

# 36-707: Regression Analysis Homework 6 Solutions

Fall 2007

## Problem 1

a.  $h^*(x) = I_{[0,.1]}(x) \times I_{[.9,1]}$  OR  $h^*(x) = I_{[0,.1]}(x) \times I_{[.9,10]}$  depending which problem you chose to do.

b. If  $x \sim Unif(0, 10)$ ,  $R^* = 0.108$ . If  $x \sim Unif(0, 1)$ ,  $R^* = 0.18$ .

c. If  $x \sim Unif(0, 10)$ ,  $R^* = 0.116$ , which corresponds to  $x = 0.9$ . If  $x \sim Unif(0, 1)$ ,  $R^* = 0.26$ , which corresponds to  $x = 0.9$ .

## Problem 2

We assume that  $\pi = P(Y = 1) = 1/2$ .

$$h^*(x) = \begin{cases} 1 & \text{if } \frac{f_1(x)}{f_0(x)} > 1 \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

$$h^*(x) = \begin{cases} 1 & \text{if } f_1(x) > f_0(x) \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

$$\begin{aligned}
R^* &= P(h^*(X) \neq Y) \\
&= \int P(h^*(X) \neq Y | X = x) f(x) dx \\
&= \int (P(f_1(x) > f_0(x), Y = 0 | X = x) + P(f_1(x) < f_0(x), Y = 1 | X = x)) f(x) dx \\
&= \int (I(f_1(x) > f_0(x)) P(Y = 0 | X = x) + I(f_1(x) < f_0(x)) P(Y = 1 | X = x)) f(x) dx \\
&= \int (I(f_1(x) > f_0(x)) \frac{f(x|Y = 0) P(Y = 0)}{f(x)} \\
&\quad + I(f_1(x) < f_0(x)) \frac{f(x|Y = 1) P(Y = 1)}{f(x)}) f(x) dx \\
&= \int (I(f_1(x) > f_0(x)) 1/2 f_0(x) + I(f_1(x) < f_0(x)) 1/2 f_1(x)) dx \\
&= 1/2 \int (f_0(x) I(f_1(x) > f_0(x)) + f_1(x) (1 - I(f_1(x) > f_0(x)))) dx \\
&= 1/2 \int (f_0(x) I(f_1(x) > f_0(x)) + f_1(x) - f_1(x) I(f_1(x) > f_0(x))) dx \\
&= 1/2 \int f_1(x) dx - 1/2 \int (f_1(x) - f_0(x)) I(f_1(x) > f_0(x)) dx \\
&= 1/2 - 1/2 \int_{f_1(x) > f_0(x)} (f_1(x) - f_0(x)) dx
\end{aligned}$$

We also show that:

$$\begin{aligned}
\int |f_1(x) - f_0(x)|dx &= \int_{f_1(x) > f_0(x)} (f_1(x) - f_0(x))dx + \int_{f_1(x) < f_0(x)} (f_0(x) - f_1(x))dx \\
&= \int_{f_1(x) > f_0(x)} f_1(x)dx - \int_{f_1(x) > f_0(x)} f_0(x)dx + \int_{f_1(x) < f_0(x)} f_0(x)dx \\
&\quad - \int_{f_1(x) < f_0(x)} f_1(x)dx + \int_{f_1(x) > f_0(x)} f_1(x)dx - \int_{f_1(x) > f_0(x)} f_1(x)dx \\
&\quad + \int_{f_1(x) > f_0(x)} f_0(x)dx - \int_{f_1(x) > f_0(x)} f_0(x)dx \\
&= 2 \int_{f_1(x) > f_0(x)} f_1(x)dx - 2 \int_{f_1(x) > f_0(x)} f_1(x)dx \\
&\quad + \int f_0(x)dx - \int f_1(x)dx \\
&= 2 \int_{f_1(x) > f_0(x)} (f_1(x) - f_0(x))dx + 1 - 1 \\
&= 2 \int_{f_1(x) > f_0(x)} (f_1(x) - f_0(x))dx
\end{aligned}$$

Therefore, we can substitute this into our equation for  $R^*$  to obtain the result:

$$R^* = 1/2 - 1/4 \int |f_1(x) - f_0(x)|dx \quad (3)$$

The interpretation of this equation is that the risk of any classifier where  $\pi = 0.5$  will be  $1/2$  minus a quarter of the integral of the absolute difference between the densities of  $x|Y=1$  and  $x|Y=0$ . Thus, if we have identical distributions for  $x|Y=1$  and  $x|Y=0$ , we have the maximum risk possible of  $1/2$ . As these two densities get more and more dissimilar, our risk decreases by the amount of difference between the densities. The risk reaches a minimum of  $0$  when the two densities are completely disjoint.

### Problem 3

- a. The misclassification rates should have all been between approximately 4-6% assuming you did not do cross validation. If you did, they were slightly higher.
- b. The misclassification rates were much higher - between 30-50%. There were

a few different ways to 'be creative' with the regressions so as long as you tried something, I gave full credit.

#### Problem 4

The misclassification rates for SVM and LDA are 6.25% and 18.6% respectively for the training set. They were 17.7% and 23.8% for the test set indicating that SVM worked better in this case.