ASSIGNMENT – 39
MACHINE LEARNING

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

Answer:) 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

Answer: 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

Answer: 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

Answer: A) Regression

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

Answer: 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

B) Predictive modal

7. Lasso and Ridge regression techniques belong to _____?
A) Cross validation B) Removing outliers
C) SMOTE **D) Regularization**

Answer: D) Regularization

8. To overcome with imbalance dataset which technique can be used?
A) Cross validation B) Regularization
C) Kernel D) SMOTE

Answer:A) Cross validation

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

Answer: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**

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

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.

B) It becomes slow when number of features is very large.
C) We need to iterate.

ASSIGNMENT – 39

MACHINE LEARNING
Q13 and Q15 are subjective answer type questions, Answer them briefly.

13. Explain the term regularization?

While training a machine learning model, the model can easily be overfitted or under fitted. To avoid this, we use regularization in machine learning to properly fit a model onto our test set. Regularization techniques help reduce the chance of overfitting and help us get an optimal model. In this article titled 'The Best Guide to Regularization in Machine Learning', is know about regularization.
What Are Overfitting and Underfitting?
To train our machine learning model, we give it some data to learn from. The process of plotting a series of data points and drawing the best fit line to understand the relationship between the variables is called Data Fitting. Our model is the best fit when it can find all necessary patterns in our data and avoid the random data points and unnecessary patterns called Noise.

If we allow our machine learning model to look at the data too many times, it will find a lot of patterns in our data, including the ones which are unnecessary. It will learn really well on the test dataset and fit very well to it. It will learn important patterns, but it will also learn from the noise in our data and will not be able to predict on other datasets.
A scenario where the machine learning model tries to learn from the details along with the noise in the data and tries to fit each data point on the curve is called Overfitting.
In the figure depicted below, we can see that the model is fit for every point in our data. If given new data, the model curves may not correspond to the patterns in the new data, and the model cannot predict very well in it.

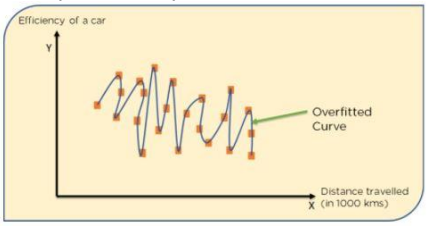


Figure 1: Overfitted Model

Conversely, in a scenario where the model has not been allowed to look at our data a sufficient number of times, the model won't be able to find patterns in our test dataset. It will not fit properly to our test dataset and fail to perform on new data too.
A scenario where a machine learning model can neither learn the relationship between variables in the testing data nor predict or classify a new data point is called Underfitting.
The below diagram shows an under-fitted model. We can see that it has not fit properly to the data given to it. It has not found patterns in the data and has ignored a large part of the dataset. It cannot perform on both known and unknown data.

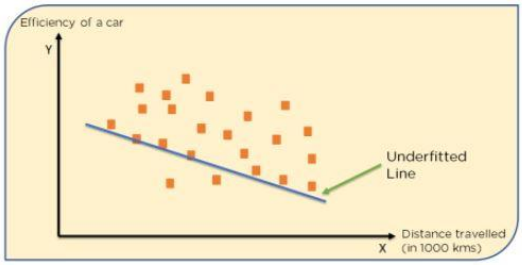


Figure 2: Underfitted Model

What are Bias and Variance?

A Bias occurs when an algorithm has limited flexibility to learn from data. Such models pay very little attention to the training data and oversimplify the model therefore the validation error or prediction error and training error follow similar trends. Such models always lead to a high error on training and test data. High Bias causes underfitting in our model.

Variance defines the algorithm's sensitivity to specific sets of data. A model with a high variance pays a lot of attention to training data and does not generalize therefore the validation error or prediction error are far apart from each other. Such models usually perform very well on training data but have high error rates on test data. High Variance causes overfitting in our model.

