ASSIGNMENT -39

MACHINE LEARNING

Answer Find Below

Ans 1- Least Square Error

Ans 2-Linear regression is sensitive to outliers

Ans 3-Negative

Ans 4- Both of them

Ans- 5-High bias and high variance

Ans 6-Predictive modal

Ans-7-Regularization

Ans-8-Cross validation

Ans- 9-TPR and FPR

Ans 10- True

Ans 11-Apply PCA to project high dimensional data

Ans 12- A,B,C Option Are Correct

Ans 13- 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.

It is one of the most important concepts of machine learning. This technique prevents the model from over fitting by adding **extra information** to it.

It is a form of regression that shrinks the coefficient estimates towards zero. In other words, this technique forces us not to learn a more complex or flexible model, to avoid the problem of over fitting.

Now, let’s understand the **“How flexibility of a model is represented?”**. For regression problems, **the increase in flexibility of a model is represented by an increase in its coefficients**, which are calculated from the regression line.

In simple words, **“In the Regularization technique, we reduce the magnitude of the independent variables by keeping the same number of variables”.** It maintains accuracy as well as a generalization of the model.

How does Regularization Work?

Regularization works by adding a penalty or complexity term or shrinkage term with Residual Sum of Squares (RSS) to the complex model.

Let’s consider the **Simple linear regression** equation:

Here Y represents the dependent feature or response which is the learned relation. Then,

Y is approximated to $\boldsymbol{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p}$

Here, $\boldsymbol{X_1, X_2, \ldots X_p}$ are the independent features or predictors for Y, and

$\boldsymbol{\beta_0, \beta_1, \ldots\ldots \beta_n}$ represents the coefficients estimates for different variables or predictors(X), which describes the weights or magnitude attached to the features, respectively.

In simple linear regression, our optimization function or loss function is known as the **residual sum of squares (RSS).**

We choose those set of coefficients, such that the following loss function is minimized:

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

Now, this will adjust the coefficient estimates based on the training data. If there is noise present in the training data, then the estimated coefficients won't generalize well and are not able to predict the future data.

This is where regularization comes into the picture, which shrinks or regularizes these learned estimates towards zero, by adding a loss function with optimizing parameters to make a model that can predict the accurate value of Y.

Techniques of Regularization

Mainly, there are two types of regularization techniques, which are given below:

1. Ridge Regression
2. Lasso Regression

Ridge Regression

Ridge regression is one of the types of linear regression in which we introduce a small amount of bias, known as **Ridge regression penalty** so that we can get better long-term predictions.

In Statistics, it is known as the **L-2 norm**.

In this technique, the cost function is altered by adding the penalty term (shrinkage term), which multiplies the lambda with the squared weight of each individual feature. Therefore, the optimization function(cost function) becomes:

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

In the above equation, the penalty term regularizes the coefficients of the model, and hence ridge regression reduces the magnitudes of the coefficients that help to decrease the complexity of the model.

Usage of Ridge Regression:

- When we have the independent variables which are having high collinearity (problem of multicollinearity) between them, at that time general linear or polynomial regression will fail so to solve such problems, Ridge regression can be used.
- If we have more parameters than the samples, then Ridge regression helps to solve the problems.

- **Limitation of Ridge Regression:**

- **Not helps in Feature Selection:** It decreases the complexity of a model but does not reduce the number of independent variables since it never leads to a coefficient being zero rather only minimizes it. Hence, this technique is not good for feature selection.
- **Model Interpretability:** Its disadvantage is model interpretability since it will shrink the coefficients for least important predictors, very close to zero but it will never make them exactly zero. In other words, the final model will include all the independent variables, also known as predictors.

Lasso Regression

Lasso regression is another variant of the regularization technique used to reduce the complexity of the model. It stands for **Least Absolute and Selection Operator**.

It is similar to the Ridge Regression except that the penalty term includes the absolute weights instead of a square of weights. Therefore, the optimization function becomes:

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

In statistics, it is known as the **L-1 norm**.

In this technique, the L1 penalty has the effect of forcing some of the coefficient estimates to be exactly equal to zero which means there is a complete removal of some of the features for model evaluation when the tuning parameter λ is sufficiently large. Therefore, the lasso method also performs **Feature selection** and is said to yield **sparse models**.

Limitation of Lasso Regression:

- **Problems with some types of Dataset:** If the number of predictors is greater than the number of data points, Lasso will pick at most n predictors as non-zero, even if all predictors are relevant.
- **Multicollinearity Problem:** If there are two or more highly collinear variables then LASSO regression selects one of them randomly which is not good for the interpretation of our model.

Ans-14

There are three algorithms are used for regularization, 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.

