Technical University of Denmark

02450 Introduction to Machine Learning and Data Mining

Year 2019-2020

Group Project 2

PROJECT 2

Group Members

Name	Student Number
Altug Tosun	s181314
Mahsa Eskandarzadeh	s192933
Sukru Han Sahin	s192136

Contributions

Section	Author
1	Altug Tosun, Mahsa Eskandarzadeh, Sukru Han Sahin
2	Altug Tosun, Mahsa Eskandarzadeh, Sukru Han Sahin
3	Altug Tosun, Mahsa Eskandarzadeh, Sukru Han Sahin
4	Altug Tosun, Mahsa Eskandarzadeh, Sukru Han Sahin

1 Regression, Part A:

1 :

In LA Ozone data, there are 10 attributes that has been explained in the first report. Ozone attribute is selected as a response variable. Day of the year attribute is not used in analysis and the remaining 8 attributes selected as explanatory variables according to findings from first report. Since the day of the year is related with time series, it was set to date object in previous report. Therefore, the inspection with day of the year can be made by considering ozone attribute as a time series but it is not relevant for report so it will not be done. LA ozone data-set does not have any categorical attribute so one-of-K coding isn't applied. Nonetheless, since regularization will be used, all of the attributes in the data-set are standardized to have standard deviation 1 and mean 0. It is done by subtracting mean of each column from their column and dividing these columns to their standard deviation values. In addition to that, the column of ones added to matrix **X** in order to include the mean in the model.

2 :

In the regression the following model will be used :

$$y = Xw + \epsilon$$

Estimation of w will be made with L_2 regularization by using following:

Cost function :

$$E_\lambda(w) = ||y - Xw||^2 + \lambda w^T w$$

Solution :

$$w^* = (X^T X + \lambda I)^{-1} X^T y$$

Lambda values are chosen as the powers of 10 between 10^{-5} and 10^{9} in order to find the optimal lambda value. $K = 10$ cross validation is applied and the average of the validation error is considered as the estimation of generalization error. Figure 1 shows the change in generalization error with the lambda values. Figure 2 shows the change in weight array (w) with respect to lambda.

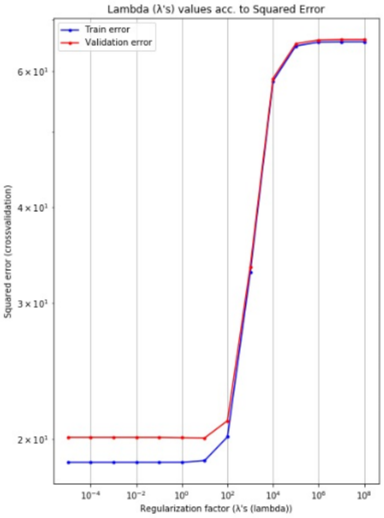


Figure 1: Lambda vs Squared Error(Generalization Error)

The values of the generalization error stays stationary until 10^1 and it sees the lowest value at there and it starts to increase afterwards. Hence, optimal lambda value for the model is 10^1.

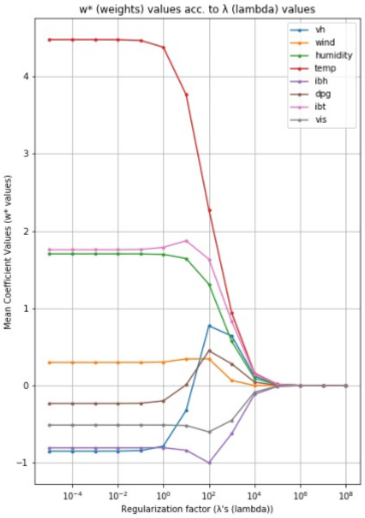
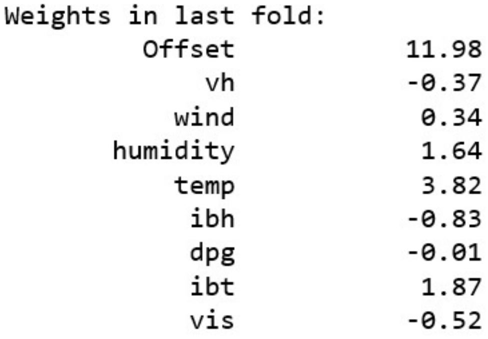


Figure 2: W values with respect to Lambda

At the high lambda values, the values of W is converging to zero. This is because the cost of having error become less than cost of increasing W. Therefore, we can say that the regularization works as expected.