An optimal model is one in which the model is sensitive to the pattern in our model, but at the same time can generalize to new data. This happens when Bias and Variance are both optimal. We call this Bias-Variance Tradeoff and we can achieve it in over or under fitted models by using Regression.

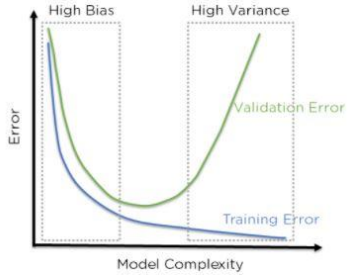


Figure 3: Error in testing and training datasets with high bias and variance

In the above figure, we can see that when bias is high, the error in both testing and training set is also high. When Variance is high, the model performs well on our training set and gives a low error, but the error in our testing set is very high. In the middle of this exists a region where the bias and variance are in perfect balance to each other, and here, but the training and testing errors are low.

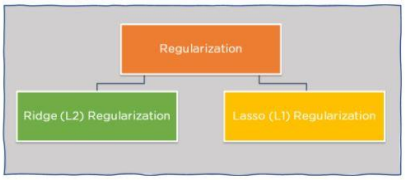
What is Regularization in Machine Learning?

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

Using Regularization, we can fit our machine learning model appropriately on a given test set and hence reduce the errors in it.

Regularization Techniques

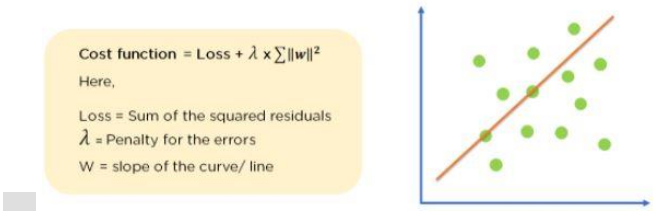
There are two main types of regularization techniques: Ridge Regularization and Lasso Regularization.



Ridge Regularization :

Also known as Ridge Regression, it modifies the over-fitted or under fitted models by adding the penalty equivalent to the sum of the squares of the magnitude of coefficients.

This means that the mathematical function representing our machine learning model is minimized and coefficients are calculated. The magnitude of coefficients is squared and added. Ridge Regression performs regularization by shrinking the coefficients present. The function depicted below shows the cost function of ridge regression :



In the cost function, the penalty term is represented by Lambda λ. By changing the values of the penalty function, we are controlling the penalty term. The higher the penalty, it reduces the magnitude of coefficients. It shrinks the parameters. Therefore, it is used to prevent multicollinearity, and it reduces the model complexity by coefficient shrinkage.
Consider the graph illustrated below which represents Linear regression :

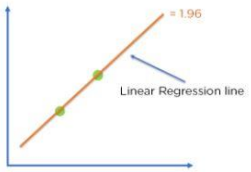


Figure 8: Linear regression model
Cost function = Loss + λ x∑||w||^2
For Linear Regression line, let's consider two points that are on the line,
Loss = 0 (considering the two points on the line)
λ= 1
w = 1.4
Then, Cost function = 0 + 1 x 1.42
= 1.96
For Ridge Regression, let's assume,
Loss = 0.32 + 0.22 = 0.13
λ = 1
w = 0.7
Then, Cost function = 0.13 + 1 x 0.72
= 0.62

Figure 9: Ridge regression model
Comparing the two models, with all data points, we can see that the Ridge regression line fits the model more accurately than the linear regression line.

