Home Work # 5. AMS 380

Name:_____SBU ID:_____

Dear all, the homework is due on Tuesday, Oct 5, 2021, at 11:59 PM. Please submit your homework to the Blackboard in a pdf or word (.doc) document. Rmarkdown is highly recommended.

Please include (1) R code; (2) Output from R; (3) Answers to all the questions asked.

Please refer to the following website website for steps for Penalized Regressions in R:
http://www.sthda.com/english/articles/37-model-selection-essentials-in-r/153-penalized-regression-essentials-ridge-lasso-elastic-net/

Penalized Regression with the Ames Housing Data

The accompanying csv file contains data on various aspects of houses sold. Our goal is to predict the SalePrice using all the other variables. The variables are:

- SalePrice - the property's sale price in dollars. This is the target variable that you're trying to predict.
- ID – the ID number we assigned to each house. There are a total of 1460 houses in this data set.
- LotArea: Lot size in square feet
- OverallQual: Overall material and finish quality
- OverallCond: Overall condition rating
- YearBuilt: Original construction date
- YearRemodAdd: Remodel date
- CentralAir: Central air conditioning
- 1stFlrSF: First Floor square feet
- 2ndFlrSF: Second floor square feet
- GrLivArea: Above grade (ground) living area square feet
- FullBath: Full bathrooms above grade
- HalfBath: Half baths above grade
- Bedroom: Number of bedrooms above basement level
- Kitchen: Number of kitchens
- TotRmsAbvGrd: Total rooms above grade (does not include bathrooms)
- Fireplaces: Number of fireplaces
- GarageCars: Size of garage in car capacity
- GarageArea: Size of garage in square feet
- YrSold: Year Sold

1. Please perform data cleaning by checking whether there are any missing values and if so, please delete observations with missing values. Please report how many observations with missing values we have in our dataset.

2. Please use the random seed 123 to divide the data into 75% training and 25% testing.

3. Please first find the best Ridge Regression model using the training data. Please (a) find the best **λ** value through cross-validation and display this value; (b) display the coefficients of the fitted model; and (c) make prediction on the testing data, plot the observed response variable on the x-axis, and the estimated response variable on the Y-axis, and report the RMSE and the Coefficient of Determination R^2.

4. Please first find the best LASSO model using the training data. Please (a) find the best **λ** value through cross-validation and display this value; (b) display the coefficients of the fitted model; and (c) make prediction on the testing data, plot the observed response variable on the x-axis, and the estimated response variable on the Y-axis, and report the RMSE and the Coefficient of Determination R^2.

5. Please first find the best Elastic Net model using the training data. Please (a) find the best **tuning parameter** values through cross-validation and display these values; (b) display the coefficients of the fitted model; and (c) make prediction on the testing data, plot the observed response variable on the x-axis, and the estimated response variable on the Y-axis, and report the RMSE and the Coefficient of Determination R^2.

6. Please discuss which penalized regression method is the best for the Ames Housing data, and why.