**Regularization Lasso |Ridge Regression**

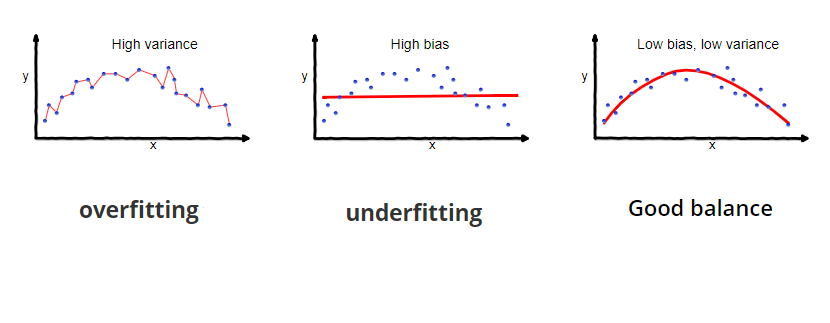
[](https://medium.com/@alielagrebi?source=post_page-----105f426b749c----------------------)

[ALi Elagrebi](https://medium.com/@alielagrebi?source=post_page-----105f426b749c----------------------)

[Feb 26, 2019](https://medium.com/@alielagrebi/regularization-lasso-ridge-regression-105f426b749c?source=post_page-----105f426b749c----------------------) · 6 min read

Mastering the trade-off between Bias and Variance is the hardest thing to become a machine Learning Champion.

The definition is about how well the model fits to the training data to build the model , and how well the model will generalize in the testing data in order to minimize the expected test error to achieve low variance and low bias.

**The concept of bias-variance trade-off is very helpful to understand the overfitting and underfitting .**

**[Balancing Bias and Variance to Control Errors in Machine Learning](https://towardsdatascience.com/balancing-bias-and-variance-to-control-errors-in-machine-learning-16ced95724db)**

**[In the world of Machine Learning, accuracy is everything. You strive to make your model more accurate by tuning and…](https://towardsdatascience.com/balancing-bias-and-variance-to-control-errors-in-machine-learning-16ced95724db)**

**[towardsdatascience.com](https://towardsdatascience.com/balancing-bias-and-variance-to-control-errors-in-machine-learning-16ced95724db)**

After i took “applied machine learning with python” course in coursera offered by university of Michigan.I decide to share some topics and dive deeper into it in this post .

**[Applied Machine Learning in Python | Coursera](http://Applied Machine Learning in Python | CourseraThis course will introduce the learner to applied machine learning, focusing more on the techniques and methods than on…www.coursera.org)**

**[This course will introduce the learner to applied machine learning, focusing more on the techniques and methods than on…](http://Applied Machine Learning in Python | CourseraThis course will introduce the learner to applied machine learning, focusing more on the techniques and methods than on…www.coursera.org)**

**[www.coursera.org](http://Applied Machine Learning in Python | CourseraThis course will introduce the learner to applied machine learning, focusing more on the techniques and methods than on…www.coursera.org)**

Any Machine learning model can fit the in-sample data very well and the out-sample data very poorly.This is know as overfitting (low Bias and high variance).

Similarly,it could fit the in-sample and out-sample data very poorly(high bias and low variance).This is know as underfitting .

In this post , I will cover a technique to improve our model when it overfits in regression problems.I try to answer those questions:

**what is Regularization ?**

**what problem do Regularization methods solve ?**

**what are L1 and L2 Regularization methods ?**

**Let the war against overfitting begin 😂**

The famous Overfitting happens when our model tries too hard to capture the random noise in our training data instead of genearalize it .this mean that the statistical model starts memorizing the data points that they don’t really represent the true properties of the data- the errors, and all the slightly specific characteristics of your sample. This turns our model to be more flexible at the risk of overfitting.

**But what we can do in this case to avoid overfitting?** 😧



The main concept behind avoiding overfit is simplifying the models as much as possible.and pay attention the to the trade-off between overfitting and underfitting a model.

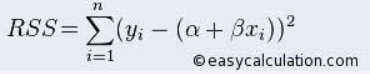
Let’s talk about **Regularization** :

is an extremely important concept in machine learning. It’s a way to prevent overfitting, and thus, improve the likely generalization performance of a model by reducing the complexity of the final estimated model.But how ?

we all know that this simple relation of linear regression :



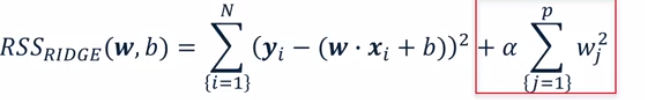
The fitting procedure involves the loss function as residual sum of squares or RSS.The cofficients β chosen to minimize this loss function based on the training data :



So if the model tries to memorize the noise in trainning data, then the estimated cofficients to minimize the RSS will never generalize to the out-sample data.This is where Regularization comes in by adding a second element( a constraint) and The reason for doing that is **to “*punish*”,** **“*shrink”* or “*regularize”* the high values of the coefficients β towards zero.** The practical effect of using regularization is to find the feature weights **β** that fit the data well.We don’t see this effect with a single variable linear regression example,but for regression problems with dozens or hundreds of features,the accuracy improvement from using regularized linear regression like ridge regression or lasso regression could be significant.

