

f (<https://www.facebook.com/AnalyticsVidhya>) | t (<https://twitter.com/analyticsvidhya>)

g+ (<https://plus.google.com/+Analyticsvidhya/posts>)

in (<https://www.linkedin.com/groups/Analytics-Vidhya-Learn-everything-about-5057165>)



(<http://datahack.analyticsvidhya.com/contest/date-your-data>)

Home (<http://www.analyticsvidhya.com/>) > Business Analytics (<http://www.analyticsvidhya.com/blog/category/business-analytics/>)

# How to avoid Over-fitting using Regularization?

BUSINESS ANALYTICS ([HTTP://WWW.ANALYTICSVIDHYA.COM/BLOG/CATEGORY/BUSINESS-ANALYTICS/](http://www.analyticsvidhya.com/blog/category/business-analytics/))

SHARE f (<http://www.facebook.com/sharer.php?u=http://www.analyticsvidhya.com/blog/2015/02/avoid-over-fitting-regularization/&t=How%20to%20avoid%20over-fitting%20using%20Regularization?>) t (<https://twitter.com/home?status=How%20to%20avoid%20over-fitting%20using%20Regularization?+http://www.analyticsvidhya.com/blog/2015/02/avoid-over-fitting-regularization/>) g+ (<https://plus.google.com/share?url=http://www.analyticsvidhya.com/blog/2015/02/avoid-over-fitting-regularization/>) p (<http://pinterest.com/pin/create/button/?url=http://www.analyticsvidhya.com/blog/2015/02/avoid-over-fitting-regularization/&media=http://www.analyticsvidhya.com/wp-content/uploads/2015/02/How-to-avoid-Over-fitting-using-Regularization.jpg&description=How%20to%20avoid%20over-fitting%20using%20Regularization?>)

		<b>JOIN OUR PREDICTIVE BUSINESS ANALYTICS PROGRAM</b> Taught by Industry Experts          For Working Professionals	<b>PLACEMENT ASSURED</b> <a href="#">KNOW MORE</a>
--	--	--	---

([http://admissions.bridgesom.com/pba-new/?utm\\_source=AV&utm\\_medium=Banner&utm\\_campaign=AVBanner](http://admissions.bridgesom.com/pba-new/?utm_source=AV&utm_medium=Banner&utm_campaign=AVBanner))

Occam's Razor, a problem solving principle states that

**“Among competing hypotheses, the one with the fewest assumptions should be selected. Other, more complicated solutions may ultimately prove correct, but—in the absence of certainty—the fewer assumptions that are made, the better.”**



(<http://i1.wp.com/www.analyticsvidhya.com/wp-content/uploads/2015/02/How-to-avoid-Over-fitting-using-Regularization.jpg>)

## Business Situation:

In the world of analytics, where we try to fit a curve to every pattern, Over-fitting is one of the biggest concerns. However, in general models are equipped enough to avoid over-fitting, but in general there is a manual intervention required to make sure the model does not consume more than enough attributes.

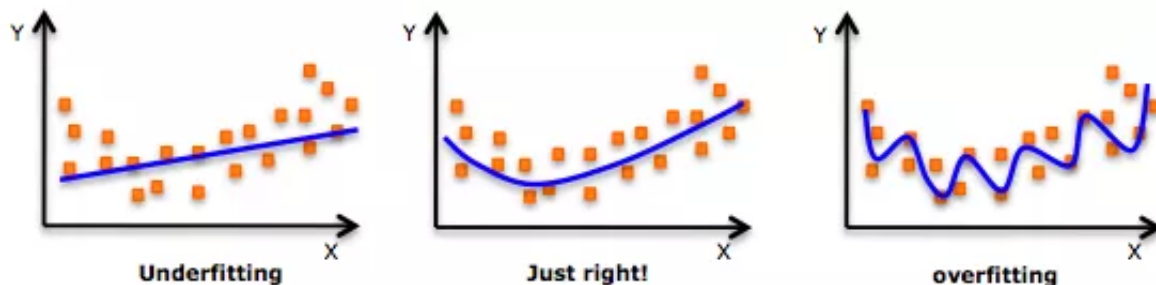
Let's consider an example here, we have 10 students in a classroom. We intend to train a model based on their past score to predict their future score. There are 5 females and 5 males in the class. The average score of females is 60 whereas that of males is 80. The overall average of the class is 70.

