Machine Learning – ES 654

Spring 2019

IITGN

Here is the link for the secret gist containing all the answer codes for homework 5.

1st Ans:

(a) Code was provided in secret gist

(b) Code was provided in secret gist. Runtime of KNN is O(knd), where 'n' are the no.of training instances, 'd' are the no.of dimensions of dataset and 'k' are the no.of nearest neighbours that we consider. The plot of KNN runtime *vs* 'n' shows that runtime linearly increases with increase in 'n' which obeys the theoretical runtime. Whereas with increase in no.of dimensions of the doesn't much give a linear increase in KNN runtime

2nd Ans:

(a) Code was provided in secret gist

Error type (RMSE)	Train Error			Validation error			Test error		
Fold\Model	Ridge	Lasso	KNN	Ridge	Lasso	KNN	Ridge	Lasso	KNN
1	9.34	9.19	4.95	9.15	9.26	7.66	7.31	7.12	8.24
2	8.98	8.95	5.01	7.23	7.24	6.60	9.47	9.48	7.83
3	9.44	9.27	0.98	8.20	8.29	7.58	7.60	7.57	8.92
4	8.12	7.96	6.17	7.83	7.98	7.14	11.74	11.60	10.65
5	9.40	9.26	7.04	8.10	8.25	8.001	7.84	7.75	6.99

The above table depicts train, validation and test errors for 5-fold cross validation of real estate valuation dataset.

KNN performs better on train, validation and test set compared to Lasso and Ridge regression.

(b) Code was provided in secret gist. No all features are not on same scale. Yes scaling does impact KNN. For this dataset, it is observed that normalising dataset gives high error that the original dataset. Normalisation of data can increase or decrease the accuracy of the model (RMSE in case of regression) depending on the data samples. Sometimes normalisation may remove important feature differences resulting accuracy to go down. It can also happen that it reduces noise in data, which results in decrease of incorrect classifications and hence increases the accuracy of the model

(c) Code for plot is provided in secret gist. The train error kept on increasing with increase in K-value, whereas the test error kept decreasing till k=5 and it slightly increased for k>5. The train error increased because we know that with increase in ‘k’ we increase the bias in the model. From the test error we can see that optimal value for ‘k’ is 5.

(i) The maximum error value for a house is occurred when that data point is an outlier that is the distance to its nearest neighbours is too large compared to other data points. If data point is really an outlier then we can reduce the error value by feature scaling . If the point is not an outlier then increasing ‘k’ may decrease the error.

3rd Ans: Code for the Voronoi diagram is provided in the secret gist

‘Setosa’ flowers are forming a continuous region/cluster on the graph. Whereas the other ‘Versicolor’ and ‘Virginica’ are data points are having discontinuous clusters/regions on the graph (they regions/clusters are being mixed). Hence ‘Setosa’ test samples could be more correctly predicted than ‘Versicolor’ and ‘Virginica’ test points.