PES University, Bengaluru

UE18CS312 - Data Analytics

Session: Aug – Dec 2020
Week 5 – Worksheet 2(b) (for Unit 2)

Compiled by: Ms. Richa and Ms. Mainaki Saraf
VII CSE, PES University RR Campus

3.0 Getting Started

The “Hitters” dataset which is a part of R’s ISLR package will be used for this worksheet. The dataset is also available at https://gist.github.com/keeganhines/59974f1ebef97bbaa44fb19143f90bad .

1. Read this dataset
2. Remove/Fill in the missing values using a suitable technique
3. The main motive is to build a regression model using ridge and lasso regression techniques to predict the salary of a hitter given all the other variables.
4. Remove the columns which have nothing to do with the target variable.

3.1 Ridge Regression

1. Explore the **glmnet** library in R for this exercise.
2. Alter the arguments on the glmnet() to fit a ridge regression model to the data. Which argument has to be changed and what should it be set to implement this model?
3. Are the variable values standardized? What can be done to avoid the standardization? Hint: **Check glmnet arguments**
4. How does the model select the values for ƛ? Is it possible to restrict the possible values to a particular range?
5. Try retrieving the coefficient estimates. In what form are they? What is the shape of the same? When are the estimates expected to be of a larger value? When are they of smaller values?
6. Split the data into test and train sets and fit the model with the training set. Predict the values for the test set and record the MSE value.
7. Try changing the lambda value used for prediction to see how the MSE value varies. What will the value of lambda be if the model was to replicate a least-square fit?

3.2 Lasso Regression

1. From the above exercise of ridge regression, get the best possible lambda value and use that lambda value to fit a lasso regression model to the train set.
   Hint: **Make changes in the parameter of glmnet()**
2. Try printing the predictors which were kept by the model and the ones which were discarded.

3. Predict the values for the test set. Calculate the MSE and compare to that of the ridge regression model with the same lambda value.

4.1 Polynomial Regression

For this part of the worksheet, the “Wages” dataset in R will be used. It is also available at https://www.picostat.com/dataset/r-dataset-package-islr-wage
Read the dataset to the file. The model which will be built will predict the wage of a person given certain input variables.

1. Plot the “Age” and “Wage” columns and observe their relationship. What kind of a relationship is it?
2. Taking only the “Age” column as input. Try fitting a linear regression model. Now, change the degree of this column to 2 and fit another linear regression model. Hint: Explore **poly()**. What does poly() return? Determine the new equation for the model.
3. Keep increasing the degree upto 5 and compare how each of the models performed and record the results. Which degree upto 5 performed the best? Play around with different degrees and see how the model behaves.
4. Explore other methods to compare how the various models with different degrees performed. Hint: **compare p-values using anova()** for all the fitted models. Which model provides a reasonable fit based on this test? Give reasons.

4.2 Logistic Regression

For this part of the worksheet the “Smarket” dataset from R’s ISLR library will be used. Its also available at https://github.com/selva86/datasets/blob/master/Smarket.csv
Read the dataset to the file. A logistic regression model is to be fit to predict the Direction using Lag1 to Lag5 and the Volume variable.

1. Split the data into test and train sets. How will you perform the split? Will a random split give the required results? If not random, then which approach will you use? Hint: Look into the **Year** column
2. Try fitting a logistic regression model for the target variable by giving the above-mentioned variables as input for the training set. Hint: Explore **glm()**
   Which argument is altered to fit the logistic regression model and what is the value given?
3. Get a summary of the model built. Get the coefficients and residuals too and observe how different input variables play a role in determining the target value.
4. Try predicting the values for the test set. In which form are the predictions? What can be done to get the class labels instead as the predicted output? Hint: **Set a threshold** on the output value to classify it as one class.
5. Check how the model performed for the test set. Which metric/metrics will you use to evaluate the model’s performance?