Now, there are several ways to make the prediction:

- Predict the score as 70 for the entire class.
- Predict score of males = 80 and females = 60. This a simplistic model which might give a better estimate than the first one.
- Now let's try to overkill the problem. We can use the roll number of students to make a prediction and say that every student will exactly score same marks as last time. Now, this is unlikely to be true and we have reached such granular level that we can go seriously wrong.

The first case here is called under fit, the second being an optimum fit and last being an over-fit.

Have a look at the following graphs,



(<http://i2.wp.com/www.analyticsvidhya.com/wp-content/uploads/2015/02/underfitting-overfitting.png>)

Image source: pingax.com

The trend in above graphs looks like a quadratic trend over independent variable X. A higher degree polynomial might have a very high accuracy on the train population but is expected to fail badly on test dataset. We will briefly touch up on various techniques we use to avoid over-fitting. And then focus on a special technique called **Regularization**.

## Methods to avoid Over-fitting:

Following are the commonly used methodologies :

1. **Cross-Validation** : Cross Validation in its simplest form is a one round validation, where we leave one sample as in-time validation and rest for training the model. But for keeping lower variance a higher fold cross validation is preferred.
2. **Early Stopping** : Early stopping rules provide guidance as to how many iterations can be run before the learner begins to over-fit.
3. **Pruning** : Pruning is used extensively while building CART models. It simply removes the nodes which add little predictive power for the problem in hand.
4. **Regularization** : This is the technique we are going to discuss in more details. Simply put, it introduces a cost term for bringing in more features with the objective function. Hence, it tries to push the coefficients for many variables to zero and hence reduce cost term.

## Regularization basics

A simple linear regression is an equation to estimate  $y$ , given a bunch of  $x$ . The equation looks something as follows :

$$y = a_1x_1 + a_2x_2 + a_3x_3 + a_4x_4 + \dots$$

In the above equation,  $a_1, a_2, a_3 \dots$  are the coefficients and  $x_1, x_2, x_3 \dots$  are the independent variables. Given a data containing  $x$  and  $y$ , we estimate  $a_1, a_2, a_3 \dots$  based on an objective function. For a linear regression the objective function is as follows :

$$\min_f \sum |Y_i - f(X_i)|^2$$

([http://i0.wp.com/www.analyticsvidhya.com/wp-content/uploads/2015/02/initial\\_eq.png](http://i0.wp.com/www.analyticsvidhya.com/wp-content/uploads/2015/02/initial_eq.png))

Now, this optimization might simply overfit the equation if  $x_1, x_2, x_3$  (independent variables) are too many in numbers. Hence we introduce a new penalty term in our objective function to find the estimates of co-efficient. Following is the modification we make to the equation :

$$\min_{f \in H} \sum_{i=1}^n |Y_i - f(X_i)|^2 + \lambda \|f\|_H^2$$

(<http://i1.wp.com/www.analyticsvidhya.com/wp-content/uploads/2015/02/late.png>)

The new term in the equation is the sum of squares of the coefficients (except the bias term) multiplied by the parameter lambda. Lambda = 0 is a super over-fit scenario and Lambda = Infinity brings down the problem to just single mean estimation. Optimizing Lambda is the task we need to solve looking at the trade-off between the prediction accuracy of training sample and prediction accuracy of the hold out sample.

## Understanding Regularization Mathematically

There are multiple ways to find the coefficients for a linear regression model. One of the widely used method is gradient descent. Gradient descent is an iterative method which takes some initial guess on coefficients and then tries to converge such that the objective function is minimized. Hence we work with partial derivatives on the coefficients. Without getting into much details of the derivation, here I will put down the final iteration equation :

$$\theta_j := \theta_j - \alpha \frac{1}{m} \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)}) x_j^{(i)}$$

(<http://i1.wp.com/www.analyticsvidhya.com/wp-content/uploads/2015/02/gradient.png>)

