

MACHINE LEARNING**WORK SHEET SET-1**

In Q1 to Q11, only one option is correct, choose the correct option

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

- A) Least Square Error
- B) Maximum Likelihood
- C) Logarithmic Loss
- D) Both A and B

Ans: - A) Least Square Error

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

- A) Linear regression is sensitive to outliers
- B) linear regression is not sensitive to outliers
- C) Can't say
- D) none of these

Ans: - A) Linear regression is sensitive to outliers

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

- A) Positive
- B) Negative
- C) Zero
- D) Undefined

Ans: - B) Negative

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

- A) Regression
- B) Correlation
- C) Both of them
- D) None of these

Ans: - B) Correlation

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

- A) High bias and high variance
- B) Low bias and low variance
- C) Low bias and high variance
- D) none of these

Ans: - C) Low bias and high variance

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

- A) Descriptive model
- B) Predictive modal
- C) Reinforcement learning
- D) All of the above

Ans: - B) Predictive modal

7. Lasso and Ridge regression techniques belong to _____?

- A) Cross validation
- B) Removing outliers
- C) SMOTE
- D) Regularization

Ans: - D) Regularization

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

- A) Cross validation
- B) Regularization
- C) Kernel
- D) SMOTE

Ans: - D) SMOTE

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

- A) TPR and FPR
- B) Sensitivity and precision
- C) Sensitivity and Specificity
- D) Recall and precision

Ans: - A) TPR and FPR

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

A) True

B) False

Ans: - B) False

11. Pick the feature extraction from below:

A) Construction bag of words from a email

B) Apply PCA to project high dimensional data

C) Removing stop words

D) Forward selection

Ans: - B) Apply PCA to project high dimensional data

In Q12, more than one options are correct, choose all the correct options

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.

C) We need to iterate.

D) It does not make use of dependent variable

Ans: - A) We don't have to choose the learning rate.

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

Q13 AND Q15 ARE SUBJECTIVE ANSWER TYPE QUESTIONS, ANSWER THEM BRIEFLY.**13. Explain the term regularization?**

Ans: -

First have to know that what is Regularization?

>>Regularization techniques that are used to calibrate machine learning models in order to minimize the adjusted loss function and prevent overfitting or underfitting

>>Regularization is a technique used to reduce the error by fitting the function appropriately on the given training set data and avoid overfitting the commonly used to regularization techniques

Terms: -

>>When a machine learning model, models Training data well fails to perform well one the testing data i.e., was not able to predict test data it is called as overfitting and this situation can be dealt with regularization in machine learning

>>Regularization adds to a regularization term in order to prevent the coefficients to fit so perfectly to overfit

Example: - Home sales analysis

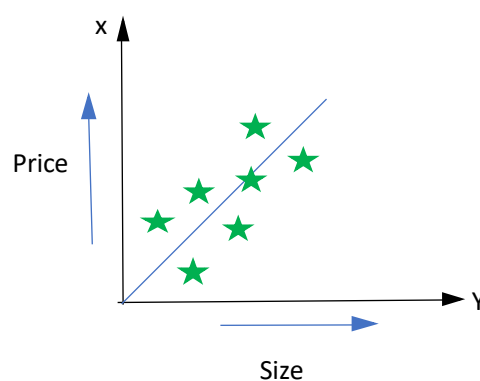



Fig (a)

The fig shows “under fitting “

Here,

Train error 

Test error ↑

Linear Regression

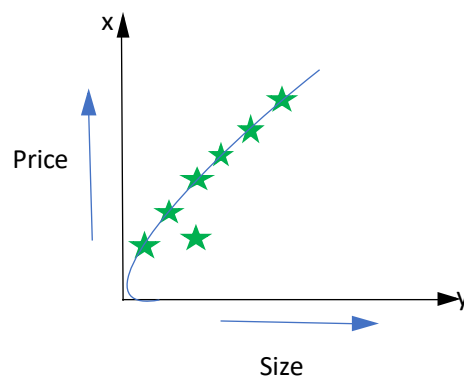


Fig (b)

The above fig shows best polyneme of degree

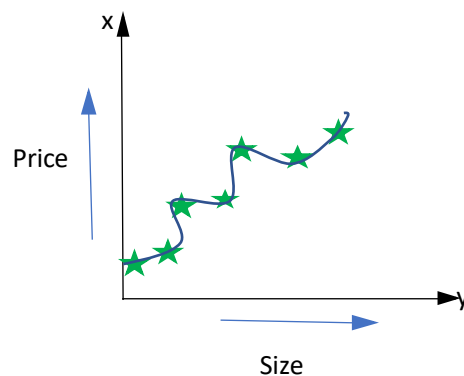


Fig (c)

