

Project 1: Linear Regression using K-Fold Cross Validation on IRIS dataset

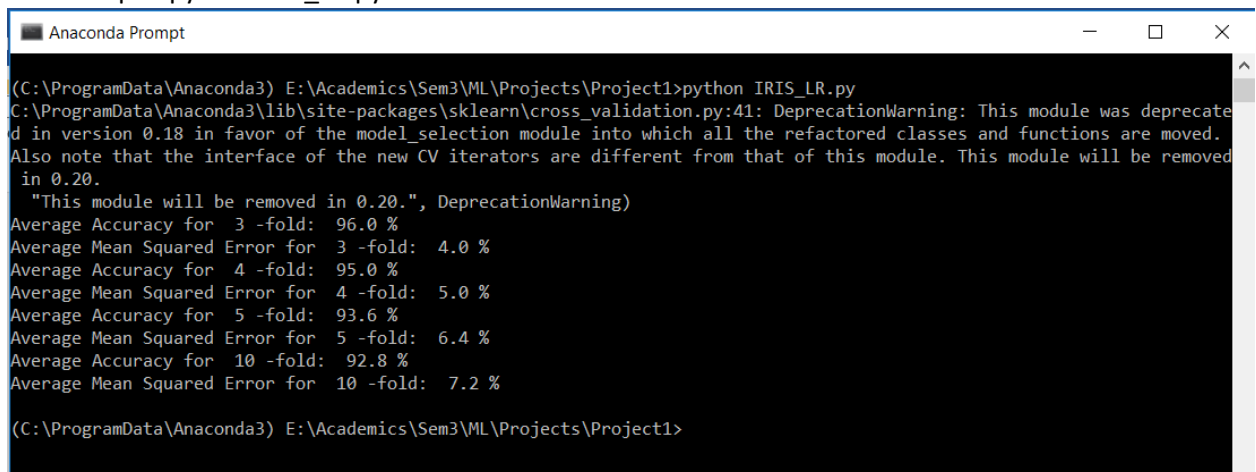
Rathna Sindura Chikkam

Procedure:

I have used K-Fold cross validation library from Scikit-Learn package to split the (150,4) IRIS data set into train and test data. I have tested [3,4,5,10] folds to perform linear regression on the data. For each fold, I would use the train data of both X(observations) and Y(class labels) to calculate the Beta vector, and then used the Beta Vector to do a dot product of Beta Vector($B_{\hat{}}$) with test data of X, to find the predictions of Y (classified X test data by performing dot product of Beta Vector and X test data). Using metrics library of Scikit-Learn, for each fold, I have calculated average accuracy score and average mean squared error values and I am printing the same as the output of the program.

To Run the file:

- Unzip my folder and use either Spyder 3.0 software or anaconda prompt or cmd prompt to run the program from the unzipped folder example
- Type "python <filename>" as shown below:
- For example: python IRIS_LR.py



```
(C:\ProgramData\Anaconda3) E:\Academics\Sem3\ML\Projects\Project1>python IRIS_LR.py
C:\ProgramData\Anaconda3\lib\site-packages\sklearn\cross_validation.py:41: DeprecationWarning: This module was deprecated in version 0.18 in favor of the model_selection module into which all the refactored classes and functions are moved. Also note that the interface of the new CV iterators are different from that of this module. This module will be removed in 0.20.
  "This module will be removed in 0.20.", DeprecationWarning)
Average Accuracy for 3 -fold: 96.0 %
Average Mean Squared Error for 3 -fold: 4.0 %
Average Accuracy for 4 -fold: 95.0 %
Average Mean Squared Error for 4 -fold: 5.0 %
Average Accuracy for 5 -fold: 93.6 %
Average Mean Squared Error for 5 -fold: 6.4 %
Average Accuracy for 10 -fold: 92.8 %
Average Mean Squared Error for 10 -fold: 7.2 %

(C:\ProgramData\Anaconda3) E:\Academics\Sem3\ML\Projects\Project1>
```

Results:

Below are my findings on training the IRIS data set:

Average Accuracy for 3 -fold: 96.0 %

Average Mean Squared Error for 3 -fold: 4.0 %

Average Accuracy for 4 -fold: 95.0 %

Average Mean Squared Error for 4 -fold: 5.0 %

Average Accuracy for 5 -fold: 93.6 %

Average Mean Squared Error for 5 -fold: 6.4 %

Average Accuracy for 10 -fold: 92.8 %

Average Mean Squared Error for 10 -fold: 7.2 %

With these results, the least-fold, $K=3$ seems to be perfect choice as gives highest accuracy and least error among the folds tested for.