CS 482 MACHINE LEARNING

ASSIGNEMNT III

Ch5 and Ch6

Dimensionality Reduction and Non Linear Models

For each step below, provide tables, visuals with explanation of what is being shown and results. Submit the report and the code. Do NOT copy and paste code into the report.

Provided is a dataset for regression. Also there is a data description text file describing the predictors. The task is to predict sale price of homes.

1. **MEET THE DATA** :Provide a Meet the Data Section to introduce the data. Include the following in the Meet the Data section

a) Number of features (comes from the load function above)
b) Names of the features (comes from load function above)
c) Name of target (comes from load function above)
d) Number of samples (comes from the load function above)
e) First five rows of the data

2. **DATA PREPROCESSING:**

2A) Delete all columns that have unique values (list if you deleted any)

2.B) Converting Words to numeric values: List the columns with non-numeric values and explain what encoding you used and why.

Column with word values	Encoding Used	Reasons for your choice

2B) Fill in missing information in the dataset if there are any. Explain why you chose to fill the data the way you chose. Provide a table with which columns had missing information and how you filled it in and why you chose this particular scheme for filling in missing values

Column with missing values	Percentage of values missing	How you chose to fill in missing values	Reasons for your choice

2C) State the number of features after the steps above

3. LEARNING FROM DATA:

Show visual displays with histograms for first two features. Make sure you label and write about the display. Use only training data for this purpose. Split 80:20 with Random_state =42 and shuffle set to True for splitting the data.

4. FEATURE EXTRACTION

4A) Remove 30% of the features that have least correlation to the target. Show visual display of correlation numbers. State the number of features removed. State the correlation cutoff value for removing 30% of the features.

4B) Using model based feature selection using remaining features, (use lasso regression as your base model with best alpha value between 1 to 10 with increment of 5), select top 70% of the remaining features that highest value of coefficient β. State the number of features removed. Display the co-efficient values as in the chart below:

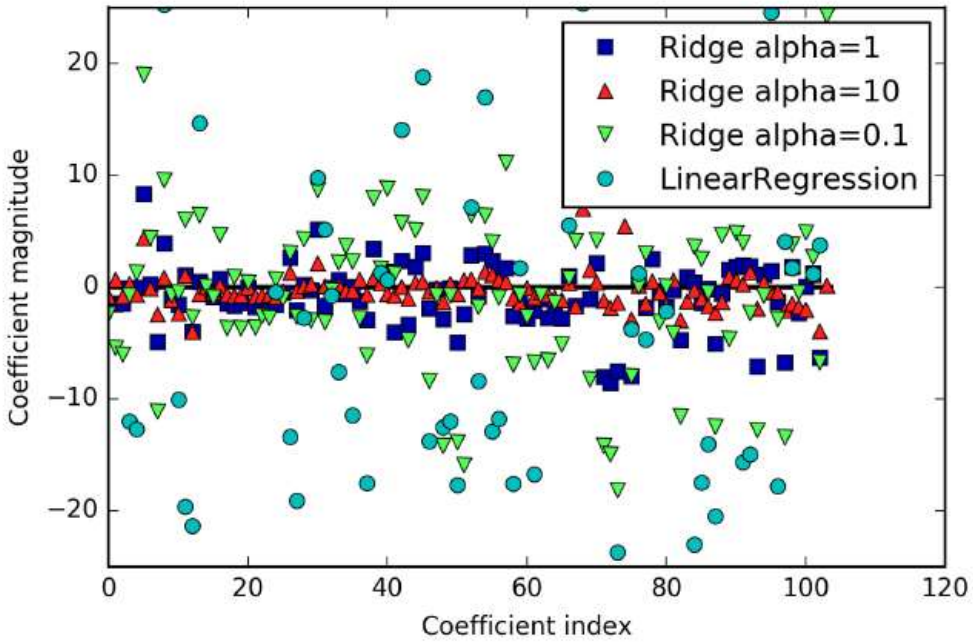

Figure 2-12. Comparing coefficient magnitudes for ridge regression with different values of alpha and linear regression

List the remaining features (if you have less than 50 features, otherwise simply state the number).

4C) Use PCA (Principal component Analysis) on the remaining features. Use 10% of features as the number of components of PCA.

5. **MODEL DEVELOPMENT:** Now use SVM (rbf kernel with gamma from 1 to 10 with increment of 5 and Cost from 10 to 100 with increment of 10) and Neural network (with 2 hidden layers with varying number of units and random initial weights, random_seed = 42) with parameter tuning to get the model with best performance.

6. **MODEL EVALUATION** Show a chart indicating the best performance of each model. Use model metrics that you learned in the previous assignment for comparing the models.