The above fig shows “over fitting”

Under fitting: -

The machine learning, we have underfitting problem as shows fig(a) we cant get good out for training data and test data and also its high bias

The equation we write for under fitting

$$\theta + \theta_1 X$$

The machine learning, we have under fitting problem we can't get good output for training data test data

Linear regression: -

The above fig(b) best fit polynome degree and the equation we write for linear regression

$$\theta + \theta_1 X + \theta_2 X^2$$

Over fitting: -

The above fig© shows over fitting curve Here training data set is cos function is near about 0 so this is fails to training set data output

the equation we write for linear regression

$$\theta + \theta_1 X + \theta_2 X^2 + \theta_3 X^3 + \theta_4 X^4 + \theta_5 X^5 + \theta_6 X^6$$

$$J(\theta) = 0.$$

Fails to generates new examples due to high failure

Here the overfitting in machine learning is helps to bring good training data predict but test data output we can't get good predict output

To reduces the outfitting

>>Reduces numbers of features, manually or by model solution algorithms using some important features may thrown out

>>Regularization keeps all the features but reduces magnitude or values of parameter θ , Regularization works well when we have a lot of features and each of which contributes a bit to predicting while that means target out

>>Regularization regularizes or shrinks the coefficient estimates towards to zero

>>Regularization are techniques used to reduces error by fitting a function appropriately on the given training set on avoid the over fitting

>>Regularization is a technique used for tuning the machine learning model by adding on additional penalty term in the error function

>>the additional term (penalty)controls the excessively fluctuating function such that coefficient don't take extreme values

>>the algorithms on the right is fitting the noise in above example a way to reduces overfitting is turn to artificially penalize higher degree polynomial this ensures that a higher degree polynomial is suited only it reduces the error significantly compared to simple model to over come to

14. Which particular algorithms are used for regularization?

Ans: -

There are three types

1)Lasso regularization (least absolute shrinkage and selection operator)

2)Ridge regularization

3)dropout

Lasso regularization (least absolute shrinkage and selection operator)

>>it adds a penalty to the error function and the penalty is the sum of the absolute values of weights

$$\text{Min } (\sum_{i=1}^n (Y_i - W_i X_i)^2 + P \sum_{i=1}^n |W_i|)$$

Where P is tuning parameter which decides how much we want to penalize the model

$$\text{cost function} = (\sum_{i=1}^n (Y_i - \hat{Y}_i)^2 + \alpha P \sum_{i=1}^n |B_i|)$$

$$\text{where } (Y_i - \hat{Y}_i)^2 = (\text{actual output} - \text{predicted output})^2$$

$$Y_i = B_i + B_1 X_{i1} + \dots + B_{10} X_{i10} \text{ here only using coefficient}$$

>>Lasso shrinks the less important features coefficient to zero the removing some features all together

>>Lasso adds absolute values of magnitude of coefficients as penalty term to the loss function

>>Lasso works well for features selection in case we have a large number of features

>>cross validation works well with a small set of features but for large set of features these techniques beneficial

>>So only a subset of the most important features are left with non-zero weights it also makes the model easier to interpret

>> As the value of p/α increases it reduces the value of coefficients hence avoiding overfitting but after certain value the model starts losing important properties of the data (under fit problem) therefore the value of p/α should carefully selected by using some hyper parameter tuning techniques

Ridge regularization

>>Its also adds a penalty to the error function but in Ridge regularization the penalty is the sum of the squared values of weights

The equation of ridge regularization

$$\text{Min } (\sum_{i=1}^n (Y_i - W_i X_i)^2 + P \sum_{i=1}^n |W_i|^2)$$

Where P is the tuning parameter which decide how much we want to penalize the model

