**REGULARIZATION IN MACHINE LEARNING: WHAT, WHY AND HOW?**

**ANSWERED!**

Hello, today, we will talk about regularization in Machine Learning. Specifically, we will focus on Ridge and Lasso Regression techniques followed by some basic information on other regularization techniques used in the field of Machine Learning. So, without wasting much time, let’s get started….

**WHAT AND WHY OF REGULARIZATION:**

When we are working on any data science project, the main aim is to get a good performance estimate (High accuracy or low error rate) on both, training and testing data. Some simpler models tend to “underfit” the data, while some complex models tend to “overfit” a lot. In simpler words, overfitting is a situation in which our model learns too much about the training data and hence, it almost fits perfectly to the training data and gives a great accuracy score or lower cost function on training data. However, learning the training data too perfectly makes the scenario worse since our model loses its ability to generalize on the trend of the data resulting in very poor performance on test data. In ML terms, this overfitting case is best described as *LOW BIAS HIGH VARIANCE* scene. On the other hand, underfitting occurs when model is too simple, that is, model neither learns about training data nor performs better on test data. We say that our model is suffering from underfitting when a random guess is more accurate than our model. Underfitting is best described as *HIGH BIAS HIGH VARIANCE* scene. Our main aim while developing any machine learning model is to get to the point of *LOW BIAS LOW VARIANCE.* However, as most of the things in life, we can’t get best of both the worlds 😉 and hence, we have to settle in between while working out our way towards the *LOW BIAS LOW VARIANCE* scenario. Therefore, we always try to find a sweet spot in between of overfitting and underfitting. Regularization is a technique which tries to help us during our quest of finding this sweet spot. This adjustment of settling in between of bias and variance is also called as **Bias Variance Tradeoff** and therefore, it is often said that regularization is a way to achieve correct bias variance tradeoff. This bias variance tradeoff can be best visualized on line as follows: (Apologies for the bad drawing, but I guess it will serve the purpose anyway.)



Overfitting sweet spot Underfitting



**“HOW” OF REGULARIZATION**

As you might be aware of the fact that, regression analysis is mainly performed to predict continuous or in more simpler terms, numerical response variable. Most often, we use root mean squared error or absolute mean error to calculate error function and to get estimate of how good our model is performing. To get better insight, consider following example where our model is overfitting the training data. The blue curve denotes the scenario of overfitting, where if our model is provided with data that don’t lie on blue curve, it will give high error as compared to the model described by black line. To counter this drawback of overfitting, regularization tries to minimize the slopes of model. When we talk about slope in regression analysis, it is generally the weights of the attributes that we are using to predict response variable. Regularization tries to penalize models having high weight to the attributes by adding a penalty term to the cost function of model.

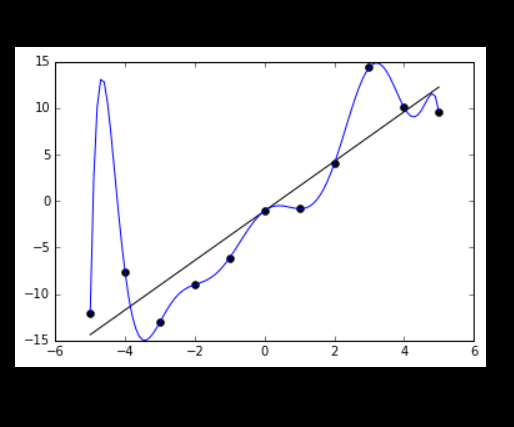


Image Source: <https://en.wikipedia.org/wiki/Overfitting>

So, if we are using cost function as root mean squared error for our regression analysis, then regularization adds an extra term to cost function which is a function of weights. This extra term then increases the cost of the overfitted model and hence, we try to minimize this newly found cost function. In this way, we try to find more generalized model for our data and try to get out from the trap of overfitting.

There are two types of regularization techniques in regression:

1)Ridge regression (L2 Regularization)

Formula: + λ\*()

2) LASSO Regression (L1 Regularization)

Formula: + λ\*()

,where, wi is weight of each feature.

As you can see, weights in ridge regression(wi^2) can go towards zero, but will never be exactly zero, while in case of lasso regression, weight can become zero(wi). Hence, we can have zero weight features in lasso regression and hence, lasso regression can also be used for feature selection.

**Regularization used in other tasks in Machine learning:**

1. Bagging and boosting are regularization techniques used in decision trees.
2. Dropout layer in neural network is implemented as a mean of regularization.

**General Idea based on my experience:**

1. Always standardize data before using regularization techniques.
2. For smaller dataset, ridge regression performs better while for larger datasets, lasso performs better.
3. Value of λ drives regularization. If the model is overfitting, try to use high values for lambda as it will lead to drastic increase in cost function. If the model is underfitting, try to use smaller lambda as it will denote that we are giving less importance to decreasing weight of features and hence, it will emphasize more on reducing root mean squared error rather than penalty terms of weights.

Feel free to give suggestions …Cheers !