Hwk #2: Logistic Regression and Linear Discriminant Analysis
Biostat 216: Machine Learning in R for the Biomedical Sciences

For this homework we will continue to use NHANES data that exists in a package for R.

NHANES consists of survey data collected by the US National Center for Health Statistics (NCHS) which has conducted a series of health and nutrition surveys since the early 1960's. Since 1999 approximately 5,000 individuals of all ages are interviewed in their homes every year and complete the health examination component of the survey. The health examination is conducted in a mobile examination center (MEC).

Note that there is a an NHANES warning that comes from their website.
"For NHANES datasets, the use of sampling weights and sample design variables is recommended for all analyses because the sample design is a clustered design and incorporates differential probabilities of selection. If you fail to account for the sampling parameters, you may obtain biased estimates and overstate significance levels."

We are going to ignore this warning and just apply our analyses to the data as if they were randomly sampled! There is a resampled dataset in the package that is re-sampled to represent a more random sample, but we will ignore that here. We will be using the data called `NHANESraw`.

For homework questions below that correspond to tasks in R please provide the code that you used along with relevant output. You can submit your homework based `RMarkdown` if you are comfortable using that, or you can simply cut and paste into this word document.

Note that when you are asked to comment in questions below just one or two sentences should suffice in each case.

Data Preparation

1. Install the package `NHANES` into `R` if you don't already have it installed, load the NHANES package, and then run the command `data(NHANES)` which will load the NHANES data.
2. Make an object `nhanes` that is a subset version `NHANESraw` that does not include any missing data for `Diabetes`, `BPSysAve`, `BPDiaAve`, or `Age` using the following (or your own alterative approach):

```
nhanes <- subset(NHANESraw,subset=is.na(Diabetes)==FALSE &
is.na(BPSysAve)==FALSE & is.na(BPDiaAve)==FALSE & is.na(Age)==FALSE)
```

3. Further subset the data such the observations with `BPDiaAve` equal to zero are removed.

```
nhanes <- subset(nhanes, subset=BPDiaAve>0.1)
```

4. Make an object `nhanes09` that is a subset of `nhanes` to only the 2009_10 data. This will be your training dataset. Also make an object `nhanes11` that is a subset of `nhanes` to only the 2011_12 data. This will be your test dataset.

```
nhanes09 <- subset(nhanes, subset=SurveyYr=="2009_10")
nhanes11 <- subset(nhanes, subset=SurveyYr=="2011_12")
```

Logistic Regression

5. (2 points). Fit a logistic regression model (call it `glm1`) with `Diabetes` as outcome and averaged systolic blood pressure (`BPSysAve`) as a single predictor and use the `summary` command to examine the fitted model. Generate the 95% confidence intervals for the `BPSysAve` coefficient.
6. (1 points). Generate the estimate and 95% confidence interval for the odds-ratio associated with `BPSysAve`. Summarize the result.
7. (1 points). Predict the probabilities of diabetes associated with each of the training observations of `BPSysAve`. Make a vector of predictions for diabetes based on whether the predictions are above or below 0.5.
8. (1 point) Generate a table of predictions vs. actual in the training data.
9. (1 point) Find the proportion of correctly classified observations in the training data.
10. (2 points) Now repeat questions 7 to 9 but for predicting the test dataset.
11. (1 points) Comment on the difference in results between the training and test prediction tables and classification accuracies.
12. (1 points) Generate sensitivity and specificity estimates for the test dataset based on the 0.5 threshold
13. (2 points) Repeat questions 7 to 9 and 12 on the test data but for thresholds of 0.1 and 0.2.
14. (2 points) Comment on the results of the analyses for the different thresholds in terms of the tables, classification accuracies, and sensitivity and specificity. Under what circumstances might you prefer each of the thresholds?
15. (2 points) Fit a multiple predictor logistic regression (call it `glm2`) with `Diabetes` as outcome and predictors: `BPSysAve`, `BPDiaAve`, `Age`. Use the `summary` command to examine the fitted model and determine the estimated coefficients, odds-ratios, and 95% confidence intervals thereof.
16. (2 points) Repeat questions 7 to 9 and 12 on this multiple predictor model for the test dataset (using 0.1, 0.2 and 0.5 thresholds).

17. (1 points) Would you prefer the single predictor or multiple predictor model if your objective was to maximize classification accuracy, and which threshold level would you choose? Comment on the reason for your choices.

Linear discriminant analysis

18. Load the package `MASS`.
19. (2 points) Fit a linear discriminant analysis (`lda1`) with `Diabetes` as outcome and predictors of `BPSysAve`, `BPDiaAve`, and `Age` in the training dataset. Examine the fit by typing `lda1`.
20. (2 points) Generate the test set prediction table, classification accuracy, sensitivity and specificity.
21. (2 points) How do these measures compare with that of the logistic regression model with these predictors and 0.5 threshold?
22. (3 points) Redo questions 19-21 but with prior probabilities set to 0.5 for diabetes (call this `lda2`)
23. (2 points) Comment on the difference between `lda1` and `lda2` with respect to the test set prediction table, classification accuracy, sensitivity and specificity.