Here, theta are the estimates of the coefficients. Alpha is the learning parameter which will guide our estimates to convergence. Now let's bring in our cost terms. After taking the derivative of coefficient square, it reduces down to a linear term. Following is the final iteration equation you get after embedding the penalty/cost term.

$$\theta_j := \theta_j (1 - \alpha \frac{\lambda}{m}) - \alpha \frac{1}{m} \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)}) x_j^{(i)}$$

([http://i0.wp.com/www.analyticsvidhya.com/wp-content/uploads/2015/02/gradient\\_mod.png](http://i0.wp.com/www.analyticsvidhya.com/wp-content/uploads/2015/02/gradient_mod.png))

Now if you look carefully to the equation, the starting point of every theta iteration is slightly lesser than the previous value of theta. This is the only difference between the normal gradient descent and the gradient descent regularized. This tries to find converged value of theta which

is as low as possible.

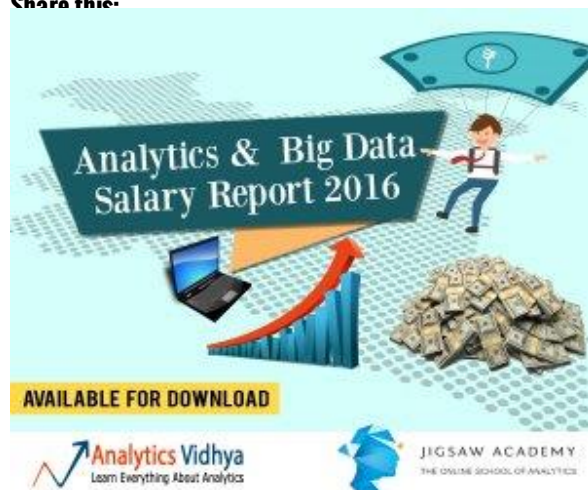
## End Notes

In this article we got a general understanding of regularization. In reality the concept is much deeper than this. In few of the coming articles we will explain different types of regularization techniques i.e. L1 regularization, L2 regularization etc. Stay Tuned!

Did you find the article useful? Have you used regularization to avoid over-fit before? Share with us any such experiences. Do let us know your thoughts about this article in the box below.

**If you like what you just read & want to continue your analytics learning, subscribe to our emails (<http://feedburner.google.com/fb/a/mailverify?uri=analyticsvidhya>), follow us on twitter (<http://twitter.com/analyticsvidhya>) or like our facebook page (<http://facebook.com/analyticsvidhya>).**

Share this:



</avoid-over-fitting-regularization/?share=linkedin&nb=1> 135

</avoid-over-fitting-regularization/?share=facebook&nb=1> 152

</avoid-over-fitting-regularization/?share=google-plus-1&nb=1>

</avoid-over-fitting-regularization/?share=twitter&nb=1>

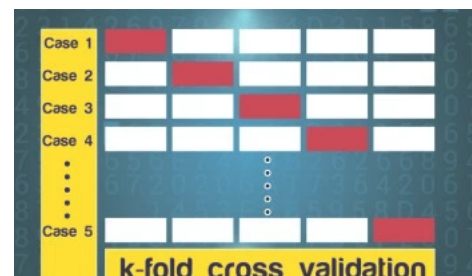
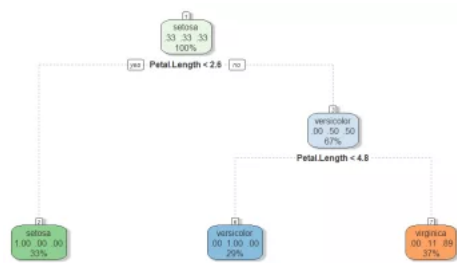
</avoid-over-fitting-regularization/?share=pocket&nb=1>

</avoid-over-fitting-regularization/?share=reddit&nb=1>

(<http://www.analyticsvidhya.com/blog/2016/02/quick-insights-analytics-big-data-salary-report-2016/>)

## RELATED





(<http://www.analyticsvidhya.com/blog/2014/06/comparing-cart-random-forest-1/>)

Comparing a CART model to Random Forest (Part 1)

(<http://www.analyticsvidhya.com/blog/2014/06/comparing-cart-random-forest-1/>)

In "Business Analytics"

(<http://www.analyticsvidhya.com/blog/2014/04/survival-analysis-model-you/>)

Is survival analysis the right model for you?

