Homework 4

CSE 802 - Pattern Recognition and Analysis
Instructor: Dr. Arun Ross
Points: 150
Due Date: April 21, 2021, 11:59pm

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 April 21, 11:59 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. [15 points] Let $\boldsymbol{x} = (x_1, \ldots x_d)^t$ 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, $\boldsymbol{\theta} = (\theta_1, \ldots \theta_d)^t$ is an unknown parameter vector, θ_i being the probability that $x_i = 1$. Let $D = \{\boldsymbol{x_1}, \ldots \boldsymbol{x_n}\}$ 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 $\boldsymbol{\theta}$.)

2. [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.

3. [10 points] Based on the notation developed in class, write down the Sequential Backward Selection (SBS) algorithm and the Sequential Floating Backward Selection (SFBS) algorithm.

4. [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

5. Generate 100 random training points from *each* of the following two distributions: N(20,5) and N(35,5). Write a program that employs the Parzen window technique with a Gaussian kernel to estimate the density, $\widehat{p}(x)$, using *all* 200 points. Note that this density conforms to a *single* **bimodal** distribution.

   (a) [15 points] Plot the estimated density function for each of the following window widths: $h = 0.01, 0.1, 1, 10$. [Note: You can estimate the density at discrete values of x in the [0,55] interval with a step-size of 1.]

   (b) [10 points] Repeat the above after generating 500 training points from each of the two distributions, and then 1,000 training points from each of the two distributions.

   (c) [5 points] Discuss how the estimated density changes as a function of the window width and the number of training points.

6. Consider the dataset available here. It consists of two-dimensional patterns, $\boldsymbol{x} = [x_1, x_2]^t$, pertaining to 3 classes $(\omega_1, \omega_2, \omega_3)$. The feature values are indicated in the first two columns while the class labels are specified in the last column. The priors of all 3 classes are the same and a 0-1 loss function is assumed. Partition this dataset into a training set (the first 250 patterns of each class) and a test set (the remaining 250 patterns of each class).

   (a) [10 points] Let

$$\begin{aligned} p([x_1, x_2]^t|\omega_1) &\sim N([0,0]^t, 4I), \\ p([x_1, x_2]^t|\omega_2) &\sim N([10,0]^t, 4I), \\ p([x_1, x_2]^t|\omega_3) &\sim N([5,5]^t, 5I), \end{aligned}$$

   where I is the 2×2 identity matrix. What is the error rate on the test set when the Bayesian decision rule is employed for classification? Report the confusion matrix as well.

(b) [10 points] Suppose $p([x_1, x_2]^t|\omega_i) \sim N(\boldsymbol{\mu}_i, \Sigma_i)$, $i = 1, 2, 3$, where the $\boldsymbol{\mu}_i$'s and Σ_i's are *unknown*. Use the training set to compute the MLE of the $\boldsymbol{\mu}_i$'s and the Σ_i's. What is the error rate on the test set when the Bayes decision rule using the *estimated parameters* is employed for classification? Report the confusion matrix as well.

(c) [10 points] Suppose the form of the distributions of $p([x_1, x_2]^t|\omega_i)$, $i = 1, 2, 3$ is unknown. Assume that the training dataset can be used to estimate the density at a point using the Parzen window technique (a spherical Gaussian kernel with $h = 1$). What is the error rate on the test set when the Bayes decision rule is employed for classification? Report the confusion matrix as well.

(d) [5 points] Describe your observations based on the error rates and confusion matrices of the 3 classifiers above.

7. [20 points] The iris (flower) dataset consists of 150 4-dimensional patterns belonging to three classes (setosa=1, versicolor=2, and virginica=3). There are 50 patterns per class. The 4 features correspond to (a) sepal length in cm, (b) sepal width in cm, (c) petal length in cm, and (d) petal width in cm. Note that the class labels are indicated at the end of every pattern.

   Design a K-NN classifier for this dataset. Choose the first 25 patterns of each class for training the classifier (i.e., these are the prototypes) and the remaining 25 patterns of each class for testing the classifier. [Note: Any ties in the K-NN classification scheme should be broken at random.]

   (a) In order to study the effect of K on the performance of the classifier, report the confusion matrix for K=1,5,9,13,17,21.

   (b) Plot the classification accuracy as a function of K. Discuss your observations.
