Problem Set 5

Jacob Miller, Jacob Shiohira, Reid Gahan

4/14/2018

Problem 1

```
# - Read in the Framinham data set
dataset <- read.csv("Data/framingham.csv", stringsAsFactors=FALSE)

# - How many observations? How many features (explanatory variables) ?

# - Explore the data (plots, tables, statistics) including feature types.

# - Summarize any NA's by feature
```

Problem 2

```
# - Resolve any NA's. Describe your method. I used mice()
# - Do you suggest scaling data or not? If so, normal or uniform scaling? If scaling is advisable, then
# - Randomly split data with 80% in the training set and 20% in the test set.
# - Provide a baseline accuracy by reviewing the TenYearCHD binary variable
# - What is the postive predictive rate (precision) of your baseline prediction ?
# - Which error type seems worse and thus should be considered along with accuracy in your predictions
```

Problem 3

```
# - Create a 'null' logistic regression model with glm() using only the intercept feature. Use glm's bi
# - Create a 'full' logistic regression model with glm() using all of the features including the interc
# - Print summary() for both models
# - Print the log likelihood, the deviance, and the AIC using logLik(model), model$deviance, and AIC(mo
# - Compare the models and comment
```

Problem 4

```
# - Create the lowest AIC ('best') model using backward elimination of parameters. (Suggest labeling th
# - Print the summary(), the log likelihood, the deviance, and the AIC for this model and compare with
# - Print these 'best' parameters (coefficients) along with their 95% confidence intervals in a single
```

Problem 5

```
# - Using the 'best' model, "predict" the 10YrCHD on the training set.
# - Display the confusion matrix, the accuracy, as well as the true positive and true negative rates. U
# - Compare to the baseline model from Question 2
```

```
# - Now predict the 10YrCHD on the test set
# - Display the confusion matrix, the accuracy, as well as the true positive and true negative rates. A
```

Problem 6

```
# Now use ROC curves to reconsider a threshold (cutoff) of .5 using the ROCR (or other) package

# - Create the prediction object for the training set
# - Plot the accuracy v. the threshold (cutoff)
# - Plot the true positive rate v. the false positive rate and label the threshold values
# - Plot sensitivity v. specificity and label the threshold values
# - Plot precision v. recall and label the threshold values
# - Comment on your findings for each plot
# - Compute AUC for the three models (null, full, and best) and comment
```

Problem 7

```
# Using gradient descent, solve for the parameters from the 'best backward eliminaton' data set and ver
# -Create the and test and training sets
# -Compute the parameters (coefficients)
# -Compare to glm()'s coefficients in a single table
```

Problem 8

```
# - look at assignment doc for this problem
```

Problem 9

```
# - Chose another predictive method that you think might produce similar or better results on the Frami
# -Implement using library functions on the training set
# -Create confusion tables for predicting on the test set. You may use any solution parameters or tunin
# -Display the accuracy and the true positive and negative error rates
# -Comment on why this method did or did not result in better accuracy or a better true positive rate (
```

Problem 10

```
# - Read the technical paper ==> http://circ.ahajournals.org/content/97/18/1837

# - How is that method different from logistic regression? One paragraph is sufficient, full understand
```