3 :



```
Weights in last fold:
         Offset           11.98
             vh           -0.37
           wind            0.34
       humidity            1.64
           temp            3.82
            ibh           -0.83
            dpg           -0.01
            ibt            1.87
            vis           -0.52
```

Figure 3: W values when Lambda = 10

The prediction of new values are made by using the following equation: $y = Xw$ where w values are given in Figure 3.

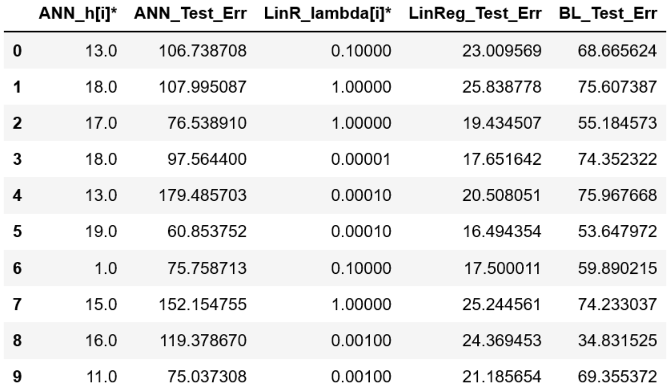
The coefficients of the attributes is an indicator of how important they are as the matrix $\mathbf{X}$ is regularized before the fit. As it can be seen from Figure 3, temp has the highest coefficient. This result is sensible since the effect of the temperature is generally high on atmospheric occasions. After temp attribute, ibt has the highest coefficient. This result is also sensible because ibt stands for inversion base temperature and it is expected to have a strong relationship with the ozone level according to meteorology. Humidity is also having strong relationship with ozone level since lower relative humidity corresponded to higher temperatures, higher solar radiation and higher ozone formation rates.

2 Regression: Part B

1 :

Linear regression, baseline model and ANN model were fitted to the data in order to get prediction on ozone values. These models fitted to data with the transformation of the the attributes to achieve a standard deviation of 1 and a mean of 0. In the light of this information, it can be seen seen that the baseline model is just an average of the training data. For the linear regression model, the cost function introduced to add a penalty directly related to the absolute values of weights multiplied by a complexity controlling variable. λ. For the ANN model, h is employed as a complexity controlling parameter. "h" is the number of hidden units in the hidden layer of the 3-layer model. 3-layer model consists of input layer, output layer and hidden layer. Generalization error of the model assessed by two-level cross validation method by using K1 and K2 as 10. Lambda values are chosen as the powers of 10 between 10^{-5} and 10^{9} as it fits the our needs after a few test-runs. Similarly, h values are chosen between 1 to 20.

2 :



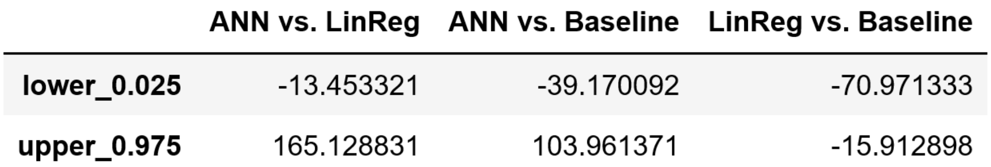
	ANN_h[i]*	ANN_Test_Err	LinR_lambda[i]*	LinReg_Test_Err	BL_Test_Err
0	13.0	106.738708	0.10000	23.009569	68.665624
1	18.0	107.995087	1.00000	25.838778	75.607387
2	17.0	76.538910	1.00000	19.434507	55.184573
3	18.0	97.564400	0.00001	17.651642	74.352322
4	13.0	179.485703	0.00010	20.508051	75.967668
5	19.0	60.853752	0.00010	16.494354	53.647972
6	1.0	75.758713	0.10000	17.500011	59.890215
7	15.0	152.154755	1.00000	25.244561	74.233037
8	16.0	119.378670	0.00100	24.369453	34.831525
9	11.0	75.037308	0.00100	21.185654	69.355372

