Homework 1

CSE 802: Pattern Recognition and Analysis
Instructor: Dr. Arun Ross
Total Points: 100
Due Date: February 3, 2020, 12:40 PM

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 hard-copy of the homework must be submitted by February 3, 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. **Code developed as part of this assignment must be included as an appendix to your submission or inline with your solution.**

---

1. 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 (x_1), (b) lower right boundary (x_2), (c) upper right boundary (x_3), (d) lower left boundary (x_4), (e) middle left boundary (x_5), (f) middle right boundary (x_6), (g) middle upper boundary (x_7), and (h) middle lower boundary (x_8). Note that the class labels (1, 2, 3 or 4) are indicated at the end of every pattern.

   (a) [4 points] Compute and report the mean pattern vector, i.e., the centroid, of *each class*.

   (b) [4 points] For *each class*, determine the pattern (i.e., vector) from that class which is the farthest from the class mean. You can use the Euclidean distance metric for this problem.

   (c) [8 points] For *each feature*, plot the histograms pertaining to the 4 classes. Your output should contain 8 graphs corresponding to the 8 features; *each graph* should contain 4 histograms corresponding to the 4 classes (choose a bin size of your choice for the histograms). Based on these plots, indicate (a) the *features* that are likely to be useful for distinguishing the 4 classes, and (b) the *classes* that are likely to overlap with each other to a great extent. Provide an *explanation* for your answer.

   (d) [5 points] Assume that each pattern can be represented by features x_1 and x_2. This means, each pattern can be viewed as a point in 2-dimensional space. Draw a scatter plot showing all 192 patterns (use different labels/markers to distinguish between classes). Draw another scatter plot based on features x_3 and x_4. Based on these scatter plots, *explain* which of the two feature *subsets* ((x_1, x_2) or (x_3, x_4)) is likely to be useful for separating the 4 classes.

(e) [4 points] Assume that each pattern can be represented by features (x_1, x_2, x_4). Draw a 3-dimensional scatter plot showing all 192 patterns. Based on this scatter plot, *explain* which classes overlap with each other to a great extent.

2. [10 points] What type of learning scheme - supervised, unsupervised, or reinforcement - can be used to address each of the following problems. You must *justify* your answer.

   (a) Teaching a self-driving rover to navigate the terrain of Mars.

   (b) Given a large set of unknown flowers, discover categories of flowers based on color, geometry, texture and scent of the flowers.

   (c) Determining the identity of a person in a video based on their voice.

   (d) Predicting whether it would rain or not in the next 24 hours based on current weather conditions such as precipitation, humidity, temperature, wind, pressure, etc.

   (e) Given a large set of photos, group all similar looking faces together.

3. [15 points] Describe each of the following terms with an example: (a) overfitting, (b) loss function, (c) decision boundary, (d) segmentation, (e) invariant representation.

4. [20 points] The paper Classification Accuracies of Physical Activities Using Smartphone Motion Sensors by Wu et al. discusses a pattern classification system that determines the physical activity of an individual based on data gleaned from the iPod Touch.

   (a) Briefly describe this system based on the pattern recognition terminology developed in class: (i) sensors used; (ii) pre-processing of raw data; (iii) features extracted; and (iv) classification scheme. How many features (i.e., d) and classes (i.e., c) are present?

   (b) How was classifier training accomplished? How many patterns were available in the training set? How were the training patterns labeled?

   (c) How was the performance of the pattern recognition system evaluated? What metrics were used to evaluate classifier performance?

   (d) In your opinion, did the proposed pattern recognition system perform well? Why or why not?

5. [5 points] Consider the following probability density function which is non-zero only in the range $0 \leq x \leq 10$:

$$p(x) = K.x^3(10 - x).$$

   Here, K is a constant. Determine the value of the constant K.

6. Consider the problem of classifying two-dimensional patterns of the form $\boldsymbol{x} = (x_1, x_2)^t$ into one of two categories, ω_1 or ω_2. Using the labeled patterns presented in this data set[1], do the following.

   (a) [8 points] Plot the histograms (bin size = 1) corresponding to $(x_1|\omega_1)$ and $(x_1|\omega_2)$ in a graph. Also, plot the histograms (bin size = 1) corresponding to $(x_2|\omega_1)$ and $(x_2|\omega_2)$ in a separate graph. Is x_1 more discriminatory than x_2?

[1] The text file has 3 columns. The first two columns correspond to the feature vector of a pattern and the third column corresponds to its class label.

(b) [7 points] Plot the two-dimensional patterns in a graph. Use markers to distinguish the patterns according to their class labels. Suppose you have the following decision rule (classifier) to classify a novel pattern $\boldsymbol{x} = (x_1, x_2)^t$:

If $x_1 + x_2 - 15 < 0$, $\boldsymbol{x} \in \omega_1$ else $\boldsymbol{x} \in \omega_2$.

In the same graph, plot the decision boundary corresponding to this rule. What is the error rate (i.e., the percentage of patterns that are misclassified) when this decision rule is used to classify the patterns in the given data set?

(c) [7 points] Repeat the above after modifying the decision rule (classifier) as follows:

If $x_1 + x_2 - 12 < 0$, $\boldsymbol{x} \in \omega_1$ else $\boldsymbol{x} \in \omega_2$.

(d) [3 points] Which of the two classifiers has performed well on this dataset.
