2} = [5,5]^t, \Sigma_2 = I)$.

Generate n=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 $\boldsymbol{\mu_1}$, $\boldsymbol{\mu_2}$, Σ_1, and Σ_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 n=500 and n=50,000 random training samples from each of the two densities. How do the estimated parameters and the empirical error rate change in (a) and (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., $p(\boldsymbol{x}|\omega_i) \sim$ N($\boldsymbol{\mu_i}$, $\boldsymbol{\Sigma}_i$), $i = 1,2,3$. Write a program to 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 $\boldsymbol{\mu_i}$ and $\boldsymbol{\Sigma}_i$, $i = 1,2,3$. **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 $\boldsymbol{x}$ and assigns it to one of the three classes based on the maximum posterior rule, i.e., assign $\boldsymbol{x}$ to ω_j if,

$$j = \arg \max_{i=1,2,3} \{P(\omega_i|\boldsymbol{x})\}.$$

(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.

   (c) Discuss the differences between the PCA and MDA projection vectors.

7. [20 points] Assume that the features in the 4-class 8-dimensional **IMOX** dataset described above are statistically independent. Further, assume that each feature for each of the four classes is normally distributed, i.e., $p(x_{ij}|\omega_j) \sim N(\mu_{ij}, \sigma_{ij}^2)$, where $i = 1 \ldots 8$ and $j = 1 \ldots 4$.

   (a) Report the MLE estimates of the mean and variance of each feature for each class, i.e., compute $\widehat{\mu_{ij}}$ and $\widehat{\sigma_{ij}^2}$, for $i = 1 \ldots 8$ and $j = 1 \ldots 4$.

   (b) Assuming a 0-1 loss function and equal priors (and statistically independent features having a Gaussian form), design a Bayesian classifier that inputs an 8-dimensional pattern and assigns it to one of the four classes.

   (c) Train this classifier using the first 24 patterns of each class (so, a total of 96 training patterns). Report the confusion matrix and the empirical error rate of this classifier on the remaining 24 patterns of each class (so, a total of 96 test patterns).

8. [20 points] Consider a dataset in which every pattern is represented by a set of 15 features. The goal is to identify a subset of 5 features or less that gives the best performance on this dataset. How many feature subsets would each of the following feature selection algorithms consider before identifying a solution (i.e., the number of times the criterion function, $J(.)$, will be invoked)?

   (a) SFS;

   (b) Plus-l-take-away-r with $(l, r) = (5, 3)$;

(c) SBS;

(d) Exhaustive Search