Figure 4: Two Level Cross-Validation Results (K1 = K2 = 10)

As it is required, algorithm 6 from lecture slides are implemented. In the Figure 4, index of the table which is leftmost column indicates the number of the outer folds in order. There are mainly 5 columns in the figure. 3 of them are regarding the estimated generalization errors based on mean squared error metric for the models of ANN, Linear Regression and Baseline model. Generalization error of ANN column is called as `ANN_Test_Err` and others are respectively `LinReg_Test_Err` and `BL_Test_Err`. Since it is supposed to use the optimal values of lambda and hidden layers (h* and $\lambda*$) in the outer layer of the model implementations, for each outer level these values are depicted as well. Optimal h values are for ANN model and Optimal λ values are for Linear Regression model. At the each outer fold new models are trained according to these complexity controlling parameters (h* and $\lambda*$). Then the generalization errors for each trained model in the outer model are calculated. Of course, models trained with the optimal complexity parameters were evaluated on the each outer fold's test data. Columns which are yielding these optimal complexity parameters are `ANN_h[i]`(complexity-control parameter for ANN) and `LinR_lambda[i]`(complexity-control parameter for Linear Regression). Optimal λ values of the two layer cross validation for the Linear Regression model are very similar with the previous part **Regression A** since they are less than or equal to 1. But, our optimal λ value for Ridge Regression is 10^{1}.

3 :

Setup 2 has chosen for comparisons in this part (Method 11.4.1 is applied). Pairwise comparisons on errors are made to determine similarity between performance after the models are evaluated using two-layer cross-validation. Figure 5 shows 95% credibility intervals for the differences in error between the various models. Difference between ANN model vs Linear regression model, ANN model vs Baseline model, Linear regression model vs Baseline model depicted in first, second and third columns respectively.



	ANN vs. LinReg	ANN vs. Baseline	LinReg vs. Baseline
lower_0.025	-13.453321	-39.170092	-70.971333
upper_0.975	165.128831	103.961371	-15.912898

Figure 5: Pairwise comparisons with 95% credibility intervals

The intervals for the differences between ANN model and Baseline Model include the value 0 and it reveals that they have similarity. The intervals for the differences between ANN model and Linear Regression model includes zero as well. These models shows similarity in performance. On the other hand, It doesn't hold for Linear Regression and Baseline Model. Therefore, this model show dissimilarity in performance.

Even though ANN model vs Linear Regression includes 0 in the interval, it is at the end-side of the interval. Hence, it can be concluded that the similarity between these two is very small. To sum up, there exists performance differences in between all of these 3 models. By looking at Figure 4 and Figure 5, it can be said that the Linear regression model is preferable over the baseline model . For Figure 5, the baseline model might be preferable over the ANN model. Yet, certain conclusions can not be made from two-level cross validation results. Because cross validation methods aren't enough to compare these models itself. Therefore we have implemented Correlated t-test for cross validation. When we compare these two models in pairwise comparisons table at Figure 5, since "0" is in the interval and model performances are alike, therefore we can't conclude if ANN or Baseline is a better performer or not. Because "0" is not much at the end-side of the interval. For the ANN and Linear Regression comparison in Figure 5, lower value of -13 is close to to zero. From this comment and from Figure 4 as well, we can have a comment such as, Linear Regression "might be" a better performer than ANN.

