Homework 4

Instructions

Due: 1:35pm on Wednesday, October 14th

1. Add your name between the quotation marks on the author line in the YAML above.
2. Compose your answer to each problem between the bars of red stars.
3. Commit your changes frequently.
4. Be sure to knit your .Rmd to a .pdf file.
5. Push both the .pdf and the .Rmd file to your repo on the class organization before the deadline.

Theory

Problem 1

Based on ISLR Exercise 4.3

This problem relates to the QDA model, in which the observations within each class are drawn from a normal distribution with a class specific mean vector and a class specific covariance matrix. We consider the simple case where p = 1; i.e. there is only one feature.

Suppose that we have K classes, and that if an observation belongs to the kth class then X comes from a one-dimensional normal distribution, $X \sim N(\mu_k, \sigma_k^2)$. Recall that the density function for the one-dimensional normal distribution is given in (4.11). Prove that in this case, the Bayes' classifier is not linear. Argue that it is in fact quadratic.

Hint: For this problem, you should follow the arguments laid out in Section 4.4.2

Problem 2

Based on ISLR Exercise 4.5

We examine the differences between LDA and QDA.

a. If the Bayes decision boundary is linear, do we expect LDA or QDA to perform better on the training set? On the test set? Why?

b. If the Bayes decision boundary is non-linear, do we expect LDA or QDA to perform better on the training set? On the test set? Why?

c. In general, as the sample size n increases, do we expect the test prediction accuracy of QDA relative to LDA to improve, decline, or remain unchanged? Why?

d. True or False? Even if the Bayes decision boundary for a given problem is linear, we will achieve usually achieve a superior test error rate using QDA rather than LDA because QDA is flexible enough to model a linear decision boundary. Justify your answer.

Problem 3

Based on ISLR Exercise 4.7

Suppose that we wish to predict whether a given stock will issue a dividend this year ("Yes" or "No") based on X, last year's percent profit. We examine a large number of companies and discover that the mean value of X for companies that issued a dividend was $\bar{X} = 10$, while the mean for those that didn't was $\bar{X} = 0$. In addition, the variance of X for these two sets of companies was $\hat{\sigma}^2 = 36$. Finally, 80% of companies issued dividends. Assuming that X follows a normal distribution, predict the probability that a company will issue a dividend this year given that its percentage profit was $X = 4$ last year.

Hint: Use Bayes' Theorem.

Applied

Problem 4

Based on ISLR Exercise 4.11

In this problem, you will develop a model to predict whether a given car gets high or low gas mileage based on the `Auto` data set from the `ISLR` package, which can be loaded with the following code:

```
library(ISLR)
data(Auto)
```

a. Mutate the `Auto` data frame to create a binary variable `mpg01` that takes the value 1 if `mpg` is larger then the median value of `mpg` and 0 otherwise.

b. Produce some numerical and graphical summaries of the `Auto` data in order to investigate the relationship between `mpg01` and the other features. Discuss your findings. Which variables seem most useful in predicting `mpg01`?

c. Randomly divide the data into a training and a test set using a $75 - 25$ ratio. Be sure to set a seed for reproducibility.

d. Perform LDA on the training data to predict `mpg01` using the variables that seemed most associated with `mpg01` from part b. Then create a confusion matrix and compute the test error for the model. Finally, create an ROC curve for the model.

e. Perform QDA on the training data to predict `mpg01` using the variables that seemed most associated with `mpg01` from part b. Then create a confusion matrix and compute the test error for the model. Finally, create an ROC curve for the model.

f. Perform Logistic Regression on the training data to predict `mpg01` using the variables that seemed most associated with `mpg01` from part b. Then create a confusion matrix and compute the test error for the model. Finally, create an ROC curve for the model.

g. Perform KNN on the training data (with 3 different values of K) to predict `mpg01` using the variables that seemed most associated with `mpg01` from part b. Then create a confusion matrix and compute the test error for the model.

h. If had to select *ONE* model to make predictions about `mpg01`, which would you use and why?

Problem 5

For this exercise we will use data found in `wisc-trn.csv` and `wisc-tst.csv` which contain train and test data respectively. The original data set `wisc.csv` is also provided, but not used. This is a modified version of the Breast Cancer Wisconsin (Diagnostic) dataset from the [UCI Machine Learning Repository](https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic). Only the first 10 feature variables have been provided.

We are interested in models predicting the value of `class`.

Decision Boundaries

a. Create a logistic regression model using **training** data with two predictors, `radius` and `symmetry`. Plot the **test** data with `radius` on the x axis and `symmetry` on the y axis, with points colored according to their tumor status. Use `geom_abline` to add a line representing the decision boundary for the logistic regression classifier, using 0.5 as the threshold for predicted probability. *Hint: If $p(X) = 0.5$, what must the log-odds be equal to?*

b. Based on a visual inspection of the graphic you created in the part (b), how well did the logistic regression model do?

c. Repeat parts (b) using 0.1 as the threshold for predicted probability. How did model accuracy change? Consider both the rate of false positives and the rate of false negatives.

A Handmade Model

d. Use an appropriate visualization to investigate the conditional distributions of each predictor (`symmetry`) given the two values of `class`. Does each appear to be approximately Normal? Do the two conditional distributions for `symmetry` appear to have the same variance?

e. Even if the conditions for LDA are not met, we will proceed as if they are. Create estimates for the mean for each conditional distribution and the pooled variance, and use the training proportions as estimates for the prior distribution of `class`.

f. Use the formula for the discriminant in Section 4.4 to write explicit linear equations for the discriminants for the two levels of `class`. Plot these two lines using `geom_abline`. Estimate the value of the decision "boundary" based on this graph and then compute the value exactly using the equations of your two lines.

g. Classify the points in the training set according to your decision boundary. What is your error rate?

h. **Challenge Problem** (This part is optional and requires some linear algebra) Create estimates for the mean of the conditional distribution of `radius` and estimate the covariance matrix for `symmetry` and `radius`. Then use the formula for the discriminant with $p = 2$ predictors (equation 4.19) to find explicit equations for the discriminant planes. Find the decision boundary corresponding to the intersection of these planes and add it to the scatterplot of `radius` and `symmetry` colored by `class`.