(<http://www.analyticsvidhya.com/blog/2014/04/survival-analysis-model-you/>)

In "Big data"

(<http://www.analyticsvidhya.com/blog/2015/05/k-fold-cross-validation-simple/>)

k-Fold Cross Validation made simple

(<http://www.analyticsvidhya.com/blog/2015/05/k-fold-cross-validation-simple/>)

In "Business Analytics"

TAGS: ANALYTICS ([HTTP://WWW.ANALYTICSVIDHYA.COM/BLOG/TAG/ANALYTICS/](http://www.analyticsvidhya.com/blog/tag/analytics/)), CROSS-VALIDATION

([HTTP://WWW.ANALYTICSVIDHYA.COM/BLOG/TAG/CROSS-VALIDATION/](http://www.analyticsvidhya.com/blog/tag/cross-validation/)), LINEAR-REGRESSION ([HTTP://WWW.ANALYTICSVIDHYA.COM/BLOG/TAG/LINEAR-REGRESSION/](http://www.analyticsvidhya.com/blog/tag/linear-regression/)), OVERFIT ([HTTP://WWW.ANALYTICSVIDHYA.COM/BLOG/TAG/OVERFIT/](http://www.analyticsvidhya.com/blog/tag/overfit/)), PRUNING

([HTTP://WWW.ANALYTICSVIDHYA.COM/BLOG/TAG/PRUNING/](http://www.analyticsvidhya.com/blog/tag/pruning/)), REGULARIZATION ([HTTP://WWW.ANALYTICSVIDHYA.COM/BLOG/TAG/REGULARIZATION/](http://www.analyticsvidhya.com/blog/tag/regularization/)), UNDERFIT ([HTTP://WWW.ANALYTICSVIDHYA.COM/BLOG/TAG/UNDERFIT/](http://www.analyticsvidhya.com/blog/tag/underfit/))

#### Previous Article

**Apprentice Leader - Mu Sigma - Bangalore ( 3-10 years of experience)**  
(<http://www.analyticsvidhya.com/blog/2015/02/apprentice-leaders-mu-sigma-bangalore-3-10-years-experience/>)

#### Next Article

**SAS Global Forum. The Kay Bailey Hutchison Convention Center. Dallas. TX. Apr 26 - 29**  
(<http://www.analyticsvidhya.com/blog/2015/02/sas-global-forum-kay-bailey-hutchison-convention-center-dallas-tx-apr-26-29/>)



(<http://www.analyticsvidhya.com/blog/author/tavish1/>)

Author

**Tavish Srivastava**

**(<http://www.analyticsvidhya.com/blog/author/tavish1/>)**

I am Tavish Srivastava, a post graduate from IIT Madras in Mechanical Engineering. I have more than two years of work experience in Analytics. My experience ranges from hands on analytics in a developing country like India to convince banking partners with analytical solution in matured market like US. For last two and a half years I have contributed to various sales strategies, marketing strategies and Recruitment strategies in both Insurance and Banking industry.

## 3 COMMENTS

---





Comment

Name (required)

Email (required)



Website

☐ Notify me of follow-up comments by email.

☐ Notify me of new posts by email.

**SUBMIT COMMENT**

## TOP USERS

Rank	Name	Points
1	 Nalin Pasricha ( <a href="http://datahack.analyticsvidhya.com/user/profile/Nalin">http://datahack.analyticsvidhya.com/user/profile/Nalin</a> )	3478
2	 SRK ( <a href="http://datahack.analyticsvidhya.com/user/profile/SRK">http://datahack.analyticsvidhya.com/user/profile/SRK</a> )	3364
3	Aayushmnit ( <a href="http://datahack.analyticsvidhya.com/user/profile/aayushmnit">http://datahack.analyticsvidhya.com/user/profile/aayushmnit</a> )	3075



4

binga (<http://datahack.analyticsvidhya.com/user/profile/binga>)

2623

5

vikash (<http://datahack.analyticsvidhya.com/user/profile/vikash>)

