Homework 4

CS 591Q/791V - Pattern Recognition
Instructor: Dr. Arun Ross
Due Date: April 29, 2010

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 should be placed in a zip file and sent to arun.ross at mail.wvu.edu with the subject line "CS 591Q/791V : Homework 4". Also, include a hard-copy of your 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.

2. 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.

3. [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 7 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?

   (a) Exhaustive search;

   (b) SFS;

   (c) SBS;

   (d) Plus-l-take-away-r with $(l, r) = (5, 3)$.

Note: CS 791V students have to answer the following questions in addition to the ones above.

1. [10 points] Based on the notation developed in class, write down the Sequential Floating Backward Selection (SFBS) algorithm.

2. [10 points] Consider a 2-category classification problem involving a single feature x. Assume equal class priors and a 0-1 loss function. The class-conditional densities are as follows:

$$\begin{aligned} p(x|\omega_1) &= \begin{cases} 2x & \text{for } 0 \leq x \leq 1 \\ 0 & \text{otherwise,} \end{cases} \\ p(x|\omega_2) &= \begin{cases} 2 - 2x & \text{for } 0 \leq x \leq 1 \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

   Suppose we randomly select a single point from ω_1 and a single point from ω_2, and create a 1-nearest-neighbor classifier. Suppose too we select a test point from one of the categories (ω_1 for definiteness). Integrate to find the expected error rate $P_1(e)$.
