

Q & A

Ruichen Wang

February 1, 2019

Abstract

Some basic questions worth thinking.

why L1 regulation generates sparsity? L2 regulation cause blur?

Firstly, why do we want the result matrix to be sparse?

Consider 1 million dimension, calculate the inner product between w and x need a lot of computation. If the w can be sparse, the inner product will only be performed on the non-zero columns.

Or consider another situation, in some scenario, there are free data and many features, which is often called as ‘**small n, large p problem**’. If $n \ll p$, then our model will be very complex, our w will be a singular matrix ($|w| = 0$). In other words, **overfitting**.

One way to control overfitting is adding a regularization term to the loss function. Ridge (l_2norm) and LASSO (l_1norm) regression are two very common regression ways.

$$J(w) = Loss(x) + \lambda ||w||_2^2$$

$$J(w) = Loss(x) + \lambda ||w||_1$$

Those fomula can be denoted as :

Back to the problem, intuitivly, the target loss will alway intersect at the coordinate axis when using l1 norm. Imaging high dimension situation, the angles will certainly more likely to be intersected, while the ball will not.

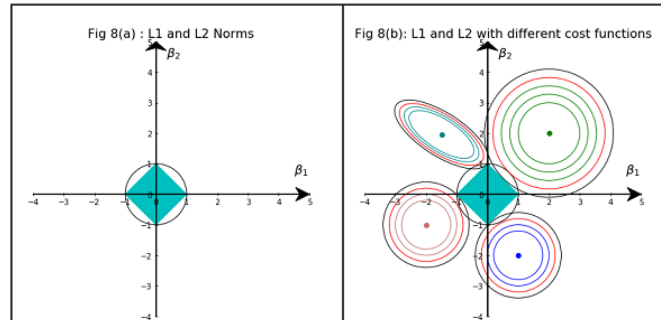


Figure 1: L1 and L2 norm.

why L2 regulation cause blur?

one-hot encoding for gbd?

xgboost vs gbd?

xgboost vs gbd prevent overfitting

bagging vs boosting

xgboost vs lightgbm

xgb rf lr difference

user-cf item-cf difference and application scenarios

svm vs lr

lstm vs gru