L1 Regularization · Regularization L2 · Overfitting (modeling) · Regression (statistics) · Data Science · +2

What is the difference between L1 and L2 regularization? How does it solve the problem of overfitting? Which regularizer to use and when?

Answer · Follow · 608 · Request · 1

29 Answers

Manjari Narayan, Ph.D. Electrical Engineering & Statistics, Rice University (2015)

Updated November 16, 2014 · Upvoted by Anant Raj, Ph.D. Machine Learning & Mathematical Optimization, Max Planck Institute for Intelligent Systems (2020) and Justin Rising, MSE in CS, PhD in Statistics

Originally Answered: What is the difference between L1 and L2 regularization?

Justin Solomon has a great answer on the difference between L1 and L2 norms and the implications for regularization.

***ℓ1* vs *ℓ2* for signal estimation:**

Here is what a signal that is sparse or approximately sparse i.e. that belongs to the ell-1 ball looks like. It becomes extremely unlikely that an ***ℓ2*** penalty can recover a sparse signal since very few solutions of such a cost function are truly sparse. ***ℓ1*** penalties on the other hand are great for recovering truly sparse signals, as they are computationally tractable but still capable of recovering the exact sparse solution. ***ℓ2*** penalization is pref... (more)

247 · 6

6 comments from Aleks Jakulin and more

YunFang Juan, Applied Machine Learning to Yahoo! Search and Facebook Ads.

Updated September 29, 2013 · Upvoted by Alberto Bietti, PhD student in machine learning. Former ML engineer and Benjamin Golub, PhD in Economics from Stanford; I work in applied probability on statistical problems related to social netwo...

Originally Answered: What is the difference between L1 and L2 regularization?

Practically, I think the biggest reasons for regularization are 1) to avoid overfitting by not generating high coefficients for predictors that are sparse. 2) to stabilize the estimates especially when there's collinearity in the data.

1) is inherent in the regularization framework. Since there are two forces pulling each other in the objective function, if there's no meaningful loss reduction, the increased penalty from the regularization term wouldn't improve the overall objective function.

Related Questions

- Why is L1-regularization better than L2-regularization, provided that all Norms are...
- What is an intuitive explanation of regularization?
- What is the advantage of combining L2 and L1 regularizations?
- When is Ridge regression favorable over Lasso regression?
- Why is L1 regularization supposed to lead to sparsity than L2?
- Are there any criteria to distinguish overfitting?

Ask question

Dima Korolev, big data guy.
Answered May 6, 2017

Originally Answered: What is the difference between L1 and L2 regularization?

There's a simple and important point that most answers don't emphasize (although the Kenneth Tran's one does touch on it).

The L1 regularization zeroes out many per-features weights.

The L2 regularization results in plenty of relatively small but nonzero ones.

Thus, if reducing the feature space is important, there's a compelling reason to pick the L1 one over the L2 one.

Natural examples of where it could be valuable to reduce the feature space could be:

- when it's important to maximize the QPS of applying the model,
- when it's important to minimize the size of the data file with feature values which

... (more)

Related Questions More Answers Below

Why is L1-regularization better than L2-regularization, provided that all Norms are equivalent?

What is an intuitive explanation of regularization?

What is the advantage of combining L2 and L1 regularizations?

Leihua Ye, Ph.D. Researcher, Ph.D. Data Science & Political Science, University of California, Santa Barbara (2020)
Answered January 15

- **Part 1: What They Are**

LASSO regression, L1 regularization, includes a hyper-parameter α times the sum of the absolute value of the coefficients as penalty term in its cost function, shown below (marked in red):

$$\sum_{i=1}^{n}\left(Y_i - \hat{b}_0 - \hat{b}_1 X_i - \cdots \hat{b} X_p\right)^2 + \alpha\left(|\hat{b}_1| + \cdots + |\hat{b}_p|\right)$$

On the one hand, if we do not apply any penalty (set α =0), the above formula turns into a regular OLS regression, which may overfit.

On the other hand, the model will probably underfit if we apply a very large penalty (or, a large α value), because we have falsely penalized all coefficients (the most important ones included).

