Introduction to Data Science Assignment 2

Due time: 11:59 pm, October 14, 2023
5 questions and 100 points in total

This assignment covers the sections of “Statistical Learning”, “Regression”, “Quantifying Uncertainty” and “Resampling Methods.”

Every student is expected to complete the assignment independently. Discussion between students is encouraged, but no plagiarism will be tolerated.

Please submit your answers in a PDF file. If you want to also submit your code separately, you can pack all files, including a readme file, into a compressed file.

Question 1 (45 points + 5 bonus points)

Let us conduct regression analysis on the bike sharing data set. The data set description is available at <https://archive.ics.uci.edu/ml/datasets/bike+sharing+dataset>, and the data set can be downloaded at <https://archive.ics.uci.edu/ml/machine-learning-databases/00275/Bike-Sharing-Dataset.zip>.

We will use only the data in file <day.csv>.

1. (5 points) Plot the marginal distribution for the count of bike rentals and the conditional count distribution given the weather situation (weathersit).
2. (5 points) Fit a linear regression for ride count as a function of the weather situation variable. Report the coefficients.
3. (5 points) What is the difference in expected ride counts when the weather is clear (1) versus wet (3)?
4. (5 points) What are the in-sample residual sum of squares (RSS), R^2, and estimated standard deviation of the residual errors for the ride-weather regression?
5. (5 points) Fit a linear regression for the ride counts onto all of the weather variables (weathersit, temp, hum, windspeed). You can ignore the date/time variables. What is the impact on expected ride count due to a 10 degree increase in the temperature?
6. (10 points) Add interactions between the continuous weather variables and the weathersit factor. For each weather situation, what is the change in expected ride count per 10 degree increase in temperature?
7. (10 points) Fit a logistic regression to model ride counts using weather variables (weathersit, temp, hum, windspeed), excluding date/time variables. To categorize the

total rental bike count, establish qualitative labels: 'Low Demand' and 'High Demand.' Specifically, designate counts less than or equal to 4000 as 'Low Demand' and counts greater than 4000 as 'High Demand.' Evaluate and report the logistic regression model's performance using 5-fold cross-validation accuracy.

8. (Optional bonus question; 5 bonus points) Choosing an appropriate threshold in logistic regression hinges on the specific context of your problem and analysis goals. In contrast to the 4000 threshold employed in the prior exercise, opt for a different threshold guided by some criteria. Clearly articulate the criterion you consider for this choice. Evaluate and report this logistic regression model's performance, including accuracy and F-1 score metrics, in comparison to the previous exercise (question 1.7). Additionally, provide a rationale for any observed improvement or lack thereof in performance.

   Note that the F-1 score is calculated using the formula: 2 * TP / (2 * TP + FP + FN), where TP represents true positives, FP represents false positives, and FN represents false negatives.

Question 2 (20 points)

For each of the following four scenarios below, determine whether a flexible statistical learning method is expected to perform better or worse than an inflexible method. Provide a rationale for your answer.

1. (5 points) When the sample size is extremely large, and the number of predictors is small.
2. (5 points) When the number of predictors is extremely large, but the number of observations is small.
3. (5 points) In cases where the relationship between the predictors and the response is highly non-linear.
4. (5 points) When the variance of the error terms, denoted as $\sigma^2 = \text{Var}(\epsilon)$, is extremely high.

Question 3 (10 points)

Explain the practical applications of regression in two real-life scenarios. In one scenario, emphasize the primary objective of inference, while in the other, focus on prediction. For each case, clearly elucidate the response variable and the predictor variables, and provide a detailed rationale for why each application is categorized as either inference or prediction.

Question 4 (15 points)

1. (5 points) Using the least squares coefficient estimates, prove that in the case of simple linear regression, the least squares line always passes through the point $(\bar{x}, \bar{y})$.
2. (5 points) Prove that in the case of simple linear regression with Y as the response variable and X as the predictor variable, the R^2 statistic is equal to the square of the correlation between X and Y. For simplicity, you may assume that the means of X and Y are both zero (i.e., $\bar{x} = \bar{y} = 0$).
3. (5 points) Prove that the logistic function representation and logit representation for the logistic regression model are equivalent. In other words, derive the odds from the logistic function.

Question 5 (10 points)

Suppose we are analyzing the performance of employees in a company, with variables: X_1 = hours spent on professional development, X_2 = years of experience, and Y = receive a promotion. We fit a logistic regression model and obtain estimated coefficients: $\hat{\beta}_0$ = -4, $\hat{\beta}_1$ = 0.03, $\hat{\beta}_2$ = 0.5.

1. (5 points) Estimate the probability that an employee who spends 30 hours on professional development and has 5 years of experience receives a promotion.
2. (5 points) How many hours of professional development should the employee in the prior exercise (question 5.1) undertake to have a 50% chance of receiving a promotion?