Machine Learning Worksheet-1 (Internship-23, Konatala Mohit ID-34)

1. Which of the following methods do we use to find the best fit line for data in Linear Regression?

a) Least Square Error

2. Which of the following statement is true about outliers in linear regression?

b) Linear regression is sensitive to outliers

3. A line falls from left to right if a slope is _____?

b) Negative

4. Which of the following will have symmetric relation between dependent variable and independent variable?

b) Correlation

5. Which of the following is the reason for over fitting condition?

c) Low bias and high variance

6. If output involves label then that model is called as

d) All of the above

7. Lasso and Ridge regression techniques belong to _____?

d) Regularization

8. To overcome with imbalance dataset which technique can be used?

d) SMOTE

9. The AUC Receiver Operator Characteristic (AUCROC) curve is an evaluation metric for binary classification problems. It uses _____ to make graph?

c) Sensitivity and Specificity

10. In AUC Receiver Operator Characteristic (AUCROC) curve for the better model area under the curve should be less

b) False

11. Pick the feature extraction from below:

a) Construction bag of words from an email

12. Which of the following is true about Normal Equation used to compute the coefficient of the Linear Regression?

a) We don't have to choose the learning rate.

b) It becomes slow when number of features is very large.

13. Explain the term regularization?

Overfitting is a common issue faced in the field of machine learning, one the techniques that addresses this issue is regularization.

Regularization is a technique used to avoid over-fitting i.e. by adding penalty term to the cost function on the number of parameters in the given model. The regularization function measures the complexity of the hypotheses. Regularization limits us from learn a more flexible model i.e. to avoid over-fitting.

Most often in regression we face the increased complexity in the model which represented by the coefficients (calculated from the regression line). Regularization helps in reducing the scale of the explanatory variables by maintaining the equivalent number of variables as it maintains accuracy and generalization of the model.

Working of Regularization:

It works by adding penalty term with residual sum of squares to the intricate model.

Let us take an example of simple linear regression

$$\boldsymbol{Y = \beta_0 + \sum_{i=1}^{n} \beta_i X_i}$$

Here, Y represents the dependent variable, $X_1, X_2, X_3, \ldots . X_i$ are the predictor of the dependent variable and $\beta_0, \beta_1, \beta_2, \ldots . \beta_i$ are coefficients of the predictors which describes weights/scale attached to the features

In linear regression, residual sum of squares (RSS) is the optimization function.

We picked the set of coefficients, such that the following optimization function is minimized:

$$\boldsymbol{RSS} = \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \left(\boldsymbol{y_i} - \boldsymbol{\beta_0} - \sum_{\boldsymbol{j=1}}^{\boldsymbol{p}} \boldsymbol{\beta_j x_{ij}} \right)^{\boldsymbol{2}}$$

The above equation is the cost function for simple linear regression.

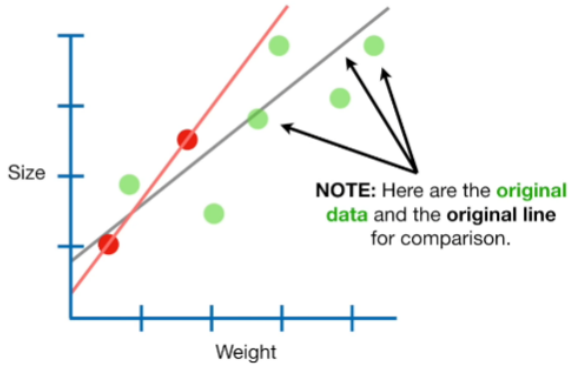
This function will adjust the coefficient estimates based on the training data as noise in the training data effects the predict data but regularization shrinks the estimates in the direction of zero i.e. by adding a loss function that regularizes the parameters to make a model that can predict the accurate value of the Y.

14. Which particular algorithms are used for regularization?

In machine learning there are two particular algorithms used for regularization:

- Ridge Regression (L2 norm)
- Lasso Regression (L1 norm)

Ridge Regression (L2 norm): In machine learning when we use linear regression to model relationship between two variables (dependent and independent) i.e. using least square to fit a line to the data that has minimum sum of squared residual this works when the collected is large (data size is proportional to the best fit), but when the training data is too small (of a particular model) then residual value is minimum and when the same model is compared with larger quantity of data the initial best fit(1st line) has high variance i.e. the Line is over-fitting.



The above graph represents size and weight of the male. The green line indicate the best fit (for set of green values which more in quantity), whereas the red line is same model with less data sample due to which there is a high variance (over-fitting). Ridge regression helps in find a new line that doesn't fit the training data that by inducing a small bias which helps in lowering the variance value. This helps in better prediction of the dependent variable in long term.

Working of Ridge regression:

$$\boldsymbol{RSS} = \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \left(\boldsymbol{y_i} - \boldsymbol{\beta_0} - \sum_{\boldsymbol{j=1}}^{\boldsymbol{p}} \boldsymbol{\beta_j x_{ij}} \right)^{\boldsymbol{2}}$$

The above cost function is altered by adding a shrinkage expression i.e. it minimizes the sum of the residual and λ x slope^2 where the square of slope is penalty and λ determines the severity of the penalty.

$$\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \left(\boldsymbol{y_i} - \boldsymbol{\beta_0} - \sum_{\boldsymbol{j=1}}^{\boldsymbol{p}} \boldsymbol{\beta_j x_{ij}} \right)^{\boldsymbol{2}} + \boldsymbol{\lambda} \sum_{\boldsymbol{j=1}}^{\boldsymbol{p}} \boldsymbol{\beta_j^2}$$

This helps in reducing the complexity of the model.

