Homework 5

Stat 435, Spring 2020
Due Friday, June 5, 11:59pm

Problem 1: (70 points) Consider a classification problem with a binary class label Y and a single continuous feature X that takes values in $(-4,-2) \cup (2,4)$. Suppose (X,Y) is generated by choosing Y at random with $P(Y = 1) = P(Y = 2) = 1/2$, and then drawing X conditional on Y according to uniform distributions. Specifically, assume that the class-conditional densities for X are

$$\begin{aligned} p(x \,|\, Y = 1) &= \frac{1}{2} \cdot \mathbf{1}_{(-4,-2)}(x) \quad \text{and} \\ p(x \,|\, Y = 2) &= \frac{1}{2} \cdot \mathbf{1}_{(2,4)}(x)\,. \end{aligned}$$

In the below we consider 0-1 loss, that is, the risk of a classifier is the probability of an error.

(a) (10 points) What is the marginal distribution of X? What is the conditional distribution of Y given X?

(b) (10 points) What is the Bayes rule $f_B(x)$ and its risk $P(Y \neq f_B(x)$? Explain!

(c) (20 points) Let $\hat{f}_1(x; S)$ be the 1-nearest neighbor classifier based on a training sample $S = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ of i.i.d observations of (X,Y). What is the risk $\Pr(Y \neq \hat{f}_1(X; S)$? Explain. (Here, the risk is computed by integrating over training data and a new independent pair (X,Y)).

(d) (20 points) Under the same scenario calculate the risk of the 3-nearest neighbor classifier.

(e) (10 points) Which method, 1-nearest neighbor or 3-nearest neighbor, has smaller risk in this problem?

2. ISLR Section 8.4 Problem 3 (20 points)

2. ISLR Section 8.4 Problem 9 (a) ... (g) (10 points each)