Homework 4
CSE 802, Spring 2017
Due: March 15, 2017

1. Solve problems 6(a), 6(b), 8, and 9(a) from Chapter 4 in the DHS book.

   Please note that the problem numbers provided in the assignment questions correspond to the printed second edition of the text book. It is your responsibility to check the version you have, and make sure you are doing the right problems. Electronic edition of the book may have different problem numbering from the printed edition.

2. The iris data set, one of the most well known data sets, has been used in pattern recognition literature to evaluate the performance of various classification and clustering algorithms. This data set consists of 150 4-dimensional patterns belonging to three types of iris flowers (setosa, versicolor, and virginica). There are 50 patterns per class. The 4 features correspond to: sepal length, sepal width, petal length and petal width (the unit of measurement is cm).

   The data can be accessed at `http://www.cse.msu.edu/~cse802/HW/HW4/iris.txt`. The class labels are indicated at the end of every pattern.

   (a) Find the non-parametric density estimates of the three classes using Parzen window density estimation method. Assume a 4-dimensional Gaussian kernel and try window widths of 0.01, 0.5, and 10. For the Gaussian kernel, assume that the covariance matrix is diagonal. The diagonal entries are the sample variance. Perform 5-fold cross validation and report the average and the variance of the error rate. Each of the three classes should be divided into five folds to have equal representation in training and testing.

   (b) Now assume that each pattern class has a multivariate Gaussian density with unknown mean vector and unknown but different covariance matrix. Design a plug-in classifier for this problem. Again, use a 5-fold cross validation to design and test the classifier. Report the average and the variance of the error rate of the quadratic classifier.

   (c) Compare the performance in (a) and (b).

3. *Bonus Question (25 points)*

   Consider a set of two-dimensional points pertaining to two classes that are linearly separable. The data can be accessed at `http://www.cse.msu.edu/~cse802/HW/HW4/two_class.txt`. Note that the class labels (+1 or -1) are indicated at the end of every pattern.

   Write a program to compute the linear decision boundary for this two-class problem by implementing the fixed-increment single-sample perceptron learning algorithm. Report the

linear decision boundary computed by the learning algorithm and the number of iterations taken by the algorithm to converge to this boundary when the weight vector is initialized as follows: (i) $(-1, -1, -1)^T$, (ii) $(-30, 1, 1)^T$, and (iii) $(10, 10, 10)^T$. In each case, plot (i) the two dimensional-points, (ii) the initial decision boundary and (iii) the final decision boundary.