Introduction to Statistical Learning, Spring 2016

Homework exercise 2

Due date: 3 May 2016 in class

1. **Short questions on classification algorithms**

   (a) For $K = 2$ classes, we compare linear regression with $\{0, 1\}$ coding to logistic regression. For each of the following, state whether it is a property of logistic regression, linear regression or both:

      i. The expected prediction error is minimized by correctly predicting $P(Y|X)$.

      ii. The predictions are always legal probabilities in the range $(0, 1)$.

   (b) In the generative 2-class classification models LDA and QDA, what type of distribution does $P(Y|X = x)$ have?

      i. Unknown — can be anything

      ii. Gaussian

      iii. Bernoulli

   (c) If the Bayes decision boundary is linear, do we expect LDA or QDA to perform better on the training set? On the test set?

   (d) In general, as the sample size n increases, do we expect the test prediction accuracy of QDA relative to LDA to improve, decline, or be unchanged? Why?

2. Suppose that we wish to predict whether a given stock will issue a dividend this year (Yes or No) based on X, last years 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 didnt was $\bar{X} = 0$. In addition, the variance of X for these two sets of companies was $^2 = 36$. 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 the formula of the normal density and basic probability rules.

3. We take a data set, divide it into equally-sized training and test sets, and then try out two different classification procedures. First we use logistic regression and get an error rate of 20 % on the training data and 30 % on the test data. Next we use 1-nearest neighbors (i.e. K = 1) and get an average error rate (averaged over both test and training data sets) of 18 %. Based on these results, which method should we prefer to use for classification of new observations? Explain your answer.

4. ISLR 5.7 (fitting logistic regression with LOOCV to the Weekly data).