Homework 3

CSE 802 - Pattern Recognition and Analysis
Instructor: Dr. Arun Ross
Points: 150
Due Date: March 29, 2021

Note:

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. Copying from *any source* constitutes academic dishonesty.**
2. **A neatly typed report is expected (alternately, you can neatly handwrite the report and then scan it). The report, in PDF format, must be uploaded as a *separate* file in D2L by March 29, 12:40 pm. Late submissions will not be graded. In your submission, 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. **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 (a) included as an appendix to your report or inline with your solution, and (b) archived in a single separate zip file and uploaded in D2L. Including the code without the outputs or including the outputs without the code will result in deduction of points.**

---

1. [10 points] Consider the multivariate normal density with mean $\boldsymbol{\mu} = (\mu_1, \mu_2, \dots \mu_d)^t$, $\sigma_{ij} = 0$, $\forall i \neq j$, and $\sigma_{ii} = {\sigma_i}^2$, that is, the covariance matrix is diagonal: $\boldsymbol{\Sigma} = \mathrm{diag}\big(\sigma_1^2, \sigma_2^2, \dots \sigma_d^2\big)$. Show that,

$$p(x) = \frac{1}{\prod_{i=1}^{d}\left(\sqrt{2\pi}\sigma_i\right)} exp\left[-\frac{1}{2}\sum_{i=1}^{d}\left(\frac{x_i - \mu_i}{\sigma_i}\right)^2\right].$$

2. Consider the following bivariate density function for the random variable $\boldsymbol{x}$:
$p_{original}(\boldsymbol{x}) \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 20 & 10 \\ 10 & 30 \end{bmatrix}\right)$.

   (a) [5 points] Computer and report the whitening transform, $\boldsymbol{A_w}$, of $\boldsymbol{x}$? (You can use matlab (or any other programming language) to compute the eigenvectors/values).

   (b) [3 points] When the whitening transformation is applied to $\boldsymbol{x}$, what is the density function, $p_{transform}$ of the resulting random variable?

   (c) [5 points] Generate 10,000 bivariate *random* patterns from $p_{original}$ (if you are using matlab, then the *mvnrnd* function can be used to generate these patterns). Plot these patterns in a graph.

(d) [5 points] Apply the whitening transform, $\boldsymbol{A}_w$, to the $10,000$ bivariate patterns generated above. Plot the transformed patterns in a separate graph.

(e) [2 points] Compare the patterns in 2c and 2d. **What do you observe?**.

3. [10 points] Consider the following class-conditional densities for a three-class problem involving two-dimensional features:

$$p(\boldsymbol{x}|\omega_1) \sim N\left((-1,-1)^t, I\right);$$
$$p(\boldsymbol{x}|\omega_2) \sim N\left((1,1)^t, I\right);$$
$$p(\boldsymbol{x}|\omega_3) \sim \frac{1}{2}N\left((0.5,0.5)^t, I\right) + \frac{1}{2}N\left((-0.5,-0.5)^t, I\right).$$

(Here, class ω_3 conforms to a **Gaussian Mixture Model (GMM)** with two components - one component is $N\left((0.5,0.5)^t, I\right)$ and the other component is $N\left((-0.5,-0.5)^t, I\right)$ - whose weights are equal (i.e., $\frac{1}{2}$)).

(a) In a 2D graph, mark the mean of ω_1, ω_2, and the two components of ω_3. In the same graph, mark the point $\boldsymbol{x} = (0.1, 0.1)^t$.

(b) Assuming a 0-1 loss function and equal priors, determine the class to which you will assign the two-dimensional point $\boldsymbol{x} = (0.1, 0.1)^t$ based on the Bayes decision rule.

4. [10 points] Suppose we have two normal distributions, corresponding to two classes (ω_1 and ω_2) with the same covariances but different means: $N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$ and $N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$. In terms of their prior probabilities $P(\omega_1)$ and $P(\omega_2)$, state the condition under which the Bayes decision boundary will *not* pass anywhere *between* the two means. You can assume a 0-1 loss function.

5. Consider a two-category classification problem involving two-dimensional feature vectors of the form $\boldsymbol{x} = (x_1, x_2)^t$. The two categories are ω_1 and ω_2, and

$$p(\boldsymbol{x} \mid \omega_1) \sim N\left((4,4)^t, I\right),$$
$$p(\boldsymbol{x} \mid \omega_2) \sim N\left((0,0)^t, I\right),$$
$$P(\omega_1) = P(\omega_2) = \frac{1}{2}.$$

(a) [5 points] Calculate the Bayes decision boundary and write down the Bayes decision rule assuming a 0-1 loss function. The Bayes decision rule has to be written in terms of the Bayes decision boundary.

(b) [10 points] What are the Bhattacharyya and Chernoff theoretical bounds on the probability of misclassification, P(error)?

(c) [5 points] Generate $n = 25$ test patterns from *each* of the two class-conditional densities and plot them in a two-dimensional feature space using different markers for the two categories (if you are using matlab, then the *mvnrnd* function can be used to generate these patterns). Draw the Bayes decision boundary on this plot for visualization purposes.

(d) [5 points] What is the confusion matrix and empirical error rate when classifying the generated patterns using the Bayes decision rule?

(e) [10 points] Can the empirical error rate exceed the theoretical bounds on the probability of misclassification? Why or why not? (Hint: To help answer this question, do the following. Compute the empirical error rate by increasing the value of n from 100 to 1000 in steps of 100. Plot the empirical error rate as a function of n. What do you observe? For each value of n, you may also want to generate the test patterns multiple times (say 3 times), to determine how the empirical error rate changes for the same value of n. *Note that n is the number of test patterns from each class.*)

6. [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.

7. [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?

8. [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., $p(\boldsymbol{x}|\omega_i) \sim \text{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$, $i = 1, 2, 3$. 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?
