**Method:**

I applied linear regression with 5 Fold cross validation because there aren’t many data points to work with and going beyond the 5 fold(say 10 fold) will lead to a small test set and a large training set which leads to overfitting.

**Code Segmentation:**

The way my code works in the following way,

1. Reads the iris.data file which shuffles and creates iris\_shuffled.data file which is then used to create 2 new files train.data, test.data containing training and test data respectively.
2. Reads the data from the train.data file and creates a numpy array named “A” which is our input matrix.
3. Also reads the output labels “Y” from the train.data within the same loop.
4. Calculates the equation(B = (AT\*A)-1 \* AT \* Y) to fit parameters for the equation of line for the provided training set. (In simple words, we just calculated the beta vector i.e., weights of the input parameters.) Note: There is a bias parameter in the beta vector for which I included a (Xo = 1) for the input matrix “A”.
5. Use the trained model to predict the data points in test data and fond the accuracy of the model.
6. Repeat the same process for each iteration of the 5-fold cross validation.
7. Find the average beta vector and the accuracy for that model.

**Report and Results:**

Based on the linear regression results I observed few outliers in the data, more prominently in the classes iris-versicolor & iris-virginica. (What I mean by outliers is that there are few data points which go beyond their range and thus are being incorrectly classified as wrong class by the model means those classes have overlapping ranges.) The range of values for classes Iris\_setosa – 0.5 to 1.5, Iris\_versicolor – 1.5 to 2.5, Iris\_virginica – 2.5 to 3.5. The average accuracy result obtained by using 5-fold cross validation is 95. 33%. Equation of the model is 1.1930372-0.11025569\*x2-0.04375696\*x2+0.22583366\*x3+0.61382056\*x4 where x1,x2,x3,x4 are the 4 parameters of the iris.data file.

**Note:**

1. I shuffled the data because without shuffling the given data(which is in ordered format i.e., First 1-50 examples as iris-setosa, 51-100 examples as iris-versicolor, 101-150 examples as iris-virginica) we would get the test set containing all the data points belonging to the same class during K-Fold. Another reason for which I used a seed is to shuffle the data to get the same desired accuracy everytime somebody runs the program. (If in case you want to change the seed, it can be changed by changing the value of "seedValue" variable on line 23 of linear\_regression.py file). Can also change the value of k (k-fold cross validation) by changing the variable (named ‘numberOfFolds’) on line 26 of the “linear\_regression.py” file.
2. For detailed explanation of the code, look into linear\_regression.py file