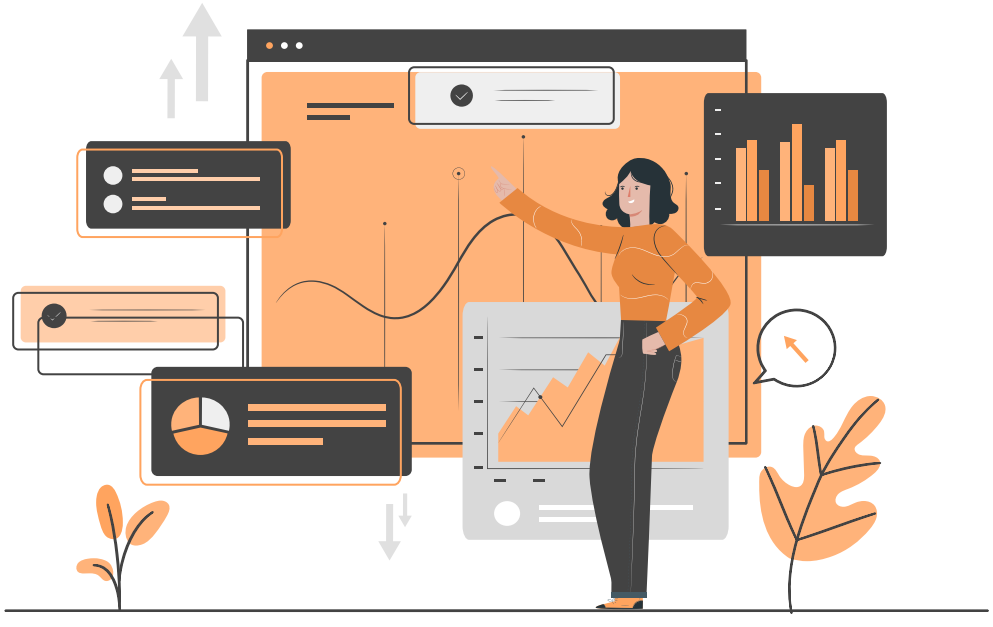


L1 & L2 Regularization Techniques

By Fireblaze AI School

Index

Index

- Key differences between Ridge and Lasso Regression
- Mathematical Formulation of Regularization Techniques
- What does Regularization Achieve?

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?

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.

Why Regularization ?

- 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?"

What is Regularization ?

- 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 $\beta 0 + \beta 1X1 + \beta 2X2 + \ldots + \beta pXp$

How Does Regularization Work?

- Here, X1, X2, …Xp are the independent features or predictors for Y, and
- β0, β1,…..βn 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).

How Does Regularization Work?

RSS = Σ(yi – ŷi)^2

where:

Σ: A greek symbol that means sum

yi: The actual response value for the ith observation

ŷi: The predicted response value based on the multiple linear regression model

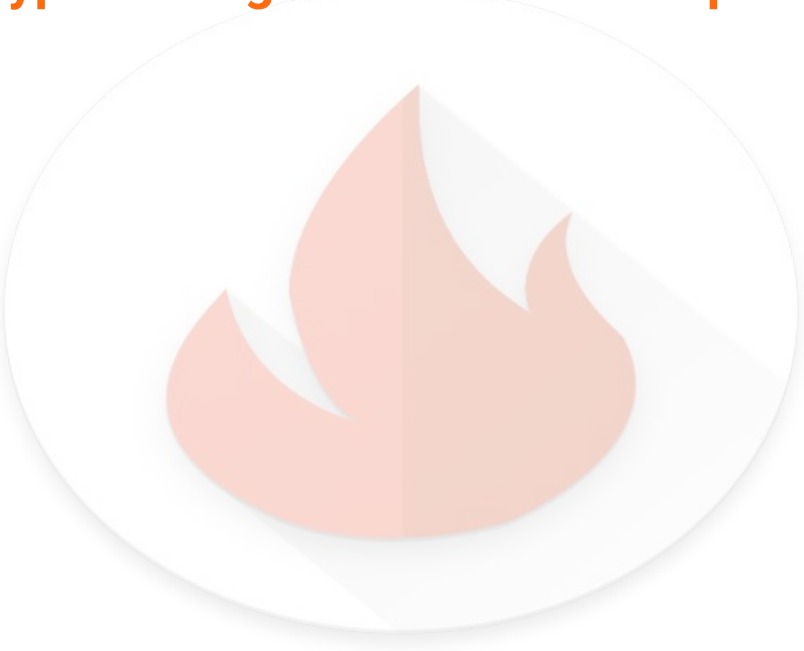
How Does Regularization Work?

- 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.

Ridge Regression

- Therefore, the optimization function(**cost function**) becomes:

$$SSE_{L_2} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{P} \beta_j{}^2$$

- 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.

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:

Lasso Regression Cost Function

- Therefore, the optimization function becomes:

$$Cost\ (W) = RSS(W) + \lambda * (sum\ of\ absolute\ value\ of\ weights)$$

$$= \sum_{i=1}^{N} \left\{ y_i - \sum_{j=0}^{M} w_j\, x_{ij} \right\}^2 + \lambda \sum_{j=0}^{M} |w_j|$$

Lasso Regression

- 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.

Difference

Key Difference Between Ridge & 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.

What Does Regularization Achieve?

- In simple linear regression, the standard least-squares model tends to have some variance in it, i.e. this model won't generalize well for a future data set that is different from its training data.
- Regularization tries to reduce the variance of the model, without a substantial increase in the bias.

Parameter λ

How λ relates to the principle of “Curse of Dimensionality”?

- As the value of λ rises, it significantly reduces the value of coefficient estimates and thus reduces the variance. Till a point, this increase in λ is beneficial for our model as it is only reducing the variance (hence avoiding overfitting), without losing any important properties in the data. But after a certain value of λ, the model starts losing some important properties, giving rise to bias in the model and thus underfitting. Therefore, we have to select the value of λ carefully. To select the good value of λ, cross-validation comes in handy.

Important Point About λ

- λ is the tuning parameter used in regularization that decides how much we want to penalize the flexibility of our model i.e, controls the impact on bias and variance.
- When λ = 0, the penalty term has no effect, the equation becomes the cost function of the linear regression model. Hence, for the minimum value of λ i.e, λ=0, the model will resemble the linear regression model. So, the estimates produced by ridge regression will be equal to least squares.
- However, as $\lambda \rightarrow \infty$ (tends to infinity), the impact of the shrinkage penalty increases, and the ridge regression coefficient estimates will approach zero.


The End