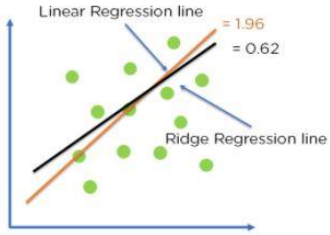
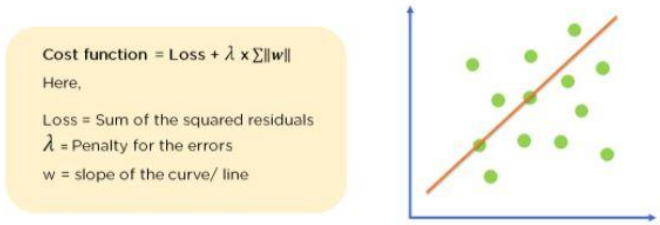


Figure 10: Optimization of model fit using Ridge Regression

Lasso Regression
It modifies the over-fitted or under-fitted models by adding the penalty equivalent to the sum of the absolute values of coefficients. Lasso regression also performs coefficient minimization, but instead of squaring the magnitudes of the coefficients, it takes the true values of coefficients. This means that the coefficient sum can also be 0, because of the presence of negative coefficients.
Consider the cost function for Lasso regression :



We then split our data into training and testing sets.

We can now use these to train our Linear Regression model. We start by creating our model and fitting the data to it. We then predict on the test set and find the error in our prediction using mean_squared_error. Finally, we print the coefficients of our Linear Regression model.

Figure 19: Ridge Regression and plotting coefficients

While training a machine learning model, the model can easily be overfitted or under fitted. To avoid this, we use regularization in machine learning to properly fit a model onto our test set. Regularization techniques help reduce the chance of overfitting and help us get an optimal model.

What Are Overfitting and Underfitting?

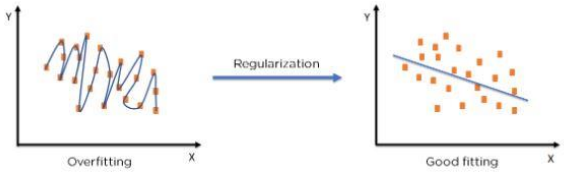
To train our machine learning model, we give it some data to learn from. The process of plotting a series of data points and drawing the best fit line to understand the relationship between the variables is called Data Fitting. Our model is the best fit when it can find all necessary patterns in our data and avoid the random data points and unnecessary patterns called Noise.

If we allow our machine learning model to look at the data too many times, it will find a lot of patterns in our data, including the ones which are unnecessary. It will learn really well on the test dataset and fit very well to it. It will learn important patterns, but it will also learn from the noise in our data and will not be able to predict on other datasets.

A scenario where the machine learning model tries to learn from the details along with the noise in the data and tries to fit each data point on the curve is called Overfitting.

In the figure depicted below, we can see that the model is fit for every point in our data. If given new data, the model curves may not correspond to the patterns in the new data, and the model cannot predict very well in it.

overfitted



Conversely, in a scenario where the model has not been allowed to look at our data a sufficient number of times, the model won't be able to find patterns in our test dataset. It will not fit properly to our test dataset and fail to perform on new data too.

A scenario where a machine learning model can neither learn the relationship between variables in the testing data nor predict or classify a new data point is called Underfitting.

The below diagram shows an under-fitted model. We can see that it has not fit properly to the data given to it. It has not found patterns in the data and has ignored a large part of the dataset. It cannot perform on both known and unknown data.

underfitted

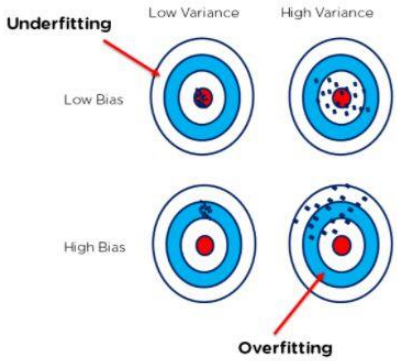


Figure 2: Underfitted Model

.

the word regularize means **to make things regular or acceptable**. This is exactly why we use it for. Regularizations are techniques used to reduce the error by fitting a function appropriately on the given training set and avoid overfitting.

Regularization consists of different techniques and methods used to address the issue of over-fitting by reducing the generalization error without affecting the training error much. Choosing overly complex models for the training data points can often lead to overfitting.

