STATS 503, Homework 4

1. **[2 POINTS]** ISLP chapter 4, conceptual exercise #2
2. **[2 POINTS]** ISLP chapter 4, conceptual exercise #3
3. **[3 POINTS]** ISLP chapter 4, conceptual exercise #5, parts a, b, and d
4. **[2 POINTS]** ISLP chapter 4, conceptual exercise #7
5. Suppose that you have one continuous predictor X and a binary categorical response Y, which can take values 1 or 2. Suppose that you collect training data from the two classes and obtain class-specific sample means $\hat{\mu}_1 = -1$ and $\hat{\mu}_2 = 3$, together with the pooled variance estimate over the two classes, $\hat{\sigma}^2 = 1$.

   (a) **[1 POINT]** Assume equal class priors and derive the LDA classification rule for this problem. Using `scipy.stats.norm.pdf` to compute the necessary probability density functions, show both of the estimated class-conditional densities in the same plot. Also show the estimated Bayes decision boundary in this plot. Make sure to label the axes. Let c denote the position of the decision boundary. Report the numerical value of c

   (b) **[1 POINT]** Suppose the estimates were in fact obtained from 100 training points, among which 40 were from class 1 and 60 were from class 2. Further, suppose that you will estimate class priors from data. Assume that the class-specific sample means are still $\hat{\mu}_1 = -1$ and $\hat{\mu}_2 = 3$, and the pooled variance estimate over the two classes is still $\hat{\sigma}^2 = 1$. Now you could obtain an estimate of the Bayes decision boundary from this new LDA model; let us call the estimate $\tilde{c}$. Without actually doing this, would you be able to tell whether $\tilde{c}$ will be the same as, less than, or greater than c, or is there no way to tell? Explain your answer without calculating $\tilde{c}$. Note: It is okay to recheck your answer once you have actually calculated $\tilde{c}$ in part (c), but your explanation must not involve the numerical value.

   (c) **[1 POINT]** Now calculate the new boundary value $\tilde{c}$ described in part (b).

   (d) **[1 POINT]** Suppose in addition to the pooled covariance value $\hat{\sigma}^2$ I now tell you that the class-specific covariances were estimated as $\hat{\sigma}_1^2 = 0.25$ and $\hat{\sigma}_2^2 = 1.5$. Based on this new information, would you recommend using LDA or QDA? Why?

   (e) **[1 POINT]** Derive the QDA rule if $\hat{\sigma}_1^2 = 0.25$ and $\hat{\sigma}_2^2 = 1.5$ and $\hat{\mu}_1 = -1$ and $\hat{\mu}_2 = 3$, assuming equal class priors. Calculate the numerical value for all points in the decision boundary of this rule.

6. The `Smarket` data set consists of percentage returns for the S&P 500 stock index over 1,250 days, from the beginning of 2001 until the end of 2005. For each date, we have recorded the percentage returns for each of the five previous trading days, `Lag1` through `Lag5`. We have also recorded the trading `Volume` (in billions of shares) for the previous trading day, and the return and direction (`Up` or `Down`) of the market on the date in question, `Today` and `Direction`. This data set can be downloaded from Canvas.

(a) [1 POINT] Produce some numerical and graphical summaries of the `smarket` data. Do there appear to be any patterns?

(b) [2 POINTS] Fit the LDA model using a training data period from 2001 to 2004, with `Direction` as the response and `Lag1` and `Lag2` as predictors. Use the model to estimate the Bayes optimal predictions on the held out test data (that is, the data from 2005). Compute the confusion matrix and the overall fraction of correct predictions for the test data.

(c) [1 POINT] Repeat (b) using QDA.

7. In this problem, you will develop a model to predict whether a given car will be classified as having high or low gas mileage based on the `Auto` data set. Download this data set from Canvas and load it as a data frame.

(a) [1 POINT] Create a binary variable, `mpg01`, that is equal to 1 if the value of `mpg` for that car is above 25, and 0 otherwise. Add this variable as a new column to your data frame.

(b) [1 POINT] Make some exploratory plots to investigate the association between `mpg01` and other variables. Describe your findings. Besides `mpg` itself, which four quantitative features do you think are most likely to be useful in predicting `mpg01`? There is no right answer here, but you should defend your argument with plots (e.g. side-by-side boxplots).

(c) [1 POINT] Split the data into a training set and a test set. Use

```
sklearn.model_selection.train_test_split
```

to split the data. Be sure to supply the following arguments: `random_state` should be 123, `train_size` should be 0.8, and `stratify` should be set to the values of your new binary feature `mpg01`. Report the number of samples in the trainset where `mpg01` is 1.

(d) [1 POINT] Fit an LDA model on the training data, using the four quantitative variables that seem most associated with `mpg01` based on (b). Use the model to estimate the Bayes optimal predictions for the responses in the training and test data. Report the misclassification rate for your predictions on training and test data. Choose two of the four variables based on (b), and make a scatterplot of the training data points. Use different colors to indicate the true values of `mpg01` and different plotting symbols (e.g., `+` and `o`) to indicate predicted values.

(e) [1 POINT] Repeat exercise 7d, but using QDA this time. Fit a QDA model on training data, use it to estimate Bayes optimal predictions, calculate training and test misclassification rates, and make a plot analogous to the one in problem 7d, using the same two variables.

(f) [1 POINT] Compare and contrast the performance of LDA and QDA. What do your results suggest about the class-specific covariances?

(g) [1 POINT] Repeat exercise 7d, but using logistic regression this time. Fit a logistic regression model on training data, use it to estimate Bayes optimal predictions, calculate training and test misclassification rates, and make a plot analogous to the one in problem 7d, using the same two variables.

(h) **[1 POINT]** Using your fitted logistic regression model, estimate the probability of a car having mpg above 25 if its values for the four predictors are all at the corresponding median values in the training dataset.

(i) **[1 POINT]** Now consider applying KNN classification to this problem. Make plots of the training classification error and the test classification error as a function of the number of neighbors K (or $1/K$; if you use $1/K$, make sure the x-axis is on the log scale). Which K gives the best performance on the training data? On the test data? Which K would you use?

(j) **[1 POINT]** Repeat exercise 7d, using KNN this time, using the number of neighbors chosen in 7i. Fit a KNN classification model on training data, use it to estimate Bayes optimal predictions, calculate training and test misclassification rates, and make a plot analogous to the one in problem 7d, using the same two variables.

(k) **[1 POINT]** Using your fitted KNN model, estimate the probability of a car having mpg above 25 if its values for the four predictors are all at the corresponding median values in the training dataset.

(l) **[1 POINT]** Compare and contrast the performance of LDA, QDA, logistic regression, and KNN on this dataset. What do your results suggest about the distribution of the data? About the nature of the boundary between classes? There is no one right answer here, but we expect you to defend your conclusions using your findings from earlier in this problem.

Total points for this assignment: 30