CSE 569 Homework #3

Total 3 points

Due Thursday, Oct. 29 **by 11:59pm**.

Problem 1. Consider histogram-based density estimation for some PDF $p(x)$ defined on the interval [0, 1]. Suppose that we are given a training set D of n samples drawn from $p(x)$, $D = \{x_1, x_2, \ldots, x_n\}$. Further suppose we use the following m equal-length bins for computing the histogram: B_1=[0, 1/m), B_2=[1/m, 2/m), …, B_m=[(m-1)/m, 1]. With this, we may count the number of samples falling into each bin, and we denote that number by Y_j for the j-th bin.

(a) Write down the histogram-based density estimate $\hat{p}(x)$, which should be a function of x and those quantities given above. Note: you need to write down a close-formed estimate so that you may evaluate its value for any x.

(b) For a given x, find the expectation of your estimate $\hat{p}(x)$, i.e., $E[\hat{p}(x)]$.

Problem 2. **(**From Problem 3 of Chapter 4 in the textbook**)**

Let $p(x) \sim U(0, a)$ be uniform from 0 to a, and let a Parzen window be defined as $\varphi(x) = e^{-x}$ for $x > 0$ and 0 for $x \leq 0$.

(a) Show that the mean of such a Parzen-window estimate is given by

$$\bar{p}_n(x) = \begin{cases} 0 & x < 0 \\ \frac{1}{a}(1 - e^{-x/h_n}) & 0 \leq x \leq a \\ \frac{1}{a}(e^{a/h_n} - 1)e^{-x/h_n} & a \leq x. \end{cases}$$

(b) Plot $\bar{p}_n(x)$ versus x for $a = 1$ and $h_n = 1, 1/4$, and $1/16$.

(c) How small does h_n have to be to have less than one percent bias over 99 percent of the range $0 < x < a$?

(d) Find h_n for this condition if $a = 1$, and plot $\bar{p}_n(x)$ in the range $0 \leq x \leq 0.05$.

Problem 3. Consider a 1-dimensional 2-class classification problem with class-conditionals as follows:

$$p(x|\omega_1) = \begin{cases} 2x, & x \in [0,\ 1] \\ 0, & otherwise \end{cases} \qquad p(x|\omega_2) = \begin{cases} 2(1-x), & x \in [0,\ 1] \\ 0, & otherwise \end{cases}$$

Assume that the priors are equal, i.e., $P(\omega_1) = P(\omega_2) = 0.5$.

(a) What is the Bayesian decision boundary for doing minimum error rate classification? What is the corresponding Bayes error?

(b) You are given two training samples: x_1 from ω_1 and x_2 from ω_2. Also, we know $x_1 < x_2$. Find the nearest-neighbor (NN) decision rule for classifying any new data point x. What is the probability of error for this NN classifier? (Note: for this Part (b), you are given a fixed training set, i.e., you view x_1 and x_2 as some given, fixed values.)

(c) More generally, suppose we randomly select a single point x_1 from ω_1 and a single point x_2 from ω_2, and create a NN classifier. Consider using this NN classifier to classify a random sample drawn from ω_1. What is the probability of error?

Problem 4. Prove that the Voronoi cells induced by the nearest-neighbor algorithm must always be convex. That is, for any two points $\mathbf{x}_1$ and $\mathbf{x}_2$ in a cell, all points on the line linking $\mathbf{x}_1$ and $\mathbf{x}_2$ must also lie in the cell.

Problem 5. **Computer Exercise.** [Review 06-Feature-Selection-Intro.pdf before working on this question. This question is essentially a simulation of the example given in Slides 11. You will need to write some code for doing part of this exercise. Include your code in the submission.]

Do the following exercise three times, with n=20 the first time, n=100 the second time, and n=600 the third time. (And feel free to try with other numbers too.)

(1) Generate the first set D_1 of n samples from the normal distribution $N(1, 1)$. Generate the second set D_2 of n samples from the normal distribution $N(1.5, 1)$.

(2) Assume D_1 is a set of i.i.d. samples of certain feature for Class 1 in a two-class problem, and accordingly D_2 is a set of i.i.d. samples of the feature for Class 2. Test if we should accept the hypothesis *that the means of this feature for the two classes are the same* for a given significant level 0.05.