3 Classification

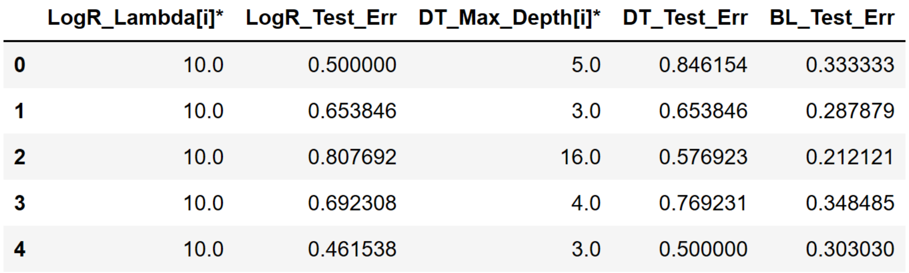
1 :

In LA Ozone data set there isn't any attribute for the classification problem. In Regression part, the main goal is to predict the values of Ozone attribute. Therefore, it has been decided to split Ozone values in value ranges and then create classes from these ranges. To do this split in a more statistical way, mean and standard deviation values of Ozone attribute is being used. Standard deviation is subtracted from mean and values which are less than 3 are classified as "low" class. Standard deviation is added to mean and values which are more 19 are classified as "high" class. And observations which doesn't belong to these classes are classified as "medium" class. Then our problem is being **multi-class** classification problem.

2 :

For the multi class classification problem (as it is defined above), *method 2* is being decided as to be Decision Tree algorithm. Therefore three algorithms are applied which are multi-nominal logistic regression, decision tree and baseline methods. Outcomes of those model are going to be compared. For baseline method there isn't any complexity controlling parameter. For the logistic regression; optimal lambda value ($\lambda *$), for decision tree; optimal maximum depth of tree parameters are going to be found out as complexity controlling parameter. Range of λ is each power of ten from 10^-5 to 10^9. Range of max. depth parameter is all integer values from 3 to 25. Again baseline method is being considered as the model that return the class with largest member in the test data. (Again a constant model.)

3 :



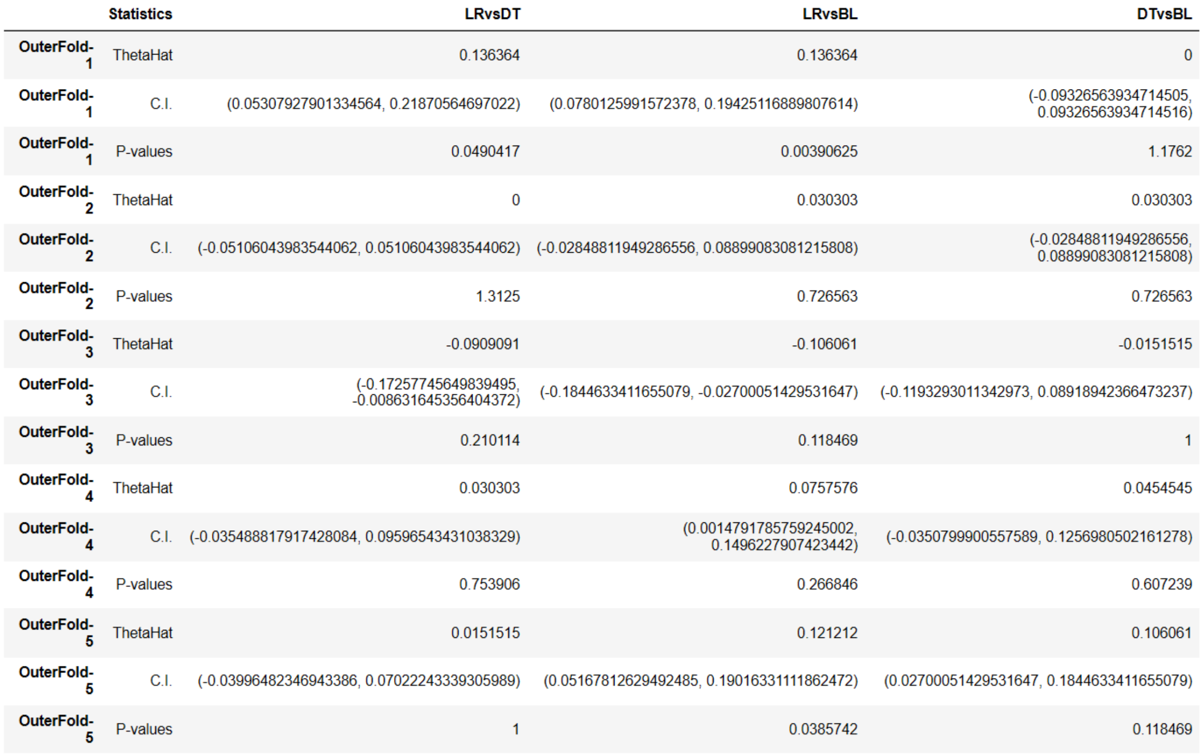
	LogR_Lambda[i]*	LogR_Test_Err	DT_Max_Depth[i]*	DT_Test_Err	BL_Test_Err
0	10.0	0.500000	5.0	0.846154	0.333333
1	10.0	0.653846	3.0	0.653846	0.287879
2	10.0	0.807692	16.0	0.576923	0.212121
3	10.0	0.692308	4.0	0.769231	0.348485
4	10.0	0.461538	3.0	0.500000	0.303030

