CS589 Machine Learning

Homework 1

Submitted by- Ravi Agrawal Due on- October 2nd, 2017

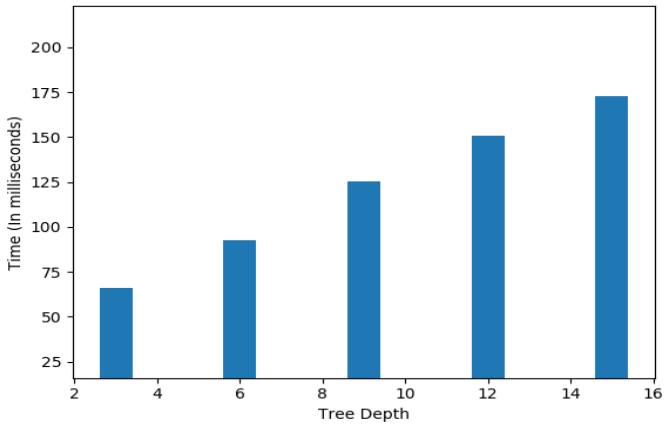
1 a: What is the criteria used to select a variable for a node when training a decision tree? Is it optimal? If yes, explain why it is optimal. If no, explain why is the optimal ordering not used?

Solution: - MAE is used as a criterion for selecting variable for a node when training a decision tree. MAE seems optimal choice here because as we saw in the class MAE is less sensitive to the outlier as compared to the MSE, In the same time I experimented using MSE and MAE in the decision Tree Classifier in the power plant data set and it turns out that the MAE outperforms the MSE almost all of the time. In general the use of MAE and MSE is user specified and depends upon the type of data and application we are working.

1 b: Sample error for models with depths [3, 6, 9, 12, 15] on the test set the model with tree depth 9 generates the MAE of 0.16175

Depth	MAE Error
3	0.21271945700778683
6	0.17242285381744882
9	0.1613646844452259
12	0.1612619518395852
15	0.16248072478924033

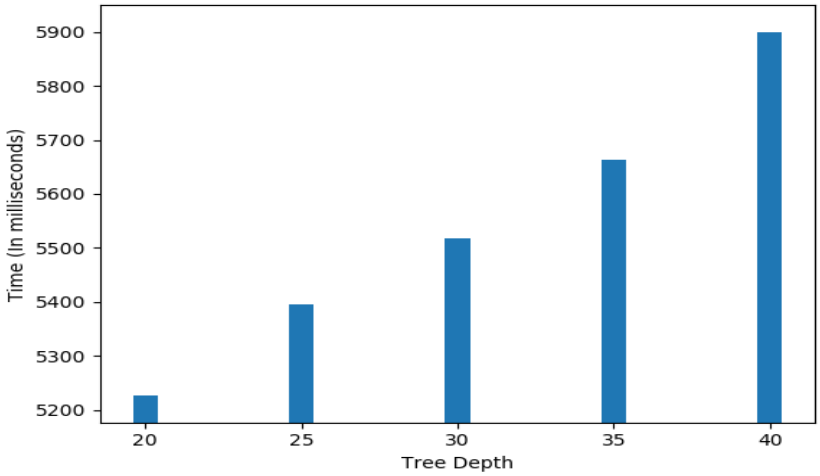
Time it took for each model to perform cross validation is shown below in the graph



1 c: Sample error for models with tree depth [20, 25, 30, 35, 40] the model with the tree depth 40 generates the MAE of 9.19 on the test data

Depth	MAE Error
40	4.0996470608394775
35	4.301234840287302
30	4.573045613472926
25	5.217632625792342
20	6.5688798038685

Time it took for each model to perform cross validation is shown below in the graph



2 a: Given, for training a model on a dataset with k samples takes K unit of time and there are N Samples.

Also, the we are splitting the data in the chunks of M samples so, that gives us

$$k\_fold = N / M$$

Lets try to define the time complexity assuming few constraint and then we can generalize the time complexity,

Lets Say,

$$N = 100$$

$$M = 20$$

$$K = N / M = 5$$

At a time we will train 4 folds of data that is 80 or $N - M$ and we will train 80 data points or $(N - M)$ data points for 5 times or N/M times

So,

$$Time\_complexity = 80\ x\ 5$$

$$Time\_complexity = (N - M) * (N/M)$$

The Big O time complexity of cross validation is **O (($\boldsymbol{N - M}$) * ($\boldsymbol{N/M}$))**.

