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:- (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

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

ANSEWER:- 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:-(A) High 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

ANSWER:-(B) Descriptive model

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

ANSWER:- (A) TRUE

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

ANSWER:-(D) FORWARD SELECTION

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.

ANSWER:-(C) We need to iterate.

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

13. Explain the term regularization ?

Regularization is a technique used to reduce the errors by fitting the function appropriately on the given training set and avoid overfitting.
The commonly used regularization techniques are :

1. L1 regularization
2. L2 regularization
3. Dropout regularization

This article focus on L1 and L2 regularization.

A regression model which uses **L1 Regularization** technique is called **LASSO(Least Absolute Shrinkage and Selection Operator)** regression. A regression model that uses **L2 regularization** technique is called **Ridge regression**.
Lasso Regression adds *"absolute value of magnitude"* of coefficient as penalty term to the loss function(L).

$$||\mathbf{w}||_1 = |w_1| + |w_2| + ... + |w_N|$$

Ridge regression adds "*squared magnitude*" of coefficient as penalty term to the loss function(L).

$$\|\mathbf{w}\|_2 = \left(|w_1|^2 + |w_2|^2 + \ldots + |w_N|^2\right)^{\frac{1}{2}}$$

NOTE that during Regularization the output function(y_hat) does not change. The change is only in the loss function.
The output function:

$$\hat{y} = w_1 x_1 + w_2 x_2 + ... + w_N x_N + b$$

The loss function before regularization:

$$Loss = Error(y, \hat{y})$$

The loss function after regularization:

$$Loss = Error(y, \hat{y}) + \lambda \sum_{i=1}^{N} |w_i|$$

$$Loss = Error(y, \hat{y}) + \lambda \sum_{i=1}^{N} w_i^2$$

We define Loss function in Logistic Regression as :

```
L(y_hat,y) = y log y_hat + (1 - y)log(1 - y_hat)
```

Loss function with no regularization :

$$L = y \log (wx + b) + (1 - y)\log(1 - (wx + b))$$

Lets say the data overfits the above function.

Loss function with L1 regularization :

$$L = y \log (wx + b) + (1 - y)\log(1 - (wx + b)) + \text{lambda} * ||w||_1$$

Loss function with L2 regularization :

$$L = y \log (wx + b) + (1 - y)\log(1 - (wx + b)) + \text{lambda} * ||w||_2^2$$

lambda is a Hyperparameter Known as regularization constant and it is greater than zero.

$$\text{lambda} > 0$$

14. Which particular algorithms are used for regularization?

This article was published as a part of the Data Science Blogathon

Introduction

One of the most common problems every Data Science practitioner faces is **Overfitting**. Have you tackled the situation where your machine learning model performed exceptionally well on the train data but was not able to predict on the unseen data or you were on the top of the competition in the public leaderboard, but your ranking drops by hundreds of places in the final rankings?

Well – this is the article for you!

Avoiding overfitting can single-handedly improve our model's performance.

In this article, we will understand how regularization helps in overcoming the problem of overfitting and also increases the model interpretability.

This article is written under the assumption that you have a basic understanding of **Regression models** including Simple and Multiple linear regression, etc.

Table of Contents

Why Regularization?

Sometimes what happens is that our Machine learning model performs well on the training data but does not perform well on the unseen or test data. It means the model

is not able to predict the output or target column for the unseen data by introducing noise in the output, and hence the model is called an **overfitted model**.

Let's understand the meaning of **"Noise"** in a brief manner:

By noise we mean those data points in the dataset which don't really represent the true properties of your data, but only due to a random chance.

So, to deal with the problem of overfitting we take the help of regularization techniques.

What is Regularization?

- It is one of the most important concepts of machine learning. This technique prevents the model from overfitting 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 overfitting.
- 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 $\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$

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

$\beta_0, \beta_1, \ldots \beta_n$ 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:

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

Fig. Cost Function For Simple Linear Regression

Image Source: link

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:

- Ridge Regression
- 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:

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

Fig. Cost Function for Ridge Regression

Image Source: link

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:

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

Fig. Cost Function for Lasso Regression

Image Source:link

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.

Key Differences between Ridge and Lasso Regression

Ridge regression helps us to reduce only the overfitting in the model while keeping all the features present in the model. It reduces the complexity of the model by shrinking the coefficients whereas Lasso regression helps in reducing the problem of overfitting in the model as well as automatic feature selection.

Lasso Regression tends to make coefficients to absolute zero whereas Ridge regression never sets the value of coefficient to absolute zero.

Mathematical Formulation of Regularization Techniques

Now, we are trying to formulate these techniques in mathematical terms. So, these techniques can be understood as solving an equation,

For ridge regression, the total sum of squares of coefficients is less than or equal to s and **for Lasso regression**, the total sum of modulus of coefficients is less than or equal to s.

Here, s is a constant which exists for each value of the shrinkage factor λ.

These equations are also known as **constraint functions**.

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

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:

$$\text{where:}$$

$$\text{Constant parameters}$$

$$\text{Independent variables}$$

$$\text{Error term}$$

$$Y=\alpha X+\beta\rho+\epsilon$$

where:

$$\alpha,\beta=\text{Constant parameters}$$

$$X,\rho=\text{Independent variables}$$

$$\epsilon=\text{Error term}$$

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.