2190

[More Rankings \(http://datahack.analyticsvidhya.com/users\)](http://datahack.analyticsvidhya.com/users)

**TRANSFORM**  
YOUR CAREER

**Enrol to India's No.1 Analytics Course**
<http://pgpba.greatlakes.edu.in/?>

- Chosen by **500 +** professionals
- **1,75,000+** learning hours delivered
- Designed for **working professionals**.  
LEARN while you WORK


[utm\\_source=AVM&utm\\_medium=Banner&utm\\_campaign=Pgpba\\_decjan](http://pgpba.greatlakes.edu.in/?utm_source=AVM&utm_medium=Banner&utm_campaign=Pgpba_decjan)

## POPULAR POSTS

- Free Must Read Books on Statistics & Mathematics for Data Science  
(<http://www.analyticsvidhya.com/blog/2016/02/free-read-books-statistics-mathematics-data-science/>)
- A Complete Tutorial to Learn Data Science with Python from Scratch  
(<http://www.analyticsvidhya.com/blog/2016/01/complete-tutorial-learn-data-science-python-scratch-2/>)
- Essentials of Machine Learning Algorithms (with Python and R Codes)  
(<http://www.analyticsvidhya.com/blog/2015/08/common-machine-learning-algorithms/>)

- A Complete Tutorial on Time Series Modeling in R  
(<http://www.analyticsvidhya.com/blog/2015/12/complete-tutorial-time-series-modeling/>)
- Complete guide to create a Time Series Forecast (with Codes in Python)  
(<http://www.analyticsvidhya.com/blog/2016/02/time-series-forecasting-codes-python/>)
- 4 tricky SAS questions commonly asked in interview  
(<http://www.analyticsvidhya.com/blog/2013/11/4-sas-tricky-analytics-interview/>)
- SAS vs. R (vs. Python) – which tool should I learn?  
(<http://www.analyticsvidhya.com/blog/2014/03/sas-vs-vs-python-tool-learn/>)
- 7 Important Model Evaluation Error Metrics Everyone should know  
(<http://www.analyticsvidhya.com/blog/2016/02/7-important-model-evaluation-error-metrics/>)



**S,A,P BI Training**

S,A,P Business Intelligence -  
Realtime Training Centre in Chennai

[www.vicaprilabs.com](http://www.vicaprilabs.com)



([http://imarticus.org/programs/business-analytics-](http://imarticus.org/programs/business-analytics-professional/)

professional/)

## RECENT POSTS

---



(<http://www.analyticsvidhya.com/blog/2016/02/complete-tutorial-learn-data-science-scratch/>)

### **A Complete Tutorial to learn Data Science in R from Scratch**

(<http://www.analyticsvidhya.com/blog/2016/02/complete-tutorial-learn-data-science-scratch/>)

MANISH SARASWAT , FEBRUARY 28, 2016



(<http://www.analyticsvidhya.com/blog/2016/02/guide-build-predictive-models-segmentation/>)

### **Guide to Build Better Predictive Models using Segmentation**

(<http://www.analyticsvidhya.com/blog/2016/02/guide-build-predictive-models-segmentation/>)

GUEST BLOG , FEBRUARY 26, 2016



(<http://www.analyticsvidhya.com/blog/2016/02/quick-insights-analytics-big-data-salary-report-2016/>)

### **Quick Insights: India Analytics and Big Data Salary Report 2016**

(<http://www.analyticsvidhya.com/blog/2016/02/quick-insights-analytics-big-data-salary-report-2016/>)

KUNAL JAIN , FEBRUARY 24, 2016



(<http://www.analyticsvidhya.com/blog/2016/02/analytics-big-data-salary-report-2016/>)

### **India Exclusive: Analytics and Big Data Salary Report 2016**

(<http://www.analyticsvidhya.com/blog/2016/02/analytics-big-data-salary-report-2016/>)

KUNAL JAIN , FEBRUARY 22, 2016



([http://www.edvancer.in/certified-business-analytics?](http://www.edvancer.in/certified-business-analytics?utm_source=AV&utm_medium=AVads&utm_campaign=AVads1&utm_content=cbapavad)

