MACHINE LEARNING

Answers of Question 1 to 11

1.) A) Least Square Error
2.) A) Linear regression is sensitive to outliers
3.) B) Negative
4.) B) Correlation
5.) C) Low bias and high variance
6.) B) Predictive modal
7.) D) Regularization
8.) D) SMOTE
9.) A) TPR and FPR
10.) B) False
11.) A) Construction bag of words from a email

12.) A) We don't have to choose the learning rate.
B) It becomes slow when number of features is very large.

Q.13) Explain the term regularization?

Ans It is a technique to prevent the model from overfitting by adding extra information to it. Sometimes the machine learning model performs well with the training data but does not perform well with the test data. It means the model is not able to predict the output when deals with unseen data by introducing noise in the output, and hence the model is called overfitted. This problem can be deal with the help of a regularization technique.

This technique can be used in such a way that it will allow to maintain all variables or features in the model by reducing the magnitude of the variables. Hence, it maintains accuracy as well as a generalization of the model.

It mainly regularizes or reduces the coefficient of features toward zero. In simple words, "*In regularization technique, we reduce the magnitude of the features by keeping the same number of features."*

How does Regularization Work?

Regularization works by adding a penalty or complexity term to the complex model. Let's consider the simple linear regression equation:

$y= \beta0+\beta1x1+\beta2x2+\beta3x3+\cdots+\beta nxn +b$

In the above equation, Y represents the value to be predicted

X1, X2, …Xn are the features for Y.

$\beta0,\beta1,\ldots..\beta n$ are the weights or magnitude attached to the features, respectively. Here represents the bias of the model, and b represents the intercept.

Linear regression models try to optimize the $\beta0$ and b to minimize the cost function.

Now, we will add a loss function and optimize parameter to make the model that can predict the accurate value of Y. The loss function for the linear regression is called as **RSS or Residual sum of squares.**

Q.14) Which particular algorithms are used for regularization?

Ans There are three main regularization techniques, namely:

1. Ridge Regression (L2 Norm)
2. Lasso (L1 Norm)
3. Dropout

Ridge and Lasso can be used for any algorithms involving weight parameters, including neural nets. Dropout is primarily used in any kind of neural networks e.g. ANN, DNN, CNN or RNN to moderate the learning. Let's take a closer look at each of the techniques.

Ridge Regression (L2 Regularization)

Ridge regression is also called L2 norm or regularization.

When using this technique, we add the sum of weight's square to a loss function and thus create a new loss function which is denoted thus:

$$\text{Loss} = \sum_{j=1}^{m}\left(Yi - Wo - \sum_{i=1}^{n} Wi\,Xji\right)^2 + \lambda \sum_{i=1}^{n} Wi^2$$

As seen above, the original loss function is modified by adding normalized weights. Here normalized weights are in the form of squares.

You may have noticed parameters λ along with normalized weights. λ is the parameter that needs to be tuned using a cross-validation dataset. When you use λ=0, it returns the residual sum of square as loss function which you chose initially. For a very high value of λ, loss will ignore core loss function and minimize weight's square and will end up taking the parameters' value as zero.

Now the parameters are learned using a modified loss function. To minimize the above function, parameters need to be as small as possible. Thus, L2 norm prevents weights from rising too high.

Lasso Regression (L1 Regularization)

Also called lasso regression and denoted as below:

$$\text{Loss} = \sum_{j=1}^{m}\left(Yi - Wo - \sum_{i=1}^{n} Wi\,Xji\right)^2 + \lambda \sum_{i=1}^{n} |Wi|$$

This technique is different from ridge regression as it uses absolute weight values for normalization. λ is again a tuning parameter and behaves in the same as it does when using ridge regression.

As loss function only considers absolute weights, optimization algorithms penalize higher weight values.

In ridge regression, loss function along with the optimization algorithm brings parameters near to zero but not actually zero, while lasso eliminates less important features and sets

respective weight values to zero. Thus, lasso also performs feature selection along with regularization.

Dropout

Dropout is a regularization technique used in neural networks. It prevents complex co-adaptations from other neurons.

In neural nets, fully connected layers are more prone to overfit on training data. Using dropout, you can drop connections with *1-p* probability for each of the specified layers. Where *p* is called **keep probability parameter** and which needs to be tuned.

With dropout, you are left with a reduced network as dropped out neurons are left out during that training iteration.

Dropout decreases overfitting by avoiding training all the neurons on the complete training data in one go. It also improves training speed and learns more robust internal functions that generalize better on unseen data. However, it is important to note that Dropout takes more epochs to train compared to training without Dropout (If you have 10000 observations in your training data, then using 10000 examples for training is considered as 1 epoch).

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

Ans In statistics, an error term is the sum of the deviations of each actual observation from a model regression line. Regression analysis is used to establish the degree of correlation between two variables, one independent and one dependent, the result of which is a line that best "fits" the actually observed values of the dependent value in relation to the **independent variable** or variables. Put another way, an error term is the term in a model regression equation that tallies up and accounts for the unexplained difference between the actually observed values of the independent variable and the results predicted by the model. Hence, the error term is a measure of how accurately the regression model reflects the actual relationship between the independent and dependent variable or variables. The error term can indicate either that the model can be improved, such as by adding in another independent variable that explains some or all of the difference, or by randomness, meaning that the dependent and independent variable or variables are not correlated to any greater degree.

Also known as the residual term or disturbance term, according to mathematical convention, the error term is the last term in a model regression equation and is represented by the Greek letter epsilon (ε). Economists and financial industry professionals regularly make use of regression models, or at least their results, to better understand and forecast a wide range of relationships, such as how changes in the money supply are related to inflation, how stock market prices are related to unemployment rates or how changes in commodity prices affect specific companies in an economic sector. Hence, the error term is an important variable to keep in mind and keep track of in that it measures the degree to which any given model does not reflect, or account for, the actual relationship between the dependent and independent variables.

There are actually two types of error terms commonly used in regression analysis: absolute error and relative error. Absolute error is the error term as previously defined, the difference between the actually observed values of the independent variable and the results predicted by

the model. Derived from this, relative error is defined as the absolute error divided by the exact value predicted by the model. Expressed in percentage terms, relative error is known as percent error, which is helpful because it puts the error term into greater perspective. For example, an error term of 1 when the predicted value is 10 is much worse than an error term of 1 when the predicted value is 1 million when attempting to come up with a regression model that shows how well two or more variables are correlated.