Ridge regression adopts a "squared magnitude" of coefficient tim... (more)

Add a comment... Add Comment

Deepu Raveendran, Passion for Machine Learning and Statistics
Answered August 14, 2015

Originally Answered: What is the difference between L1 and L2 regularization?

L1 and L2 regularization are used to avoid overfitting of data. L1 and L2 penalized estimation methods shrink the estimates of the regression coefficients towards zero relative to the maximum likelihood estimates. The purpose of this shrinkage is to prevent overfit arising due to either collinearity of the covariates or high-dimensionality. Although both methods are shrinkage methods, the effects of L1 and L2 penalization are quite different in practice. Applying an L2 penalty tends to result in all small but non-zero regression coefficients, whereas applying an L1 penalty tends to result in m... (more)

9 5

5 comments from Youssef Kashef and more

Harsh Gupta, CEO at protonAutoML (2019-present)
Answered 17h ago

L1 Regularization -

L1 or lasso regularizer is a regularization technique that requires us to minimize the sum of absolute values between features and target variable. The word "absolute" target, and other smaller values within the same feature will be ignored(or removed) when training models using this model.

13

The cost function as follows: Where C (λ) is penalty parameter λ; W i are the weights of each feature-target pair; x i are input/output data for each sample; β 0 is a constant term for intercept . Also, not... (more)

Add a comment... Add Comment

7

Regularization is a process that helps to solve the problem of over-fitting by introducing in the model with fewer rules. Regularization introduces a cost term for including more features and will try to reduce the weights of many features in your model to zero or very small values.

In L1 regularization, the weights for each parameter are assigned as a 0 or 1 (binary value). Whereas in L2 regularization the error is spread across the feature and the resultant weights for every feature has the possibility that some weights are really small values close to 0.

Overfitting is caused by several reaso... (more)

Add a comment... Add Comment

358 10

machine learning and Yang Yao, M.S. Statistics, Stanford University (2018)

Originally Answered: What is the difference between L1 and L2 regularization?

There are already a couple of pretty good answers that explain L1, L2 regularization from a theoretical perspective. In my answer I am going to try to complement them from a purely practical point of view. And, in doing so I am going to somewhat disagree with Andrew Ng.

New message

Messages

No messages

Connect with others on Quora by beginning a new

some situations. However, beyond that particular reason I have never seen L1 to perform better than L2 in practice. And, to be clear, I don't think I am the o... (more)

10 comments from Manjari Narayan and more

71 1 1

Originally Answered: What is the difference between L1 and L2 regularization?

Why regularize machine learning models?

Lots of good answers here. For example, YunFang Juan mentions how regularization helps avoid overfitting as well as the potential of learning extremely large model parameters. Aleks Jakulin discusses regularization from a Bayesian perspective where it is interpreted as adding prior information about the distribution of parameters to the model likelihood. And, Justin Solomon in his response uses regularization as a way of finding (unique) solutions to underdetermined systems of equations. These are all very useful ways of thinking about regularization, and... (more)

1 comment from Jo Helmuth

425 1 12

Machine Learning, University College London (2018) and Carl Shan, data scientist, 2014 Eric Schmidt Data Science Fellow

Originally Answered: What is the difference between L1 and L2 regularization?

There are many ways to understand the need for and approaches to regularization. I won't attempt to summarize the ideas here, but you should explore statistics or machine learning literature to get a high-level view. In particular, you can view regularization as a prior on the distribution from which your data is drawn (most famously Gaussian for least-squares), as a way to punish high values in regression coefficients, and so on. I prefer a more naive but somewhat more understandable (for me!) viewpoint.

Let's say you wish to solve the linear problem . Here, is a matrix and is a ve... (more)

12 comments from Manjari Narayan and more

103 1 2

PhD stats and Don van der Drift, In PhD Physics program for 2.5 years at Technische Universiteit Eindhoven, with a focus on statistics and dat...

Originally Answered: What is the difference between L1 and L2 regularization?

The main difference between L1 and L2 regularization is that L1 can yield sparse models while L2 doesn't. Sparse model is a great property to have when dealing with high-dimensional data, for at least 2 reasons.

