MSiA 420, HW #3
See Canvas for Due Date

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) Reconsider the ischemic heart disease data that you analyzed in HW2. As in HW2, for this and subsequent problems, let the response variable be the log (base 10) of the total cost, and do NOT take the log of the predictors. This problem involves using nearest neighbors to predict cost.

   (a) Use *n*-fold CV to find the best *K* for predicting cost using K-NN. What are the pros and cons of using *n*-fold CV, versus say 10-fold CV, for nearest neighbors?
   (b) For the optimal *K* from part (a), what is the CV estimate of the prediction error standard deviation?
   (c) What is the predicted cost for a person with age=59, gend=0, intvn=10, drugs=0, ervis=3, comp=0, comorb=4, and dur=300?

2) This problem involves fitting a generalized additive model to predict cost for the ischemic heart disease data, for which you can use the gam() function of package:gmcv.

   (a) Fit a GAM model without interactions, and construct plots of the component functions. Which predictors appear to be the most relevant for predicting cost?
   (b) For the model from part (a), what is the CV estimate of the prediction error standard deviation? What are the pros and cons of using *n*-fold CV, versus say 10-fold CV, for GAMs?
   (c) What is the predicted cost for a person with age=59, gend=0, intvn=10, drugs=0, ervis=3, comp=0, comorb=4, and dur=300?

3) This problem involves using kernel methods to predict cost for the ischemic heart disease data. Since the loess() function in R allows at most four predictors, and the ischemic heart disease data involves eight predictors, you will have to choose a good subset of four predictors to work with. From HW2, the best regression tree had the three predictors intvn, comorb, and dur, so these three should probably be in the model. These three also had the smallest P-values from the linear regression fit from HW2 and from the GAM fit in problem 2 above, and the VIFs were all small, so the other larger P-values are not distorted by multicollinearity. The predictor comp had the next smallest P-value in HW2 and in the GAM fit above, but it is discrete, which will cause an error in loess(). So as the fourth predictor for this problem, use ervis, which appears to be the next most relevant predictor from the GAM fit above.

   (a) Use CV to find the best combination of span and degree (0 for local average, 1 for local linear, and 2 for local quadratic regression) for a kernel method.

(b) Use C_p to find the best combination of span and degree (0 for local average, 1 for local linear, and 2 for local quadratic regression) for a kernel method. Is this in agreement with what CV said was the best span and degree?
(c) For the optimal model from part (a), what is the CV estimate of the prediction error standard deviation?
(d) What is the predicted cost for a person with age=59, gend=0, intvn=10, drugs=0, ervis=3, comp=0, comorb=4, and dur=300?

4) This problem involves using projection pursuit regression to predict cost for the ischemic heart disease data, for which you can use the ppr() function.

(a) Use CV to find the best number of terms for the PPR model.
(b) For the optimal model from part (a), what is the PPR model and the CV estimate of the prediction error standard deviation? Interpret the fitted model.
(c) What is the predicted cost for a person with age=59, gend=0, intvn=10, drugs=0, ervis=3, comp=0, comorb=4, and dur=300?

5) Reconsider the forensic glass data that you analyzed in HW2. In HW2 you retained the 6-category response type. But for this problem, you will convert the 6-category response into a binary response type (window glass or other) and perform a binary classification. For consistency, use the misclassification error rate as the CV measure of prediction accuracy.

(a) Use CV to find the best nearest neighbor model for classifying the type.
(b) Use CV to find the best GAM for classifying the type. You can do this using the binomial family in the gam() function.
(c) Compare the models in parts (a) and (b) with the best neural network model.

6) For this problem, you will fit a boosted tree for predicting (the log of) cost for the same ischemic heart disease data that you analyzed in the previous problems.

(a) Find the best boosted tree for predicting cost.
(b) Provide an interpretation of which predictors appear to be the most important and what are their effects. Does this agree with the results from the other methods.
(c) What is the predicted cost for a person with age=59, gend=0, intvn=10, drugs=0, ervis=3, comp=0, comorb=4, and dur=300?

7) For this problem, use the `randomForest` package and function in R to fit a random forest model for predicting (the log of) cost for the same ischemic heart disease data that you analyzed in the previous problems.

a) Fit a random forest model with `mtry=3` and `ntree = 500`. What is the out-of-bag (OOB) r^2? Note that the OOB r^2 is calculated and interpreted similarly to the CV r^2 and is produced automatically by the `print.randomForest` function. Repeat this a few times to see how much the results change from replicate to replicate, and discuss what you see.

b) In part (a), do you think `ntree` was chosen large enough? Explain.

c) Based on the numerical variable importance measure produced by the `importance.randomForest` function, what are the most important variables? Does this agree with what the boosted tree says is the most important?

d) Use the `partialPlot.randomForest` function to produce marginal plots (aka partial dependence plots) for each of the eight predictors. Interpret the plots, in terms of whether the predictors have positive, negative, nonlinear, linear, etc. effects on the response. Also discuss whether what you see in the plots agrees with the variable importance measures from part (c) and whether it agrees with the partial dependence plots for the boosted tree.

e) What is the predicted response for a person with age=59, gend=0, intvn=10, drugs=0, ervis=3, comp=0, comorb=4, and dur=300?

f) Repeat part (a) but for `mtry=1` and `mtry=2`. Which of the three random forests (for the three different `mtry` values) is the best model in terms of predictive performance? You can use the OOB r^2 as the measure of predictive performance. Explain what `mtry` is

g) Out of all the models that you fit in Problems 1—4 and 6—7, and also the tree and neural network model that you fit in HW2, which model appears to be the best for predicting cost? Explain.