Final 40015/50015

Due Dec 10, 2023 at 11:59pm

For this final assignment you are allowed to work with up to 5 people from your class (you can work by yourself if you wish). If you choose to work in a group, then only one person from the group will submit the groups solutions. All the names of those who worked together will be at the top of your submission. In addition, each person in the group will post on their submission page a grade from 0-5 for each of the members in the group. 0 meaning that the person contributed nothing and 5 meaning that the individual made good contributions to the group. Failing to submit this information will result in a grade reduction.

Do not print out datasets in your rmd file.

1. (a) Load the package alr4 into memory
   (b) From the dataset *Downer*, construct a dataframe using the variables claved, daysrec, ck, ast, urea, and pcv. Remove any rows with missing data.
   (c) Construct a logistic regression model with explanatory variables given in (1b) and outcome as the response. Use an 80/20 split of the data.
   (d) Construct a confusion matrix. How accurate is your model?

2. The *Boston* , from the **MASS** library, contains information about different suburbs in Boston and various factors that might influence house prices. Your ojective is to predict house prices (medv) based on other available features and identify the most influential predictors using Lasso regression.

   (a) Loading the Dataset:

   - Load the Boston dataset from the `MASS` package.

   (b) Lasso Regression Modeling:

   - Use Lasso regression (`glmnet` package) to build a predictive model for house prices.
   - Tune the model using cross-validation (`cv.glmnet` function) to find the optimal lambda (regularization parameter) for Lasso regression.

(c) Variable Selection:

- Extract coefficients from the Lasso model to identify which features have non-zero coefficients.
- Non-zero coefficients indicate selected features with significant predictive power.

(d) Evaluation:

- Evaluate the performance of the Lasso model using cross-validated metrics such as Mean Squared Error (MSE) or others (`cv.glmnet` output).
- Assess the selected features' importance in predicting house prices.

(e) Interpretation:

- Interpret the results, including the selected features and their coefficients, to understand their impact on house prices.
- Discuss the implications and significance of the selected features in predicting house prices in Boston suburbs.

3. Consider the `faithful` dataset in R, which records the waiting time between eruptions and the duration of eruptions of the Old Faithful geyser in Yellowstone National Park. Perform polynomial regression to model the relationship between waiting time and eruption duration.

   Utilizing 10-fold cross-validation, compute the cross-validated R^2 values for polynomial regression models with degrees ranging from 1 to 4.

   Determine the best-fitting polynomial regression model by selecting the degree that yields the highest average R^2 value across the different degrees.

   Tasks:

   (a) Load the `faithful` dataset in R.

   (b) Implement polynomial regression models with degrees from 1 to 4.

   (c) Use 10-fold cross-validation to compute R^2 values for each model.

   (d) Identify the degree that corresponds to the highest average cross-validated R^2 value.

   Provide the selected degree and its corresponding average R^2 value as the solution.

4. Suppose we have a regression in which we want to fit the mean function (3.1). Following the outline in Section 3.1, suppose that the two terms X_1 and X_2 have sample correlation equal to 0. This means that, if x_{ij}, $i = 1, \ldots, n$, and $j = 1, 2$ are the observed values of these two terms for the n cases in the data, $SX_1X_2 = \sum_{i=1}^{n}(x_{i1} - \bar{x}_1)(x_{i2} - \bar{x}_2) = 0$. Define $SX_jX_j = \sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2$ and $SX_jY = \sum_{i=1}^{n}(x_{ij} - \bar{x}_j)(y_i - \bar{y}_j)$ for $j = 1, 2$.

   (a) Give the formula for the slope of the regression for Y on X_1, and for Y on X_2. Give the value of the slope of the regression for X_2 on X_1.

   (b) Give formulas for the residuals for the regressions of Y on X_1 and for X_2 on X_1. The plot of these two sets of residuals corresponds to the added-variable plot for X_2.

   (c) Compute the slope of the regression corresponding to the addedvariable plot for the regression of Y on X_2 after X_1, and show that this slope is exactly the same as the slope for the simple regression of Y on X_2 ignoring X_1.

5. Let $H = [h_{ij}]$ be the Hat Matrix. If $X_{n\times(p+1)}$ is of full rank and contains a first column of 1's, prove that $\frac{1}{n} \leq h_{ii} \leq 1$.