For example Consider the training dataset comprising of independent variables X=(x1,x2….xn) and the corresponding target variables t=(t1,t2,…tn). X are random variables lying uniformly between [0,1]. The target dataset 't' is obtained by substituting the value of X into the function sin(2πx) and then adding some Gaussian noise into it.

The graph for this equation will be a diamond figure as shown below,

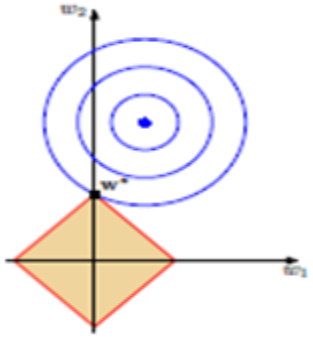


fig 3

Here, the blue circles represent the contours for the un-regularized error function (ED) and diamond shape contour is for L1 regularization term i.e (λ/2(|w1| + |w2|),. We can see in the graph that optimal value is obtained at the point where w1 term is zero i.e the basis function corresponding to w1 term will not affect the output. Here represented by w* where both the terms of cost function will take a common value of 'w' as required in the equation.

Hence we can say that for the proper value, λ the solution vector will be a sparse matrix (eg [0,w2]). So this is how the complexity of the equation (1.3) can be reduced. The solution matrix of **w** will have most of it's values as zero and the non-zero value will contain only the relevant and important information thus finding a general trend for the given dataset.

So we can use this to improve the accuracy of the models.

14. Which particular algorithms are used for regularization?

There are three main regularization techniques, namely: Ridge Regression (L2 Norm) Lasso (L1 Norm) Dropout.
Ridge regression is a regularization technique, which is used to reduce the complexity of the model.
It is also called as L2 regularization. In this technique, the cost function is altered by adding the penalty term to it.
The amount of bias added to the model is called Ridge Regression penalty.
L2 and L1 are the most common types of regularization. Regularization works on the premise that smaller weights lead to simpler models which in results helps in avoiding overfitting.
So to obtain a smaller weight matrix, these techniques add a 'regularization term' along with the loss to obtain the cost function.

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

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.

An error term is a residual variable produced by a statistical or mathematical model, which is created when the model does not fully represent the actual relationship between the independent variables and the dependent variables. As a result of this incomplete relationship, the error term is the amount at which the equation may differ during empirical analysis.

The error term is also known as the residual, disturbance, or remainder term, and is variously represented in models by the letters e, ε, or u.
Error term appears in a statistical model, like a regression model, to indicate the uncertainty in the model.
The error term is a residual variable that accounts for a lack of perfect goodness of fit.
Heteroskedastic refers to a condition in which the variance of the residual term, or error term, in a regression model varies widely.
Understanding an Error Term
An error term represents the margin of error within a statistical model; it refers to the sum of the deviations within the regression line, which provides an explanation for the difference between the theoretical value of the model and the actual observed results. The regression line is used as a point of analysis when attempting to determine the correlation between one independent variable and one dependent variable.

Error Term Use in a Formula
An error term essentially means that the model is not completely accurate and results in differing results during real-world applications. For example, assume there is a multiple linear regression function that takes the following form:
$Y=\alpha X+\beta\rho+\epsilon$
where:
α,β=Constant parameters
X,ρ=Independent variables
ϵ=Error term

When the actual Y differs from the expected or predicted Y in the model during an empirical test, then the error term does not equal 0, which means there are other factors that influence Y.

linear regression most often uses mean-square error (MSE) to calculate the error of the model. MSE is calculated by:
1. measuring the distance of the observed y-values from the predicted y-values at each value of x;
2. squaring each of these distances;
3. calculating the mean of each of the squared distances.

Linear regression fits a line to the data by finding the regression coefficient that results in the smallest MSE.
Linear regression most often uses mean-square error (MSE) to calculate the error of the model. MSE is calculated by: measuring the distance of the observed y-values from the predicted y-values at each value of x; squaring each of these distances; calculating the mean of each of the squared distances.