Homework 4. Logistic Regression, LDA, and QDA

Stat 427/627 Statistical Machine Learning

The due date is announced on Canvas.

Contents

This assignment covers classification, including Logistic regression, LDA, and QDA.

Question	1	2	3	4	Total
427/ 627	10	4	4	7	25

1 Ex.4.8.13.(with modification) Stock Market Prediction, Part 2. Logistic Regression (p.192, 10 pts).

We want to predict the behavior of the stock market in the following week. The `Weekly` data set, (from the {ISLR2} package), contains 1,089 observations with the following 9 variables.

- `Year`: The year that the observation was recorded
- `Lag1`: Percentage return for previous week
- `Lag2`: Percentage return for 2 weeks previous
- `Lag3`: Percentage return for 3 weeks previous
- `Lag4`: Percentage return for 4 weeks previous
- `Lag5`: Percentage return for 5 weeks previous
- `Volume`: Volume of shares traded (average number of daily shares traded in billions)
- `Today`: Percentage return for this week
- `Direction`: A factor with levels `Down` and `Up` indicating whether the market had a positive or negative return on a given week

(a) Perform a logistic regression with `Direction`as the response and the five lag variables plus `Volume` as predictors. Do any of the predictors appear to be statistically significant? If so, which ones?

(b) Compute the confusion matrix and overall fraction of correct predictions. Explain what the confusion matrix is telling you about the types of mistakes made by the logistic regression model.

(c) Fit a logistic regression model with `Lag2` as the only predictor using as training data the period from 1990 to 2008. Compute the confusion matrix and the overall fraction of correct predictions for the held out test data (that is the data from 2009, and 2010). How does it compare to before? What does that suggest?

(d) Plot an ROC curve for the logistic regression on the test data from (c), using different probability thresholds and add an diagonal line for FPR = TPR. Interpret the plot in terms of how useful the model might be?

2 Ex.4.8.5, LDA v QDA (p.191, 4 pts)

(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?

(b) Compare the expected performance of LDA and QDA on the training set and then on the test set if the Bayes decision boundary is *non-linear.*

(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 be unchanged? Why?

(d) **True or False**: If the Bayes decision boundary for a given problem is linear, we will probably 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.**

3 Ex.4.8.7. Non-uniform Prior. Predicting Issuance of a Stock Dividend (p.191, 4 pts)

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:

- The mean value of percent profit X for companies that issued a dividend was $\bar{x}_1 = 10$, while the mean for those that didn’t was $\bar{x}_2 = 0$.
- The variance of X for these two sets of companies was $\sigma^2 = 36$.
- 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: Recall that the density function for a Normal random variable is $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$. You will need to use the Bayes theorem. Also, use R function `dnorm(x, mean, sd)` to compute the pdf ($f(x)$) of Normal distribution.

4 Ex.4.8.13.(with modfication) Stock Market Prediction, Part 3. LDA, QDA and Summary (7 pts)

We want to continue with trying to predict the behavior of the stock market in the following week. We computed a KNN estimate and a Logistic regression estimate already. We will now look at LDA and QDA using the {`MASS`} package and compare across these prediction methods.

The `Weekly` data set, (from the {ISLR2} package), contains 1,089 observations with the following 9 variables.

- `Year`: The year that the observation was recorded
- `Lag1`: Percentage return for previous week
- `Lag2`: Percentage return for 2 weeks previous
- `Lag3`: Percentage return for 3 weeks previous
- `Lag4`: Percentage return for 4 weeks previous
- `Lag5`: Percentage return for 5 weeks previous

- `Volume`: Volume of shares traded (average number of daily shares traded in billions)
- `Today`: Percentage return for this week
- `Direction`: A factor with levels `Down` and `Up` indicating whether the market had a positive or negative return on a given week

(a) Use LDA with a training data period from 1990 to 2008, with `Lag2` as the only predictor. Compute the confusion matrix and the overall fraction of **correct predictions** for the held-out data (that is, the data from 2009 and 2010).

(b) Repeat (a) using QDA.

(c) Using the results from the previous questions (Stock Market Predictions in HW 3 and HW 4), compare the correct classification rates on the testing data obtained from the 4 methods: KNN (in HW 3), Logistic regression, LDA and QDA. Recommend one or more methods. Explain your rationale.

—— **This is the end of HW 4.** ——