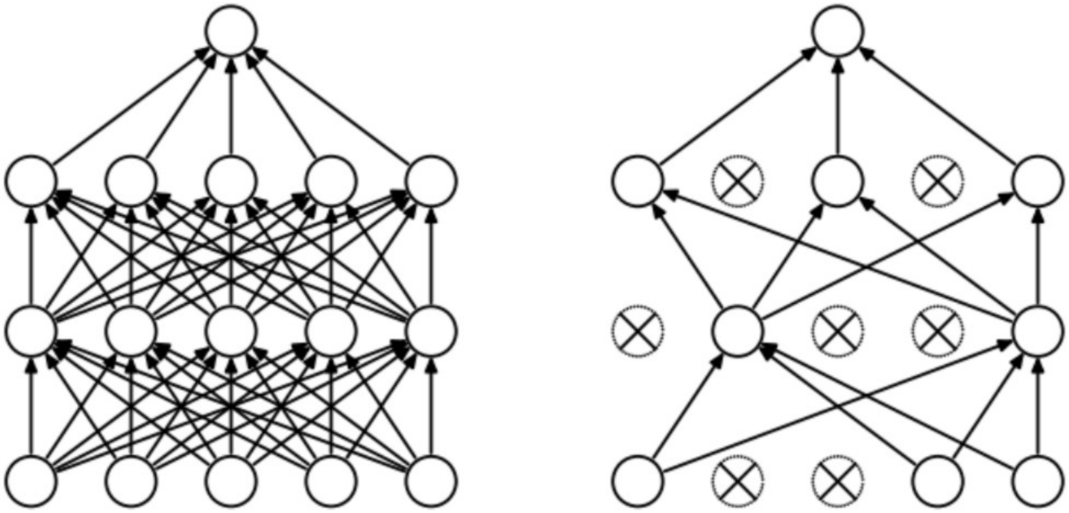
As loss function only considers absolute weights, optimization algorithms penalize higher weight values.

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.



(a) Standard Neural Net

(b) After applying dropout.

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).

Along with Dropout, neural networks can be regularized also using L1 and L2 norms. Apart from that, if you are working on an image dataset, image augmentation can also be used as a regularization method.

For real-world applications, it is a must that a model performs well on unseen data. The techniques we discussed can help you make your model learn rather than just memorize.

Conclusion

Be it an over-fitting or under-fitting problem, it will lower down the overall performance of a machine learning model. To get the best out of machine learning models, you must optimize and tune them well. At eInfochips, we deliver machine learning services that help businesses optimize the utilization AI technology. We have machine learning capabilities across cloud, hardware, neural networks, and open source frameworks.

Ans 15-

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.

KEY TAKEAWAYS

- An 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:

\begin{aligned} &Y = \alpha X + \beta \rho + \epsilon \\ &\textbf{where:} \\ &\alpha, \beta = \text{Constant parameters} \\ &X, \rho = \text{Independent variables} \\ &\epsilon = \text{Error term} \\ \end{aligned}
$Y=\alpha X+\beta\rho+\epsilon$**where:**$\alpha,\beta$=Constant parameters$X,\rho$=Independent variables$\epsilon$=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.

What Do Error Terms Tell Us

Within a linear regression model tracking a stock’s price over time, the error term is the difference between the expected price at a particular time and the price that was actually observed. In instances where the price is exactly what was anticipated at a particular time, the price will fall on the trend line and the error term will be zero.

Points that do not fall directly on the trend line exhibit the fact that the dependent variable, in this case, the price, is influenced by more than just the independent variable, representing the passage of time. The error term stands for any influence being exerted on the price variable, such as changes in market sentiment.

The two data points with the greatest distance from the trend line should be an equal distance from the trend line, representing the largest margin of error.

If a model is heteroskedastic, a common problem in interpreting statistical models correctly, it refers to a condition in which the variance of the error term in a regression model varies widely.

Linear Regression, Error Term, and Stock Analysis

Linear regression is a form of analysis that relates to current trends experienced by a particular security or index by providing a relationship between a dependent and independent variables, such as the price of a security and the passage of time, resulting in a trend line that can be used as a predictive model.

A linear regression exhibits less delay than that experienced with a moving average, as the line is fit to the data points instead of based on the averages within the data. This allows the line to change more quickly and dramatically than a line based on numerical averaging of the available data points.

The Difference Between Error Terms and Residuals

Although the error term and residual are often used synonymously, there is an important formal difference. An error term is generally unobservable and a residual is observable and calculable, making it much easier to quantify and visualize. In effect, while an error term represents the way observed data differs from the actual population, a residual represents the way observed data differs from sample population data.