Figure 6: Two Level Cross-Validation Results (K1 = K2 = 5)

Figure 6 is created in order to show results for two layer cross validation of three models. Table includes optimal values of complexity parameters ((λ) and max. depth) of Logistic Regression and Decision Tree models. Three models are abbreviated as; Logistic Regression as LogR, Decision Tree as DT, Baseline as BL in the table. `LogR_Test_Err`, `DT_Test_Err` and `BL_Test_Err` columns are the generalization errors from the models that are calculated with the optimal complexity parameters of models for each outer-fold. `DT_Max_Depth[i]` and `LogR_Lambda[i]` Optimal complexity parameters are found out from each inner fold calculation. These are the ones which yields the minimum generalization errors. Then, the calculation of the generalization error is made at the outer loop with the optimal complexity parameters by using E = (Number of misclassified observations) / N^{test}.

When the test errors are compared, baseline method is yielding less errors. Even if it seems in that way, we can't have comments about which algorithm is performing better than others since we will examine comparison of algorithms at the following step.

4 :



	Statistics	LRvsDT	LRvsBL	DTvsBL
OuterFold-1	ThetaHat	0.136364	0.136364	0
OuterFold-1	C.I.	(0.05307927901334564, 0.21870564697022)	(0.0780125991572378, 0.19425116889807614)	(-0.09326563934714505, 0.09326563934714516)
OuterFold-1	P-values	0.0490417	0.00390625	1.1762
OuterFold-2	ThetaHat	0	0.030303	0.030303
OuterFold-2	C.I.	(-0.05106043983544062, 0.05106043983544062)	(-0.02848811949286556, 0.08899083081215808)	(-0.02848811949286556, 0.08899083081215808)
OuterFold-2	P-values	1.3125	0.726563	0.726563
OuterFold-3	ThetaHat	-0.0909091	-0.106061	-0.0151515
OuterFold-3	C.I.	(-0.17257745649839495, -0.008631645356404372)	(-0.1844633411655079, -0.02700051429531647)	(-0.1193293011342973, 0.08918942366473237)
OuterFold-3	P-values	0.210114	0.118469	1
OuterFold-4	ThetaHat	0.030303	0.0757576	0.0454545
OuterFold-4	C.I.	(-0.035488817917428084, 0.09596543431038329)	(0.0014791785759245002, 0.1496227907423442)	(-0.0350799900557589, 0.1256980502161278)
OuterFold-4	P-values	0.753906	0.266846	0.607239
OuterFold-5	ThetaHat	0.0151515	0.121212	0.106061
OuterFold-5	C.I.	(-0.03996482346943386, 0.07022243339305989)	(0.05167812629492485, 0.19016331111862472)	(0.02700051429531647, 0.1844633411655079)
OuterFold-5	P-values	1	0.0385742	0.118469