[utm\\_source=AV&utm\\_medium=AVads&utm\\_campaign=AVads1&utm\\_content=cbapavad](http://www.edvancer.in/certified-business-analytics?utm_source=AV&utm_medium=AVads&utm_campaign=AVads1&utm_content=cbapavad))

## GET CONNECTED

---



**4,159**

FOLLOWERS

(<http://www.twitter.com/analyticsvidhya>)



**915**

FOLLOWERS

(<https://plus.google.com/+Analyticsvidhya>)



**11,950**

FOLLOWERS

(<http://www.facebook.com/Analyticsvidhya>)



**Email**

SUBSCRIBE

(<http://feedburner.google.com/fb/a/mailverify?uri=analyticsvidhya>)

## ABOUT US

---

For those of you, who are wondering what is "Analytics Vidhya", "Analytics" can be defined as the science of extracting insights from raw data. The spectrum of analytics starts from capturing data and evolves into using insights / trends from this data to make informed decisions.

---

## STAY CONNECTED

---



**4,159**

FOLLOWERS

(<http://www.twitter.com/analyticsvidhya>)



**915**

FOLLOWERS

(<https://plus.google.com/+Analyticsvidhya>)



**11,950**

FOLLOWERS

(<http://www.facebook.com/Analyticsvidhya>)



**Email**

SUBSCRIBE

(<http://feedburner.google.com/fb/a/mailverify?uri=analyticsvidhya>)

## LATEST POSTS

---



([http://www.analyticsvidhya.com/blog/2016/02/complete-tutorial-learn-data-](http://www.analyticsvidhya.com/blog/2016/02/complete-tutorial-learn-data-science-scratch/)

[science-scratch/](http://www.analyticsvidhya.com/blog/2016/02/complete-tutorial-learn-data-science-scratch/))

### **A Complete Tutorial to learn Data Science in R from Scratch**

(<http://www.analyticsvidhya.com/blog/2016/02/complete-tutorial-learn-data-science-scratch/>)

MANISH SARASWAT , FEBRUARY 28, 2016



([http://www.analyticsvidhya.com/blog/2016/02/guide-build-predictive-models-](http://www.analyticsvidhya.com/blog/2016/02/guide-build-predictive-models-segmentation/)

[segmentation/](http://www.analyticsvidhya.com/blog/2016/02/guide-build-predictive-models-segmentation/))

### **Guide to Build Better Predictive Models using Segmentation**

(<http://www.analyticsvidhya.com/blog/2016/02/guide-build-predictive-models-segmentation/>)

GUEST BLOG , FEBRUARY 26, 2016



([http://www.analyticsvidhya.com/blog/2016/02/quick-insights-analytics-big-data-](http://www.analyticsvidhya.com/blog/2016/02/quick-insights-analytics-big-data-salary-report-2016/)

[salary-report-2016/](http://www.analyticsvidhya.com/blog/2016/02/quick-insights-analytics-big-data-salary-report-2016/))

### **Quick Insights: India Analytics and Big Data Salary Report 2016**

(<http://www.analyticsvidhya.com/blog/2016/02/quick-insights-analytics-big-data-salary-report-2016/>)

KUNAL JAIN , FEBRUARY 24, 2016

### **India Exclusive: Analytics and Big Data Salary Report 2016**





2016/)

([http://www.analyticsvidhya.com/blog/2016/02/analytics-big-data-salary-report-](http://www.analyticsvidhya.com/blog/2016/02/analytics-big-data-salary-report-2016/)

**(<http://www.analyticsvidhya.com/blog/2016/02/analytics-big-data-salary-report-2016/>)**

KUNAL JAIN , FEBRUARY 22, 2016

## QUICK LINKS

---

Home (<http://www.analyticsvidhya.com/>)

About Us (<http://www.analyticsvidhya.com/about-me/>)

Our team (<http://www.analyticsvidhya.com/about-me/team/>)

Privacy Policy  
(<http://www.analyticsvidhya.com/privacy-policy/>)

Refund Policy  
(<http://www.analyticsvidhya.com/refund-policy/>)

Terms of Use  
(<http://www.analyticsvidhya.com/terms/>)

---

## TOP REVIEWS

---

---

© Copyright 2015 Analytics Vidhya