CPH 636 — Spring 2009 — Dr. Charnigo

Written Assignment 5

Written Assignment 5 is due on Monday 13 April at the end of lecture. You are encouraged to work in groups of two or three, though you may work individually if you prefer.

The first exercise is a theoretical problem whose solution is anticipated to be rewarding to a graduate student in statistics.

[10] 1. Before performing nearest neighbors analysis, we standardize the continuous explanatory variables. However, the ideal situation is for the continuous explanatory variables to be not only standardized but also uncorrelated. Propose a mechanism to convert standardized but correlated explanatory variables to standardized and uncorrelated explanatory variables. (You do not have to implement this mechanism in your solution to exercise 3.)

In the second exercise, you will apply a neural network to the regression problem with response variable HeartRate and explanatory variables Age, Female, Height, Weight, QRSDur, PRInt, QTInt, TInt, PInt, and ArrhythmiaAny from the rhythm data set.

[50] 2. Please perform the following tasks.

[10] a. Write out the formula for the predicted HeartRate in terms of Age, Female, Height, Weight, QRSDur, PRInt, QTInt, TInt, PInt, and ArrhythmiaAny.

[10] b. What is the average squared error on the test data set? Define $R^2_{test,NeuralNetwork}$ to be

$$1 - \frac{n_{test} \times \text{ average squared error on the test data set}}{(n_{test} - 1) \times \text{ sample variance of HeartRate on the test data set}}.$$

Calculate $R^2_{test,NeuralNetwork}$ and compare it to $R^2_{test,LinearRegression}$ from your chosen linear regression model in Written Assignment 3. Which is better at making predictions on the test data set, your linear regression model or your neural network?

[10] c. Based on your neural network, what are the predicted values of HeartRate for the observations with ID numbers 5 and 317? How far off are the predicted values from the actual values?

[10] d. The discrepancies between the predicted values and the actual values in item c may have little meaning to a person unfamiliar with the rhythm data set. With this in mind, express the discrepancies as fractions of the standard deviation of HeartRate within the test data set. Do the discrepancies seem small or large?

[10] e. Perhaps your neural network could be improved through the removal of some explanatory variables. Describe a procedure for deciding which explanatory variables to remove. (You do not need to carry out your procedure, just describe it.)

In the third exercise, you will apply a nearest neighbors analysis to the regression problem with response variable HeartRate and explanatory variables Age, Female, Height, Weight, QRSDur, PRInt, QTInt, TInt, PInt, and ArrhythmiaAny from the rhythm data set. (For this exercise, you may treat the dichotomous explanatory variables as if they were continuous!)

[40] 3. Please perform the following tasks.

[10] a. Try 16 neighbors, 8 neighbors, and 32 neighbors. Which gives you the best performance on the validation data set? Retain this number of neighbors for items b through d.

[10] b. What is the average squared error on the test data set? Define $R^2_{test,NearestNeighbor}$ to be

$$1 - \frac{n_{test} \times \text{ average squared error on the test data set}}{(n_{test} - 1) \times \text{ sample variance of HeartRate on the test data set}}.$$

Calculate $R^2_{test,NearestNeighbor}$ and compare it to $R^2_{test,LinearRegression}$ and $R^2_{test,NeuralNetwork}$. Which is best at making predictions on the test data set, your linear regression model, your neural network, or your nearest neighbors analysis?

[10] c. Based on your nearest neighbors analysis, what are the predicted values of HeartRate for the observations with ID numbers 5 and 317? How far off are the predicted values from the actual values?

[10] d. Express the discrepancies as fractions of the standard deviation of HeartRate within the test data set. Which fares better at prediction for these two specific observations, your neural network or your nearest neighbors analysis?