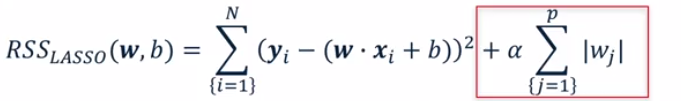
**Regularization Methods :**

**Ridge Regression or L2 Regularization:**



As shown in the above image,the RSS modified by imposing that sum of squares penalty on the size of the W coefficients. The super power of Ridge Regression is that it minimize the RSS by enforce the W coefficients to be lower, but it **does not** enforce them to be zero- **minimize their impact on the trained model to simplify the statistical model.**

**Lasso Regression or L1 Regularization:**



Another kind of regularized regression that you could use instead of ridge regression is called Lasso Regression or L1 Regularization.Like ridge regression, lasso regression adds a regularisation penalty term to the ordinary least-squares objective.but **the results are noticeably different**

With lasso regression, a subset of the coefficients are forced to be precisely zero. Which is a kind of automatic feature selection, since with the weight of zero the features are essentially ignored completely in the model. This sparse solution where only a subset of the most important features are left with non-zero weights, also makes the model easier to interpret which is a huge advantage.

**But Wait !!! what is and what did the α parameter do ?**

Now, we try to facing the critical part in Regularization 😶.

The amount of regularisation to apply is controlled by the α parameter . **significantly regularisation** **reduces the variance of the model, without substantial increase in its bias.** Remember the trade-off we’ve discussed before ,the tuning parameter **α** controls this trade-off.

As the value of **α** increases , it reduces the value of coefficients and thus reducing the variance( ***hence avoiding overfitting),***But after certain value , the model starts loosing important properties of the data ,and this increase the bias (***underfitting problem*** ). Therefore, the value of **α** should be **carefully selected by using some Hyper Parameter Tuning Techniques.**

**What about Feature scaling ? :**

**why Feature scaling ?**

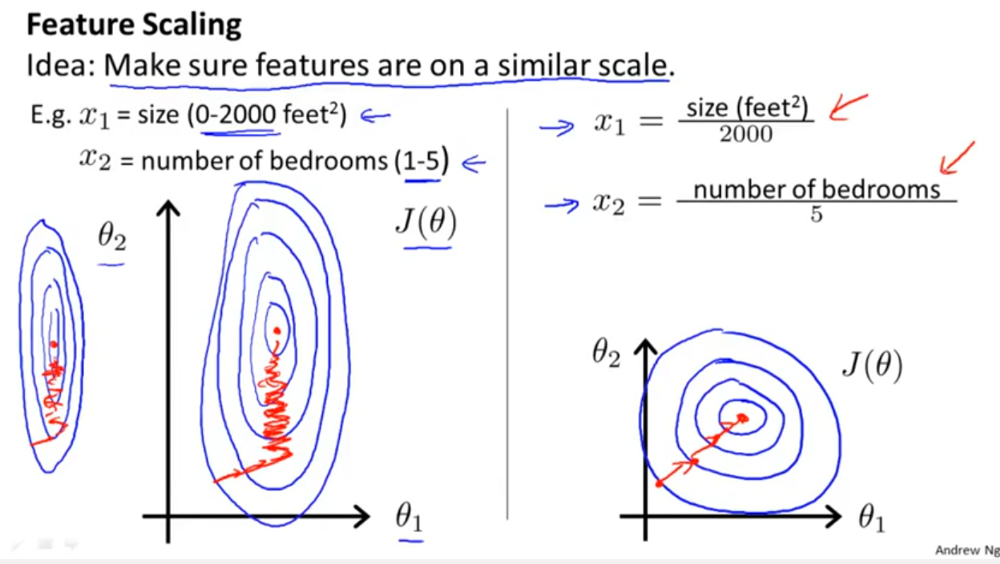


Big Monster Vs Small man

Many of the machine learning algorithms use euclidean distance between data point in their computation.Having two features with different range of numbers will let the feature with bigger range dominate the algorithm.

Andrew Ng has a great explanation in his coursera videos [here](https://www.coursera.org/learn/machine-learning).

I will illustrate the core idea here of Feature scaling using Andrew’s slides:



As you see here,suppose you have two parameters(features) and one of them take a relatively large range of values .Then,your gradients could take a long time to find the optimal solution by going back and forth .But, by scaling our features, it help in quicker converging in gradient decent algorithms. As shown above.

**When to use Feature scaling ?:**

We should scale our features before passing them to any model **that computes distance**.as the case here with Ridge and Lasso Regression

**How to scale Features ?:**

***The scikit-learn preprocessing module*** has excellent api and documentation on feature scaling [here](https://scikit-learn.org/stable/modules/preprocessing.html#standardization-or-mean-removal-and-variance-scaling).There are many common methods to perform Feature Scaling Like : **Standardisation,** **Mean Normalisation,** **Min-Max Scaling,** **Unit Vector and Robust Scaler.**