Report on MSE Vs MAE

group 27

21 April 2018

Mean Squared Error(MSE)

The mean squared error (MSE) of an estimator measures the average of the squares of the errors that is, the difference between the actuals and predicted values.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

Where, $\hat{Y}_i$ is a vector of n predictions and Y is the vector of observed values of the variable being predicted.

Mean Absolute Error(MAE)

The mean absolute error(MAE) measures the mean of the absolute errors that is, the absolute value of the difference between the forecasted value and the actual value. MAE tells us how big of an error we can expect from the forecast on average.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|Y_i - \hat{Y}_i|$$

Where, $\hat{Y}_i$ is a vector of n forcasts and Y is the vector of actual values of the variable being predicted.

MSE Vs MAE

Mean squared error has the disadvantage of heavily weighting outliers. It is a result of the squaring of each term, which effectively weights large errors more heavily than small ones. Where this kind of property is undesirable, MAE can be used in those applications by the researcher.

When dealing with outliers, it might be helpful to use MAE instead of MSE since MSE gives higher error than MAE. Yet, MSE is more popular and efficient than MAE, because MSE punishes larger errors, which tends to be useful in the real world.

The mean absolute error (MAE) has the same unit as the original data, and it can only be compared between models whose errors are measured in the same units.

Both MSE and MAE are scale-dependent.For instance, if the observed data are in km then MSE is in km^2 and MAE is always in km respectively. Often, we need to perform accuray test on predicted values across different units. In that particular context, both MSE and MAE will not be applicable because they can only be compared between models whose errors are measured in the same units.

For evenly distributed errors that is, when all of the errors have the same magnitude, then Root mean squared error(RMSE) and Mean absolute error(MAE) will give the same result.If the square of the difference between actual values and forcasted values gives a positive distance which is same as their absolute distance then, MSE = MAE.

Data collection and exploration

To calculate MSE and MAE of different regression methods we used the *Energy_efficiency.csv* dataset.This dataset has been collected from the UCI MAchine Learning *Repository*[3]. This dataset is a collection of 768 samples and 8 features, aiming to predict two real valued responses.

The dataset contains the following eight attributes or features(X1,X2,....X8) along with two response variables(Y1,Y2):

- Relative Compactness(X1)
- Surface Area(X2)
- Wall Area(X3)
- Roof Area(X4)
- Overall Height(X5)
- Orientation(X6)
- Glazing Area(X7)
- Glazing Area Distribution(X8)
- Heating Load(Y1)
- Cooling Load(Y2)

It is important to implement energy efficiency in building to mitigate the impact of climate change. Due to the high demand for energy and unsustainable supplies, energy efficiency in building plays a vital role reducing energy costs and greenhouse gas emissions. Therefore, studying this dataset to evaluate how well energy is being used there to cut out the costs which will be helpful to have a ECO-friendly environment.

Experiment and perform evaluation

We load the samples into a dataframe and took all the column attributes as factor. We randomize the data frame using . Then, we divided the dataset into a trained dataset with the top 80% of the samples, and a tested dataset with the bottom 20% of the samples respectively. So, energy train data has first *614* entries from the dataset and energy test data contains the rest *154* samples.

At first we set up a model*(rt1)* for tree regression using the *Heating.Load* as outcome variable and all the eight attribtes as input variables and fit a new dataframe with the actual and predicted value of the model based on the test data. Using Regression Tree model(rt1) and "Heating.Load" as outcome, we calculated MSE = 6.59 and MAE = 2.101.

Similarly we fit another model(rt2) for tree regression but instead of using *Heating.Load* as outcome variable now we are intersted to use *Cooling.Load* as outcome variable. And we figured out for this model(rt2), using *Cooling.Load* we got MSE = 8.461 and MAE = 2.084.

We randomize the data frame using again. Next up we fit two models namely rf1 and rf2 respectively for both *Heating.Load* and *Cooling.Load* as outcome variables using Random forest regression following the same approach as described earlier for rt1 and rt2. Then we measured the MSE and MAE and for rf1 we got, MSE = 1.36 and MAE = 0.907.

```
##                            % Inc MSE
## Glazing.Area                    74.7
## Glazing.Area.Distribution       41.0
```

```
## Relative.Compactness           24.7
## Surface.Area                   24.1
## Wall.Area                      20.8
## Roof.Area                      18.9
## Overall.Height                 17.7
## Orientation                   -19.6
```

Observing the result of *importance()* function to calculate the importance of each variable, we got to see that *Glazing.Area* was considered the most important predictor; it is estimated that, in the absence of that variable, the error would increase by 74.7%.

Whereas for model rf2, using *Cooling.Load* we got MSE = 3.698 and MAE = 1.348.

```
##                            % Inc MSE
## Glazing.Area                   75.51
## Glazing.Area.Distribution      29.36
## Relative.Compactness           24.03
## Surface.Area                   22.02
## Roof.Area                      21.45
## Wall.Area                      20.14
## Overall.Height                 18.78
## Orientation                    -4.36
```

If we look into the *importance()* function to calculate the importance of each variable, we can see that The *Glazing.Area* was considered the most important predictor for *rf2*. it is estimated that, in the absence of that variable, the error would increase by 75.51%.

If we perform overall evaluation and compare MSE and MAE for all four models we can see that using random forest regression the model, rf1 with *Heating.Load* as response variable has lower error rate for both MSE = 1.36 and MAE = 0.907 compared to other models. For regression tree model both rt1 and rt2 produced relatively higher MSE values though MAE values did not varry significantly.

References

1. https://en.wikipedia.org/wiki/Mean_squared_error
2. https://en.wikipedia.org/wiki/Mean_absolute_error
3. http://archive.ics.uci.edu/ml/datasets/Energy+efficiency#
4. A. Tsanas, A. Xifara: 'Accurate quantitative estimation of energy performance of residential buildings using statistical machine learning tools', Energy and Buildings, Vol. 49, pp. 560-567, 2012 (the paper can be accessed from weblink)