Figure 7: McNemera's Test: Comparison of Models

Evaluation of the methods for classification problem is done regarding the **McNemar test**. According to Figure 7, it is pretty obvious leftmost column of the table indicates the outer-folds of the two cross validation. Then there is the column of "Statistics" which implies the values of "ThetaHat", "C.I." (which stands for Confidence Interval) and "P-values". There are the corresponding values of each statistical parameters for each outer-fold. Then there are the columns for comparison of model. LR, DT and BL respectively stands for Logistic Regression, Decision Tree and Baseline method. For example, LRvsDT means that Logistic Regression model is compared with Decision Tree model according to statistical parameters of ThetaHat, Confidence Interval and P-value for each outer-fold.

While p-value are useful to get an indication if one classier is better than another they are less useful for determining a plausible interval of their performance difference [1]. Therefore we are using Mcnemar test to understand if one classifier is better than other or not. If **Thetahat**$(\hat{\theta}) > 0$, then model A is preferable over model B.

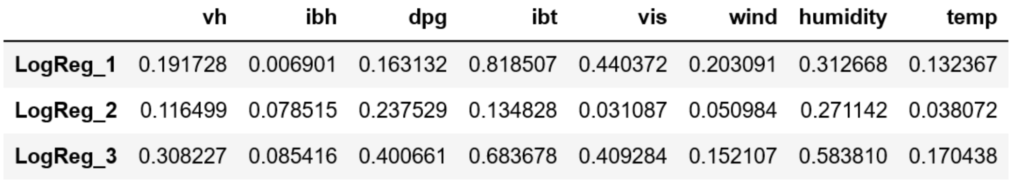
The interpretation is that the lower p is, the more evidence there is A is better than B, but only interpret the p-value together with the estimate $\hat{\theta}$ and ideally the confidence interval computed above [1]. If the p-value is less than our significance (alpha) level which is 0.05 in our evaluation, the hypothesis test is statistically significant.

According to information above, when we examine the column of Logistic Regression versus Decision Tree, first three outer-folds have $\hat{\theta}$ which are not greater than 1, last two outer-folds have really huge p-values like 1 and 0.75. Then we can say that, *Logistic Regression algoritm is not preferable over Decision Tree al-*

gorithm. When we examine column of Logistic Regression versus Baseline method, for the outer-folds of first two and last two, $\hat{\theta}$ values are greater than 0. For the p-values, first and last outer-folds are good p-values which are acceptable. Therefore we can say that *Logistic Regression algorithm is slightly preferable over baseline method.* When we examine the last column which is column of Decision Tree versus Baseline method, we see that out of five outer-folds we have 3 outer-folds which has $\hat{\theta}$ is greater than 0. For the p-values we has one p-value which tolerable. Therefore we can say that *Decision Tree algorithm is not preferable over baseline method.* Among three of the algorithms there isn't any algorithm which absolutely performs better than others. Even if the baseline method yielded less errors, we have seen that, it didn't perform better in terms of statistical comparison of models.

5 :

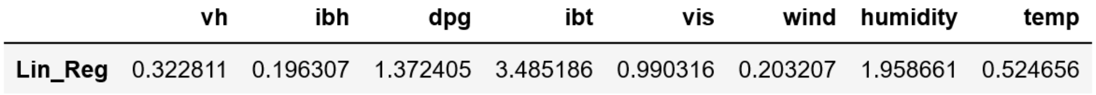
From the previous parts for linear regression $\lambda^* = 10^1$, for multinomial logistic regression $\lambda^* = 10^1$ are chosen. Then, according to these complexity parameters models are trained.



	vh	ibh	dpg	ibt	vis	wind	humidity	temp
LogReg_1	0.191728	0.006901	0.163132	0.818507	0.440372	0.203091	0.312668	0.132367
LogReg_2	0.116499	0.078515	0.237529	0.134828	0.031087	0.050984	0.271142	0.038072
LogReg_3	0.308227	0.085416	0.400661	0.683678	0.409284	0.152107	0.583810	0.170438