For M = 5:- **O ((N - 5) * (N / 5))**

For M = N/2: - $O\ ((N - N/2) * (2 * N/N))$

$O\ (N * 2/2)$

The big O complexity is $O(N)$

2 b: As seen above the making the M smaller the run time complexity of cross validation algorithms also **decreases**. Run time of the algorithms decreases or algorithm runs faster.

3 a: Sample error for the model with following nearest neighbors [3, 5, 10, 20, 25].

Neighbors	MAE
3	0.5430270380266841
5	0.5468737841816506
10	0.5810720666919433
20	0.5955868431110324
25	0.5976941253154077

The predicted out of sample error are close to the real one. The Model with the 3 neighbors was chosen and the MAE during the cross validation was 0.543 and in the Kaggle competition was 0.6012.

3 b: Sample error for the model with thte following nearest neighbors [3, 5, 10, 20, 25] for the indoor localization is reported below, The MAE from the test set is 9.19482.

Neighbors	MAE
3	3.5612284272683206
5	3.8982777410010514
10	4.756855726154411
20	5.736928836139382
25	6.035753292231853

4 a: Regularization penalty improves the conditioning of the problem and reduces the variance of the estimates.

4 b: The Sample error for ridge and lasso error are reported below:

Alpha	MAE (RIDGE)	MAE(LASSO)
10	0.19073981300360843	0.6280389865258607
0.01	0.19079079225968892	0.19090267653339069

1e-06	0.1908257962383097	0.19077023543905794
1	0.190826129299028	0.2625607633489337
0.0001	0.19084015695555887	0.19090980542288613

Ridge: - On the training set the MAE least for the alpha 10 with 0.19073 for ridge regression on the test set this gives the MAE of 0.19289.

Lasso:- The Lasso Classifier gives the least MAE 0.19077 for alpha 10**(-6). The Test set predictions although gives the MAE of 0.19280.

The Lasso regression was chosen finally for the full model with alpha 10**-6 and generated test set prediction with MAE 0.19780

4 c :- The out of Sample error for the eight models are reported below:

ALPHA	MAE(RIDGE)	MAE(LASSO)
10	18.99768795124853	35.300970591757064
0.0001	18.999177487639212	19.010584344365935
1	19.017967626558782	21.006782322888323
0.01	19.02012180647829	18.953821828476062

The LASSO Model is selected to train on the full model with alpha equals 0.01, this model generated the cross validation error as 18.95 and predicted output generates the MAE equals 29.75645.

5 a :- I am training my model finally on the decision tree with hyper parameters tree depth in Range between [6, 9, 12]. I am moving ahead with using k = 5 in k fold cross validation, which gives, MAE as below:

Depth	MAE Error
6	0.17242285381744882
9	0.1613646844452259
12	0.1612619518395852

Looking at the MAE the decision tree with depth 9 or 12 seems reasonable option, so I tested out the decision tree with the depth 9 on the test data and depth 12 on the test data and depth 9 decision tree generates the least MAE among the two (i.e.: 0.15955). I also utilized other hyperparameter in the decision tree regressor like criterion as 'MAE' and presort as TRUE and they together generate minimum MAE equals 0.159.

5 b :- The indoor localization is trained on the K nearest neighbor model producing minimum MAE among other regression models. Again for this dataset I am using K in Kfold equals 5 and tested out the

model with different nearest neighbors [3, 5, 7, 10] and MAE of the K fold is also shown below in the table.

Neighbors	MAE
3	0.5430270380266841
5	0.5468737841816506
7	0.5675434565434562
10	0.5810720666919433

On test data the model with 7 nearest neighbors performs best. I also used other hyperparameters in the model like weights for "uniform" or "distance" and algorithm with options ['auto', 'ball_tree', 'kd_tree', 'brute']. And also parameter 'p' which have options ['1' manhattan_distance (l1), '2' euclidean_distance (l2)]. The final model KNN model with 7 neighbors, weights equals 'distance', algorithm equal "ball_tree" and p equal 1(manhattan_distance or l1) produced best result with MAE 8.163.