Regularization in Deep Neural Networks

Paul Stey

March 3, 2017

Table of Contents

What is Regularization?

1. Method of calibrating the complexity of our model
2. Very general (i.e., applicable to variety of models)
3. Penalized regression
    1. Ridge regression
    2. Lasso
    3. Elastic net
4. XGBoost improvement on AdaBoost

L_2 Regularization Regression

1. Ridge regression, also called Tikhonov regularization (Tikhonov, 1963)
2. Penalize the L_2 norm
3. Constrains the Euclidian distance of β

Given the linear model

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip} + \epsilon_i$$

the ridge penalty constrains

$$\|\beta\|_2 = \sqrt{\beta_1^2 + \beta_2^2 + ... + \beta_p^2}$$

L_2 Regularization cont.

Recall the least squares solution is obtained by

$$\widehat{\beta} = \left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1} \boldsymbol{X}^T y.$$

The ridge penalty gives us

$$\widehat{\beta}^{ridge} = \left(\boldsymbol{X}^T \boldsymbol{X} + \lambda \boldsymbol{I}\right)^{-1} \boldsymbol{X}^T y$$

where λ is a penalty term, and $\boldsymbol{I}$ is the $p \times p$ identity matrix.

L_2 Regularization cont.

1. Penalty λ is in the interval $[0, \infty)$
2. Shrinks β_j values towards 0 (and each other)
3. Benefits:
   1. Reduce model variance (i.e., more generalizable)
   2. Gain stability in estimates involving correlated data
   3. Computationally efficient
4. Drawbacks:
   1. Sacrifice unbiasedness of maximum likelihood estimate

L_1 Regularization

The lasso

1. Similar to ridge regression
2. Penalize L_1 norm of β
3. Constrains taxicab distance spanned by vector of regression coefficients

$$\|\beta\|_1 = |\beta_1| + |\beta_2| + ... + |\beta_p|$$

4. As λ increases, $\beta_j \to 0$
   1. β_j for less important predictors shrink faster
   2. Can be used for "variable selection"

Elastic Net

1. Penalize *both* the L_1 and L_2 norms of β
2. Advantage of combining strengths of both approaches

Caveat: Note that we often use ridge or lasso for very specific reason (i.e., ridge for accuracy, lasso for variable selection). One might suggest we give away this clarity of purpose when using the elastic net.

Regularization More Broadly

Recall that regularization is merely a method of controlling model complexity.

The standard shrinkage method examples make this very transparent with the use of explicit penalty terms.

Broader View cont.

Consider less obvious examples:

1. Number of boosting iterations
2. Learning rate for gradient descent
3. Bagging proportion
4. Proportion in training *vs.* test set