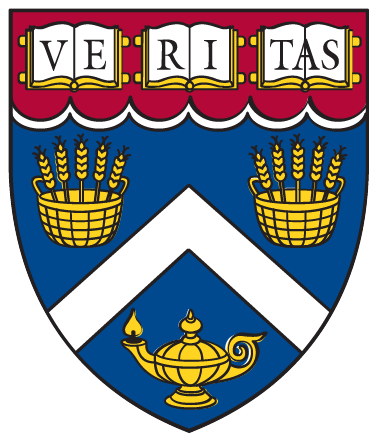
CSCI E-106: Data Modeling



Fall 2019

Dr.Hakan Gogtas

Assignment 10

Due: December 12, 2019 at 11:59pm EST

**Instructions:** Students should submit their reports on Canvas. The report needs to clearly state what question is being solved, step-by-step walk-through solutions, and final answers clearly indicated. Please solve by hand where appropriate.

Please submit two files: (1) a R Markdown file (.Rmd extension) and (2) a PDF document generated using knitr for the .Rmd file submitted in (1) where appropriate. Please, use RStudio Cloud for your solutions.

1. Refer to the Prostate Cancer data set in Appendix C.5 and Homework 9. Select a random sample of 65 observations to use as the model-building data set (use set.seed(1023)). Use the remaining observations for the test data. (10 pts)
2. Develop a neural network model for predicting PSA. Justify your choice of number of hidden nodes and interpret your model. Test the model performance on the test data.
3. Compare the performance of your neuron network model with regression tree model obtained in HW9. Which model is more easily interpreted and why? (5pts)
4. Compare the performance of your neural network model with that of the best regression model obtained in homework 8. Which model is more easily interpreted and why?
5. Refer to the Disease outbreak data set in Appendix C.10. Savings account status is the response variable and age, socioeconomic status, and city sector are the predictor variables.
6. Fit logistic regression model to predict the saving account status on the predictor variables in first-order terms and interaction terms for. all pairs of predictor variables. State the fitted response function.
7. Use the likelihood ratio test to determine whether all interaction terms can be dropped from the regression model; use α = .01. State the alternatives, full and reduced models, decision rule, and conclusion. What is the approximate P-value of the test?
8. Conduct the Hosmer-Lemeshow goodness of fit test for the appropriateness of the logistic regression function by forming five groups of approximately 20 cases each; use α = .05.
9. Refer to the Geriatric study. A researcher in geriatrics designed a prospective study to investigate the effects of two interventions on the frequency of falls. One hundred subjects were randomly assigned to one of the two interventions: education only (X1 = 0) and education plus aerobic exercise training (X1 = 1). Subjects were at least 65 years of age and in reasonably good health. Three variables considered to be important as control variables were gender (X2:0=female;1=male), a balance index (X3). and a strength index (X4). The higher balance index, the more stable is the subject and the higher the strength index, the stronger is the subject. Each subject kept a diary recording the number of falls (Y) during the six months of the study.
10. Fit the regression model. State the estimated regression coefficients, their estimated standard deviations. and the estimated response function.
11. Assuming that the fitted model is appropriate, use the likelihood ratio test to determine whether gender (X2) can be dropped from the model: State the full and reduced models. decision rule. and conclusion. What is the P-value of the test
12. Predicted the number of falls for X1=1, X2=0, X3=45, X4=70.

