STOR 565 Homework

1. Consider a classification problem in which the predictor X is uniformly distributed on the unit interval $[0,1]$ and the response $Y \in \{0,1\}$ as usual. For $x \in [0,1]$ let $\eta(x) = \mathbb{P}(Y = 1 \,|\, X = x)$. Specify the Bayes rule ϕ^* and the Bayes risk R^* in each of the following cases.

a. $\eta(x) = 1/3$ for all x

b. $\eta(x) = x$

c. $\eta(x) \in \{0,1\}$ for all x

In each of the cases above, find the prior probability $\pi_1 = \mathbb{P}(Y = 1)$, or indicate why this is not possible without more information.

2. Let $(X,Y) \in \mathbb{R}^2 \times \{0,1\}$ be a random predictor-response pair. Suppose that the predictor X is a pair (X_1, X_2) where $X_1, X_2 \in [0,1]$ are independent, X_1 is uniform on $[0,1]$, and X_2 has density $g(x_2) = 3x_2^2$ for $0 \leq x_2 \leq 1$. Suppose that $\eta(x_1, x_2) = (x_1 + x_2)/2$.

a. Find the Bayes rule ϕ^* for this problem and identify its decision boundary.

b. Find the unconditional density of X

c. Find the Bayes risk associated with (X,Y)

d. Find the prior probability that $Y = +1$.

e. Find the class-conditional density of X given $Y = 1$.

3. Show that for each number $u \in \mathbb{R}$ we have

$$\min(u, 1-u) \;=\; u\,\mathbb{I}(u < 1/2) + (1-u)\,\mathbb{I}(u \geq 1/2)$$

Hint: Consider separately the cases $u < 1/2$ and $u \geq 1/2$.

4. Consider the labeled data set $(-2,1), (-1,1), (0,0), (1,1), (2,0) \in \mathbb{R} \times \{0,1\}$.

a. Sketch the 1-nearest neighbor rule for this dataset by drawing a line and indicating which points are assigned to zero and which are assigned to one.

b. Sketch the 3-nearest neighbor rule for this dataset by drawing a line and indicating which points are assigned to zero and which are assigned to one.

5. Let X be a discrete random variable taking values in a finite (or countably infinite) set $\mathcal{X}$, and having probability mass function $p(x) = \mathbb{P}(X = x)$. Let $h : \mathcal{X} \to [a, b]$ be any function.

a. Write down the sum for $\mathbb{E}h(X)$.

b. Show that $\mathbb{E}h(X) = a$ if $p(x) > 0$ only when $h(x) = a$.

c. Establish the reverse implication: if $\mathbb{E}h(X) = a$ then $p(x) > 0$ only when $h(x) = a$. Hint: Assume to the contrary that $p(x') > 0$ for some $x' \in \mathcal{X}$ with $h(x') \neq a$. As h takes values in $[a, b]$, we have $h(x') > a$. Use this to show $\mathbb{E}h(X) > a$.

d. Following the arguments above, show that $\mathbb{E}h(X) = b$ if and only if $p(x) > 0$ implies $h(x) = b$.

6. Let $(X, Y) \in \mathcal{X} \times \mathcal{Y}$ be a jointly distributed pair. Assume that $\mathcal{X}$ and $\mathcal{Y}$ are finite. Recall that X and Y are independent if $\mathbb{P}(X = x, Y = y) = \mathbb{P}(X = x)\,\mathbb{P}(Y = y)$.

a. Show that if X and Y are independent then $\mathbb{P}(X = x \,|\, Y = y)$ does not depend on y.

b. Let $y \in \mathcal{Y}$ be fixed. Show that if $\mathbb{P}(Y = y \,|\, X = x)$ does not depend on x then it is equal to $\mathbb{P}(Y = y)$.

c. Suppose that for each $y \in \mathcal{Y}$ the conditional probability $\mathbb{P}(Y = y \,|\, X = x)$ does not depend on x. Show that X and Y are independent.

7. Suppose that you are given access to a database consisting of many email messages that have been labeled as spam or normal. You decide to construct a simple classification rule, the only feature being whether or not the word “meeting” appears somewhere in the email. Using relative frequencies to estimate probabilities you find the following:

$$\hat{P}(\text{spam}) = .3 \quad \hat{P}(\text{‘meeting’ present} \,|\, \text{spam}) = .01 \quad \hat{P}(\text{‘meeting’ present} \,|\, \text{normal}) = .04$$

Using this information, calculate a simple classification rule for spam detection. What can you say about the error rate of your rule on the database?

8. Argue as carefully as you can that if the Bayes risk R^* for a pair (X, Y) is equal to $1/2$ then Y is independent of X. Hint: Use the results of problems 5 and 6 above.