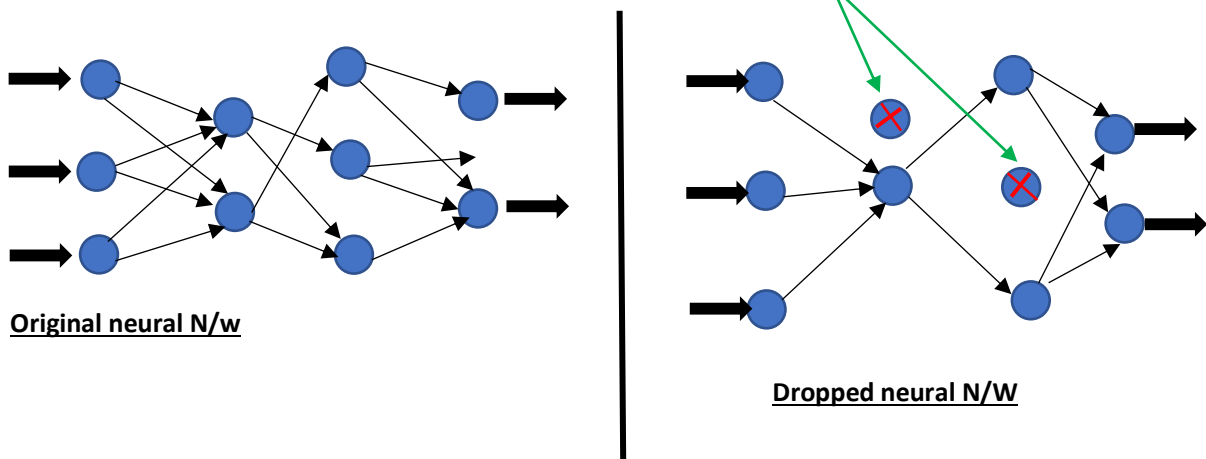
>>Ridge regression adds squared magnitude of coefficient as penalty term to the loss function

$$\text{cost function} = (\sum_{i=1}^n (\hat{Y}_i - Y_i)^2 + \alpha \sum_{i=1}^n B_i)^2$$

>>If $\alpha=0$ -----no penalty = regression without ----regularization

>>Ridge regression enforces the β coefficient to be lower but it does not enforce then to be zero that it will not remove irrelevant features but rather minimize their impact on the trained model

Drop out



>>Randomly drop units(neurons) along with their connections from the neural network during training

>>Drop out is a technique in which some models of the network are temporarily deactivated these techniques is applied in the training phase to reduces overfitting effects

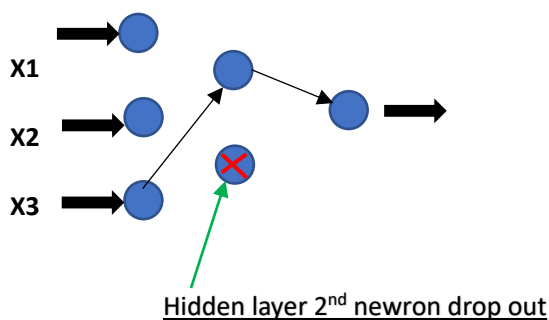
>>The basic idea behind dropout neural network to dropout nodes so that the network can concentrate on other features

>>In dropout approach we randomly choose a certain number of nodes from the input and the hidden layers which remain active and turn off the other nodes of these layers, after this we can train a part of our learn set with this network

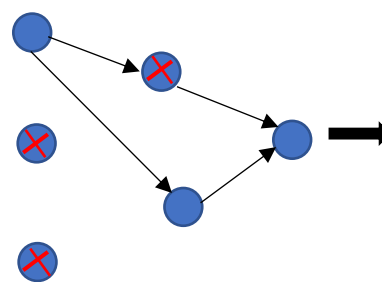
>>The next step consists in activating all the nodes again and randomly chose other nodes

>>It is also possible to train the whole training set with the randomly created dropout networks

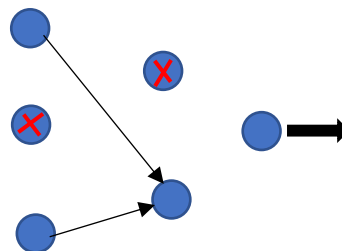
Drop out network 1st approach



Drop out network 2st approach



Drop out network 3rd approach



>>At each training stage individual nodes are either dropped out of the network with probability $1-P$ or kept with probability P , so that a reduced network is left incoming and outgoing edges to a dropped -output nodes are also removed ($p=0.5$ works were)

>>Dropout is an approach to regularization in neural network which helps reducing interdependent learning amongst the neurons

>>Using dropout approach the network become less sensitive to the specific weights of neurons this is capable of better generalization and is less likely to overfit the training data

