Q & A

Ruichen Wang

February 13, 2019

Abstract

Some basic questions worth thinking.

1. 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. Rigde ($l_2 norm$) and LASSO ($l_1 norm$) regression are two very common regression ways.

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

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

Assume we use loss using MSE, the target function can also be denoted as :

$$\min_w \frac{1}{n} ||y - Xw||^2 \quad s.t. \lambda ||w||_2^2 \leq C$$

$$\min_w \frac{1}{n} ||y - Xw||^2 \quad s.t. \lambda ||w||_1 \leq C$$

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.

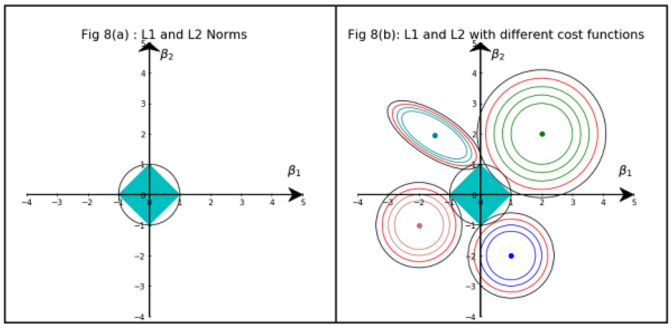

Figure 1: L1 and L2 norm.

For more math proof, see http://freemind.pluskid.org/machine-learning/sparsity-and-some-basics-of-l1-regularization

2. Why L2 regulation cause blur?

In generative models, eg.VAE, L2 norm / L2 loss / MSE tend to yield blurry images. We try to explain this in probabilistic settings. In Gaussian distribution, it defines as :

$$p(x|\mu,\sigma^2) = \frac{1}{Z} exp\left(-\frac{||\mu - x||^2}{2\sigma^2}\right)$$

$$logp(x|\mu,\sigma^2) \propto exp\left(-\frac{1}{2}||x_\mu - x||^2\right)$$

As we can see, minimizing MSE, is same as maximizing the log likelihood of gaussian, we make the assumption that our x comes from a gaussian.

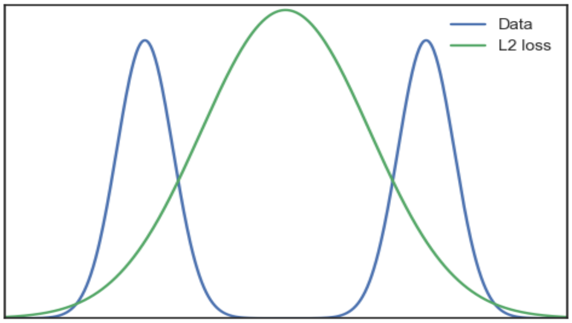

Figure 2: Gaussian and multinomial.

In reality, there often many hidden variables (multinomial) controls the x. A simple example, we have white and black dogs as dataset x. maximizing the likelihood will blur the two and generate gray dogs.

3. One-hot encoding for gbdt?

xgboost vs gbdt?

xgboost vs gbdt 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