1. [20 points] Consider a two-category classification problem involving one feature, $x$. Let both class-conditional densities conform to a Cauchy distribution as follows:

$$p(x | \omega_i) = \frac{1}{\pi b} \cdot \frac{1}{1 + \left( \frac{x - a_i}{b} \right)^2}, \quad i = 1, 2.$$ 

Assume a 0-1 loss function and equal priors.

(a) Compute the Bayes decision boundary.
(b) Show that the probability of misclassification according to the Bayes decision rule is

$$P(error) = \frac{1}{2} - \frac{1}{\pi} \tan^{-1} \left( \frac{a_2 - a_1}{2b} \right).$$

(c) Plot $P(error)$ as a function of $|a_2 - a_1|/b$.
(d) What is the maximum value of $P(error)$. Under what conditions will this occur?

2. [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:

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

(a) Derive the maximum likelihood estimate of $\lambda$, i.e., $\hat{\lambda}_{mle}$.
(b) Assume that $D = \{12, 17, 20, 25, 30\}$. Plot the distribution after computing $\hat{\lambda}_{mle}$ using $D$.

3. [15 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.

4. [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.

5. 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 class-conditional densities are:

$$p(x|\omega_1) \sim N(\mu_1 = [0, 0]^t, \Sigma_1 = I),$$

$$p(x|\omega_2) \sim N(\mu_2 = [3, 3]^t, \Sigma_2 = I).$$

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

(a) [10 points] 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) [10 points] 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) [10 points] 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) [10 points] 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 training samples increases?

6. 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) [10 points] 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) [10 points] 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) [10 points] 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?

**Bonus Points Question**

1. [10 points] In many pattern classification problems, the classifier is allowed to reject an input pattern by not assigning it to any one of the \( c \) classes. If the cost of rejection is not too high, it may be a desirable action in some cases. Let

\[
\lambda(a_i | \omega_j) = \begin{cases} 
0, & i = j \ (i, j = 1, \ldots c) \\
\lambda_r, & i = c + 1 \\
\lambda_s, & i \neq j \ (i, j = 1, \ldots c),
\end{cases}
\]

where \( \lambda_r \) is the loss incurred for rejecting an input pattern and \( \lambda_s \) is the loss incurred for misclassifying the input pattern (known as substitution error).

(a) Show that the minimum risk rule results in the following decision policy.

Assign pattern \( x \) to class \( \omega_i \) if \( P(\omega_i | x) \geq P(\omega_j | x) \forall j \) and \( P(\omega_i | x) \geq 1 - \lambda_r / \lambda_s \), else reject it.

(b) Explain what happens if \( \lambda_r = 0 \)? Also, explain what happens if \( \lambda_r > \lambda_s \)?