- Model compression: increasingly important due to the mobile growth
- Feature selection: it helps to know which features are important and which features are not or redundant.

But why does L1 yield sparse models? Despite earlier attempts [1], here's a another simple way to look at it [2].

For simplicity, let's just consider the 1-dimensional case.

-regularized loss function
... (more)

New message

Messages

No messages

Connect with others on Quora by beginning a new

30 1

Answered February 11, 2015

Originally Answered: What is the difference between L1 and L2 regularization?

There are already excellent answers discussing the sparsity of model due to l1 norm vs. l2 norm. Here I'm trying to zoom from a different angle: model robustness.

Here, model robustness refers to how robust model can handle data corruptions in prediction/scoring. In other words, training data and test data distributions are slightly different. This can happen due to subtle temporal change (say, model trained yesterday and being applied to today's data), or one feature is warped or missing due to data ETL or one bug in the system.

Consider an extreme case that you are dealing with a classificat… (more)

1 comment from Mohsen Molaei

79 4

Originally Answered: What is the difference between L1 and L2 regularization?

How to prevent overfitting
Both L1 and L2 regularization prevents overfitting by shrinking (imposing a penalty) on the coefficients.

Difference between L1 and L2
L2 (Ridge) shrinks all the coefficient by the same proportions but eliminates none, while L1 (Lasso) can shrink some coefficients to zero, performing variable selection.

Which to use
If all the features are correlated with the label, ridge outperforms lasso, as the coefficients are never zero in ridge. If only a subset of features are correlated with the label, lasso outperforms ridge as in lasso model some coefficient can be shrunken … (more)

4 comments from Håkon Hapnes Strand and more

4 1

Originally Answered: What is the difference between L1 and L2 regularization?

The L1 Norm provides a near optimal sparse solution when the underlying signal /data is sparse in some (say overcomplete) basis and the signal to noise ratio (SNR) is high

The L2 Norm is suitable for non-sparse solutions and/or bandwidth limited signals

For example, a face is sparse in a Wavelet / Harr basis; this is the basis of the JPEG 2000 algorithm

A document, however, is generally not that sparse in the bag of words representation, and L2 methods can work very well here

Notice I did not say anything about the noise distribution. Also, the optimal sparse solution is the L0 norm, not the L1 no… (more)

1 comment from Venson Wang

3

The difference between L1 and L2 regularization are as follows:

coefficients more than L2/Gaussian

.

L1 can yield sparse models while L2 doesn't

.

L1 and L2 regularization prevents overfitting by shrinking on the coefficients

.

L2 (Ridge) shrinks all the coefficient by the same proportions but eliminates none, while L1 (Lasso) can shrink some coefficients to zero, performing variable selection

.

L1 is the first moment norm |x1-x2| that is simply the absolute dıstance between two points where L2 is second moment norm corresp... (more)

Add a comment... Add Comment

17 1

Originally Answered: What is the difference between L1 and L2 regularization?

Let me also suggest reading this paper by Andrew Ng. It basically says that for a number of methods, if you expect to have a high number of irrelevant features, you will need fewer training examples to generalize well with L1 compared to L2

http://www-cs.stanford.edu/people/ang/papers/icml04-l1l2.pdf

Abstract:

We consider supervised learning in the presence of very many irrelevant features, and study two different regularization methods for preventing overfitting. Focusing on logistic regression, we show that using L1 regularization of the parameters, the sample complexity (i.e., the number of... (more)

1 comment from Eren Golge

23

Originally Answered: What is the difference between L1 and L2 regularization?

L1 is the first moment norm |x1-x2| (|w| for regularization case) that is simply the absolute dıstance between two points where L2 is second moment norm corresponding to Eucledian Distance that is |x1-x2|^2 (|w|^2 for regularization case) .

Another big difference is L1 is not differentiable thus for finding weigths that minimize your error function, you cannot use gradient based approaches. L2 give you a regularization term that is differentiable so it is suitable to be used with gradient descent.

you may find some more from this slide :http://cs.nyu.edu/~rostami/presentations/L1_vs_L2.pdf

21.3K views · View 23 upvotes

Add a comment... Add Comment

New message

Messages

No messages

Connect with others on Quora by beginning a new