Figure 8: Logistic Regression Coefficients (weights)

When multinomial logistic regression table which is Figure 8 is examined, table depicts that there are 3 rows. After multinomial logistic regression model trained and resulting arrray has three different weight (w) vectors. Because multinomial regression fits a model for each class by considering that class as 1 and the other classes as 0. Then the weight vectors and the X matrix putted into *softmax function*. Purpose of *softmax function* is to determine the probability of each class for every observation. Then classes are labeled by using the probability values resulted from this function.



	vh	ibh	dpg	ibt	vis	wind	humidity	temp
Lin_Reg	0.322811	0.196307	1.372405	3.485186	0.990316	0.203207	1.958661	0.524656

Figure 9: Linear Regression Coefficients (weights)

When linear regression table which is Figure 9 is examined, table depicts that there is 1 row which shows weights (w) array. Since the data is standardized at the beginning, the coefficients is the importance of attributes' importance. (Since standardized ranges of attributes aren't effective. Therefore their pure importance is yielded for both linear and multi-nominal logistic regression algorithm.)
It is time to explain if the same features deemed relevant as for the regression part of the report. The effect of the attributes are independent from their signs because the amount of change is important to measure the effect not the direction. Thus, by using the absolute value of the coefficients the cosine similarity value between linear regression and each logistic regression calculated. Then, the average of the cosine similarity values are considered as the similarity between the effects of attributes between these two models. The cosine similarity values in scipy takes a value between 0 and 2 and the values closer to 0 means they are more similar. The resulted similarity value is 0.1088 which is very close to zero. This value implies that the effect of the coefficients are *very similar to each other and same features deemed relevant for the models of Linear Regression and Multinominal Logistic Regression.*

4 Discussion

1 :

The applications of neural networks ,linear regression and baseline model for predicting the concentration of ground-level ozone were studied and different results were obtained from our study. These results are :

- Regression of the data displayed an ability to predict ozone levels based on the attributes which reinforced by this idea that the linear regression model performed considerably better than the baseline model. Because by comparing the amount of error indicate that they are not statistically similar because 0 does not included in the 95 percent credibility interval created by taking the difference between the generalization errors. But comparing ANN with baseline is doubtful. And more or the less most probably Linear Regression is better performer than ANN.
- The best performance achieved on an outer fold by the best model–linear regression–still conceded a mean squared error greater than 16. This shows that the average prediction of ozone levels would be greater than 4 off of the actual value.
- The range of ozone values is [1, 38], So this error value of 4 is not trivial.
- According to Figure 7, the logistic regression model and the decision tree models don't generalize in a way that is statistically significantly different than the baseline model. We have concluded in that way because of thetahat and p value results do not show any significant preference over models.
- There is a direct mapping from regression of a continuous value into categories so it expected performance is also similar to the regression models. On other hand, this classification can be seen as placing a sample into one of n categories based on its expected ozone value, but there would be no reason to expect accurate classification when accurate regression cannot be possible.
- In our data set sample size is small. There are only 330 observations. This might be a reason why there isn't a preferred model in classification. Another reason might be our split (low, medium, high) of classes which again depends on the number of observation.

2 :

There are many researches on LA Ozone Data set during the 1970s and 1980s, but they concluded that the error rates were too high for use [2]. A potential reason for this result is that the ozone values depend on more attributes than what were recorded when the data-set was being created. In addition to that, our analysis supports this prior work on LA ozone data set because the effort of performing an accurate regression or classification on the data seems infeasible for LA ozone data-set. There is not enough correlation between target values and the attributes.

5 References

[1] Tue Herlau, Mikkel N. Schmidt and Morten Morup, ‘Introduction to Machine Learning and Data Mining’,
September 24, 2019.
[2] L. Breiman, Statistical Modeling: The Two Cultures, 2001. http://staff.pubhealth.ku.dk/ tag/Teaching/share/material/Bre
two-cultures.pdf. Accessed 10 November 2019.