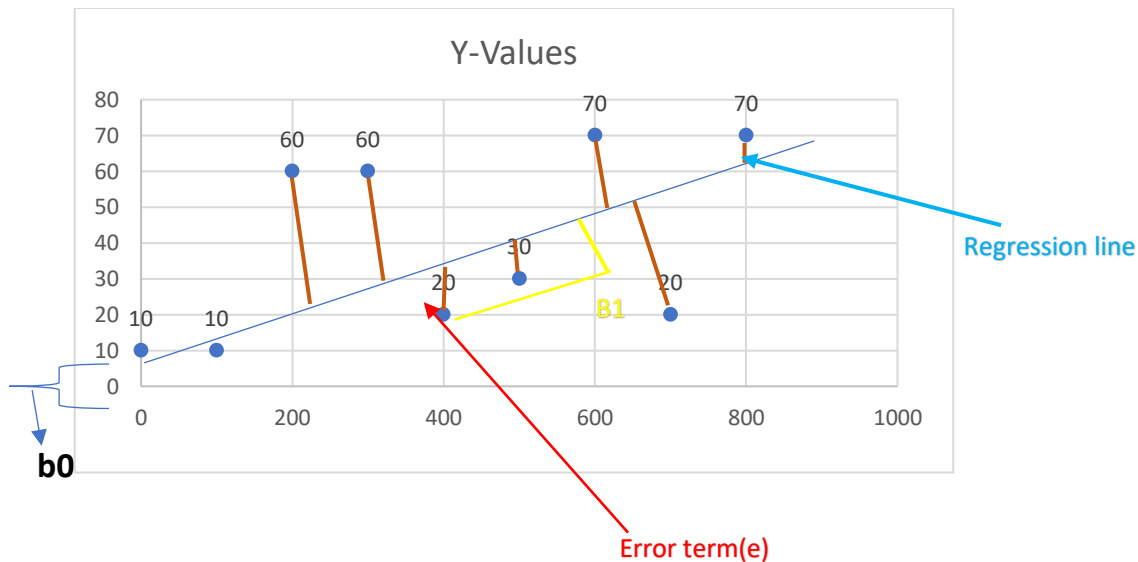
>>Dropout roughly doubles number of iterations required to conveys however training time for each approach is less

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

Ans: -

A linear regression model's main aim is to find the best fit linear line and the optimal values of intercept and coefficients such that the error is minimized

>> Error is the difference between the actual value and predicted value and the goal is to reduce this difference



In the above diagram,

>>x is our dependent variable which is plotted on the x-axis and y is the dependent variable which is plotted on the Y-axis

>>Blue dots are the data points i.e., the actual values

>>b0 is the is the intercept which is 10 and b1 is the slope of the slope of the x variable

>>The blue line is the best fit line predicated by the model i.e. the predicted values lie on the blue line

>>The vertical distance between the data point and the regression line is known as error or residual each data point has one residual and the sum of all the differences is known is the sum of Residuals/Errors

Mathematical Approach

>> Residual/Error = actual values -predicted values

Sum of Residuals/Errors= sum (Actual-predicted values)

Square of sum of Residuals/Errors=(sum(Actual-predicted values))^2

i.e.

$$\sum e_i^2 = \sum (Y_i - \hat{Y}_i)^2$$

For the in-depth understanding of the maths behind linear regression

Assumptions of Linear Regression

The basic assumptions of Linear Regression are as follows:

Linearity: It states that the dependent variable Y should be linearly related to independent variables. This assumption can be checked by plotting a scatter plot between both variables.

Normality: The X and Y variables should be normally distributed. Histograms, KDE plots, Q-Q plots can be used to check the Normality assumption

Homoscedasticity: The variance of the error terms should be constant i.e the spread of residuals should be constant for all values of X. This assumption can be checked by plotting a residual plot. If the assumption is violated then the points will form a funnel shape otherwise, they will be constant.

Independence/No Multicollinearity: The variables should be independent of each other i.e. no correlation should be there between the independent variables. To check the assumption, we can use a correlation matrix or VIF score. If the VIF score is greater than 5 then the variables are highly correlated.

The error terms should be normally distributed. Q-Q plots and Histograms can be used to check the distribution of error terms.

The error terms should be normally distributed. Q-Q plots and Histograms can be used to check the distribution of error terms.

In that linear regression there is violation off normality assumption of variables or error terms

>> To treat this problem, we can transform the variables to the normal distribution using various transformation functions such as log transformation, Reciprocal, or Box-Cox Transformation

How to deal with the Violation of any of the Assumption

>>The Violation of the assumptions leads to a decrease in the accuracy of the model therefore the predictions are not accurate and error is also high.

>>For example, if the Independence assumption is violated then the relationship between the independent and dependent variable cannot be determined precisely.

>>There are various methods are techniques available to deal with the violation of the assumptions. Let's discuss some of them below.

Violation of Normality assumption of variables or error terms

>>To treat this problem, we can transform the variables to the normal distribution using various transformation functions such as log transformation, Reciprocal, or Box-Cox Transformation.

All the functions are discussed in this article of mine: [How to transform into Normal Distribution](#)

Violation of Multicollinearity Assumption

It can be dealt with by:

>>Doing nothing (if there is no major difference in the accuracy) Removing some of the highly correlated independent variables. Deriving a new feature by linearly combining the independent variables, such as adding them together or performing some mathematical operation.

Performing an analysis designed for highly correlated variables, such as principal components analysis.