Homework 3

Stat 4540: Statistical Learning

Due on 10/18/2024

1. (5 pts) ISLR 4.8.1.
2. (5 pts) ISLR 4.8.5.
3. (20 pts) ISLR 4.8.6.
4. (20 pts) ISLR 4.8.14 (a)–(h).
5. (50 pts) We will use the modified form of the 100K MovieLens data in this question. Download the `quality_train.csv` and `quality_test.csv` data from ICON, where the former and latter files are used for training and testing logistic regression, LDA, QDA, KNN, and naive Bayes classification models. Each data set contains columns named `category`, `popular`, and `genre`. The `category` column is computed using the user ratings and the genre column is defined using the movie genres. We use `category` as the response and `popular`, `genre` as the predictors.

   (a) Answer the following questions about predicting movie category using logistic regression.
      i. Describe how would to use logistic regression for predicting category. Fit the required logistic regression models.
      ii. Obtain estimates of the regression coefficients and construct 95% confidence intervals for them.
      iii. Evaluate the performance of logistic regression on the testing data using confusion matrix, overall fraction of correct predictions, TPR, FPR, sensitivity, and specificity.

   (b) Use LDA to predict movie category. Repeat (a)iii for LDA.

   (c) Use QDA to predict movie category. Repeat (a)iii for QDA.

   (d) Use KNN to predict movie category. Use K in $\{1, 5, 10, 50\}$. Repeat (a)iii for each KNN method.

   (e) Use naive Bayes to predict movie category. Repeat (a)iii for naive Bayes.

   (f) Which of the four method performs best and why? Argue using the bias-variance tradeoff.