

Homework 3

CS 591Q/791V - Pattern Recognition

Instructor: Dr. Arun Ross

Due Date: April 7, 2009

Note: 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.** Code developed as part of this assignment must be placed in a zip file and sent to arun.ross at mail.wvu.edu with the subject line “CS 591Q/791V : Homework 3”. Also, include a hard-copy of the code when you submit the homework.

1. 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, $\hat{p}(x)$, using *all* 200 points.
 - (a) [15 points] Plot the estimated density function for the following window widths: $h = 0.01, 0.5, 10$. [Note: You can estimate the density at discrete values of x in the $[0,55]$ interval with a step-size of 1.]
 - (b) [5 points] Repeat the above after generating 1000 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. What is the most critical parameter in the Parzen window density estimation technique? Justify your answer.
2. 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.

Assume that each class can be modeled by a multivariate Gaussian density with unknown mean and covariance, i.e., $p(\mathbf{x}|C_i) \sim N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$, $i = 1, 2, 3, 4$. Design a Bayes classifier and test it as follows:

- (a) [10 points] Train the classifier: Using the first 24 patterns of each class (training data), compute the maximum likelihood estimate of the model parameters. Report these estimates, i.e., $\hat{\boldsymbol{\mu}}_i$ and $\hat{\boldsymbol{\Sigma}}_i$, $i = 1, 2, 3, 4$.
- (b) [10 points] Design the classifier: Assuming that the four classes are equally probable, write a program that inputs an 8-dimensional pattern \boldsymbol{x} and assigns it to one of the four classes based on the maximum posterior rule, i.e., assign \boldsymbol{x} to C_j if,

$$j = \arg \max_{i=1,2,3,4} \{P(C_i|\boldsymbol{x})\}.$$

- (c) [10 points] Test the classifier: Classify the remaining 24 patterns of each class (test data) using the Bayes classifier constructed above and report the confusion matrix for this four-class problem.

3. 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) [15 points] 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) [5 points] Plot the classification accuracy as a function of K .
- (c) [5 points] Discuss your observations.
-