MSiA 420, HW #2
due Wednesday 2/8/16, 10:00 am

As with all HW (unless otherwise noted), upload your solutions for this assignment on Canvas, as a Word or pdf file, by the due date/time. For all problems for which you use R, include your R script in an appendix to your homework (clearly label which parts of the script correspond to which homework problems).

1) Consider the Ischemic heart disease data described in Appendix C.9 of KNN. The data are in HW2_data.xls. The following is a description of the data, from KNN: A health insurance company collected information on 788 of its subscribers who had made claims resulting from ischemic (coronary) heart disease. Data were obtained on total costs of services provided for these 788 subscribers and the nature of the various services for the period of January 1, 1998 through December 31, 1999. Each line in the data set has an identification number and provides information on 9 other variables for each subscriber. The 10 variables are:

    1. Identification number
    2. Total cost of claims by subscriber (dollars)
    3. Age of subscriber (years)
    4. Gender (1 = male; 0 = female)
    5. Total number of Interventions or procedures carried out
    6. Number of tracked drugs prescribed
    7. Number of emergency room visits
    8. Number of complications that arose during heart disease treatment
    9. Number of other diseases (comorbidities) that the subscriber had during the period
    10. Number of days of duration of treatment condition

    For this and subsequent problems, let the response variable be the log (base 10) of the total cost. There were originally a handful of zeros in the "cost" column, which have been replaced by "1" in the Excel file (otherwise, you can't take the log). Notice that some of the predictors are very heavy tailed, so it may be better to take the log of these predictors up front (although the zeros in the discrete predictors would create problems if you took the log). But for the sake of consistency, do NOT take the log of the predictors for these problems.

    (a) Fit a linear regression model to the ischemic heart disease data. Using any and all arguments that are relevant, discuss how well the model fits the data in terms of its predictive power.
    (b) Which variables appear to have the most influence on the cost?
    (c) Construct appropriate diagnostics and residual plots to assess whether you think there are any problems with the data set that require remedial action or any nonlinearity in the relationship between the response and the predictors.

2) For this problem, your objective is to find the best neural network model for the ischemic heart disease data. For consistency, use a linear output activation function, and do NOT rescale the response (log of cost) to the [0,1] interval.

   (a) Use 10-fold cross-validation to find the best combination of shrinkage parameter λ and number of hidden nodes.
   (b) Fit the final best model and discuss how good you think the model is, in terms of its predictive power.
   (c) Which variables appear to have the most influence on the cost, and what are their effects? You can use the ALEPlot package for this.
   (d) Construct appropriate residual plots to assess whether there remains any nonlinearity not captured by the neural network model.

3) Repeat Problem 2, but for a regression tree. Specifically:

   (a) Use 10-fold cross-validation to find the best tree size or complexity parameter value.
   (b) Fit the final best model and discuss how good you think the model is, in terms of its predictive power.
   (c) Which variables appear to have the most influence on the cost, and what are their effects?
   (d) Construct appropriate residual plots to assess whether there remains any linearity not captured by the regression tree model.
   (e) Which model (the linear regression, neural network, or tree) would you recommend for this data set, and why?

4) Reconsider the forensic glass data used in lecture. These data are in HW2_data.xls and also posted on Blackboard as a text file (fgl.txt). These are data on fragments of glass in a forensic study. There are n = 214 rows with 10 variables. "type" is a factor with six levels that represents the type of glass, and the other 9 variables are predictor variables that represent the chemical composition of the glass and the refractive index (RI). The levels of type are window float glass (WinF: 70), window non-float glass (WinNF: 76), vehicle window glass (Veh: 17), containers (Con: 13), tableware (Tabl: 9) and vehicle headlamps (Head: 29).This is a classification problem, in which the objective is to classify the categorical response "type" into one of six different glass types. Type library(MASS) and ?fgl to see additional description of the data (which are in the MASS package). In lecture, we converted the 6-category response into a binary response (window glass or other). In this problem, you will retain the 6-category response and attempt to classify the glass type into one of the six categories.

   (a) Use 10-fold cross-validation to find the best neural network model for classifying the class type.
   (b) Use 10-fold cross-validation to find the best classification tree model for classifying the class type.
   (c) An alternative to a neural network or classification tree is to use nominal logistic regression, which is like the binary logistic regression with which you are familiar

from IEMS 304, except that it applies when you have more than two response categories. Sometimes this is also referred to as “multinomial” regression or “polytomous logistic regression”. See Section 14.11 of KNN and/or Section 4.4 of HTF for descriptions of the approach. In R, you can use the multinom() function (which is part of the nnet package) to fit a multinomial model. Fit a multinomial model and discuss the results.

(d) Compare the three models from parts (a)—(c) in terms of their predictive ability and interpretability. Which model do you think is the most appropriate for predicting glass type?