Ridge regression are used when the independent variable have high collinearity also when there are more parameters compared to the sample data-set.

Though ridge regression doesn't help in feature selection because even after minimizing the complexity of the model the coefficient is never leads to zero for the given set of independent variable, also its model interpretability its good as even after shrinkage of the least effective predictor it still has its value in the final model.

Lasso Regression (L1 norm): This regression is same as the L2 norm the difference being that penalty term is absolutely instead of square.

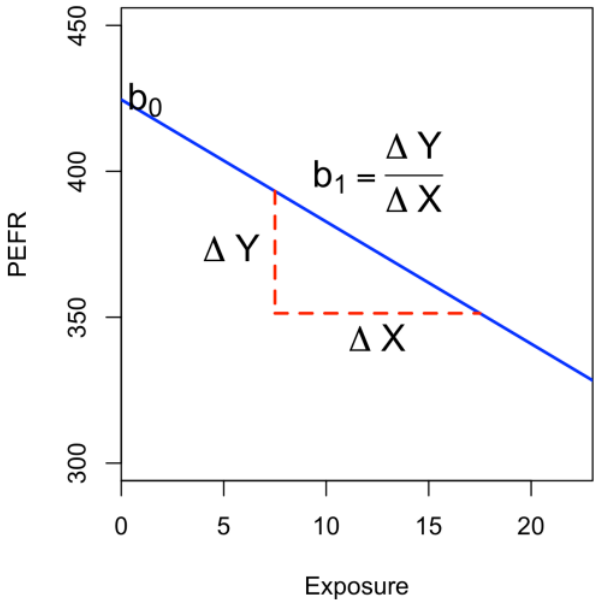
$$\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \left(\boldsymbol{y_i} - \boldsymbol{\beta_0} - \sum_{\boldsymbol{j=1}}^{\boldsymbol{p}} \boldsymbol{\beta_j x_{ij}} \right)^{\boldsymbol{2}} + \boldsymbol{\lambda} \sum_{\boldsymbol{j=1}}^{\boldsymbol{p}} \left| \boldsymbol{\beta_j} \right|$$

Lasso regression overcome the a drawback of L2 norm i.e. it at times forces few coefficients of least effective predictors equals to zero i.e. it completely negotiates some features for model while alteration of λ is adequately large due to which it can be used for features selection

15. Explain the term error present in linear regression equation?

Simple Linear Regression can help in finding the relation between the two variables i.e. finding the best line to the predict the response (PEFR) which is a function of predictor variable (Exposure)

$$PEFR = b_0 + b_1(Exposure) \quad \text{- I}$$



The above linear graph is the best line achieved using equation-I

The important concepts in regression are fitted values (prediction value) and the **residual (prediction error).** In general the data point doesn't exactly fall on the best line. So as *to correct the equation we include explicit error* $\boldsymbol{e_i}$.

$$\boldsymbol{Y_i = b_0 + b_1 X_i + e_i} \text{ - II}$$

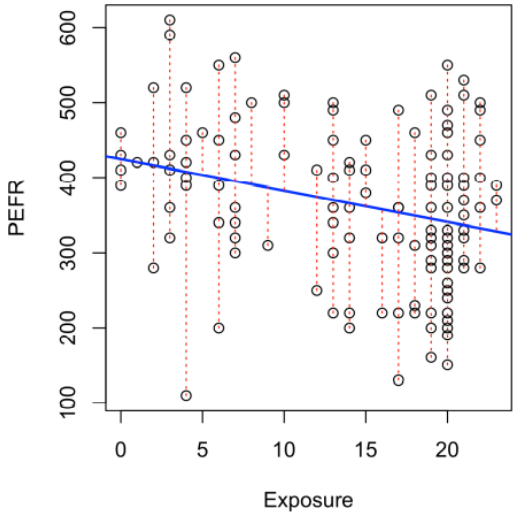
The term ***error*** *in the linear regression indicates the error in predictor which are defined using the relation with independent variables.*

The fitted values (predicted values) is usually denoted with $\hat{Y}$

$$\boldsymbol{\hat{Y}_i = \hat{b}_0 + \hat{b}_1 X_i} \text{ - III}$$

The notations $\hat{b}_0$ and $\hat{b}_1$ indicate coefficients of estimated and known respectively

We can obtain the residuals $\hat{e_i}$ by substituting equation II equals to equation III



In the above graphical figure the lengths of the vertical red dashed lines are indicating the residual.

To determine if the model is best fit to the data we use Least squares i.e. the regression line is the estimate that minimizes the sum of square residual values

$$\textbf{RSS= } \sum_{i=1}^{n}\left(Y_i - \widehat{Y}_i\right)^2$$

$$\textbf{RSS=} \sum_{i=1}^{n}\left(Y_i - \widehat{b}_0 - \widehat{b}_1 X_i\right)^2$$

The estimate $\hat{b}_0$ and $\hat{b}_1$are the values that minimize RSS.