Regularization

Authors: Mervyn Giritharan, Calie Gilmore, Martina Galvan. (PDF)

This lecture covers the benefits of regularization to avoid overfitting the testing data set.

Regularization reduces the weights to fit the test set better, however a penalty is added. There can be many possible options to use regularization, such as Lasso and Ridge, so you need to use domain knowledge to decide which regularization method is preferred to constrain your solution. Ultimately, the goal for regularization is to avoid over-fitting to result in better predictions. When a model is over fitted, it's hyper adjusting to past data that doesn't allow us to accurately predict the future data.

The regularization process imposes a penalty. The cost of the penalty: Cost = MSE + λ Penalty(f) In this formula, the MSE is the mean squared error which is the average square of the errors. Lambda is the regularization parameter and we want to find the mabda that results in the lowest variance on the test set. The regularization penalty is a function that maps each function f onto a number. Note that the penalty is not applied to the y-intercept.

Ridge Regression

In Ridge regression, the regression is the least squares regression with the ridge regression penalty added to it.

The penalty in the cost function is the sum of the squared weights/coefficients.

$$p = \sum_{i=1}^{M} W_J^2$$

The lambda penalty regaularizes the coefficients and penalizes them if they are too large. The is the formula for the penalty being added to the error.

$$E(w) = \sum_{n=1}^{N} (w^T \phi(X_n) - t_n)^2 + \frac{\lambda}{2} ||w||^2$$

This can be called shrinkage (stats) or weight decay (neaural nets). The regularization coefficient λ now controls the effective model complexity, therefore

reducing the coefficients helps decrease model complexity. Ridge is also good because it stabilizes answers and helps adjust for collinearity. The penalty added by the ridge regression creates the bias needed on the training set in order to reduce the variance on the test set and have more accurate predictions in the long run.

Lasso Regression

Lasso regression is super similar to ridge regression, but the one difference is the ridge regression squares the variable weights for the penalty, whereas the lasso regression takes the absolute value of the variable weights for the penalty.

In Lasso regression, the penalty in the cost function uses the sum of the absolute values of weights.

$$p = \sum_{i=1}^{M} |W_J|$$

A second difference is that lasso regression takes magnitudes into account which means the increase in lambda can cause some coefficients to zero out. This helps us know which features can be removed from the model, as well as reduce the over-fitting. The ridge regression cannot zero out features because it can only shrink the slope asymptotically toward zero. So lasso regression is a little better than ridge regression in reducing the variance when there are many insignificant variables. However, ridge regression is a little better when many of the variables are proven useful.

When to Use Lasso vs. Ridge?

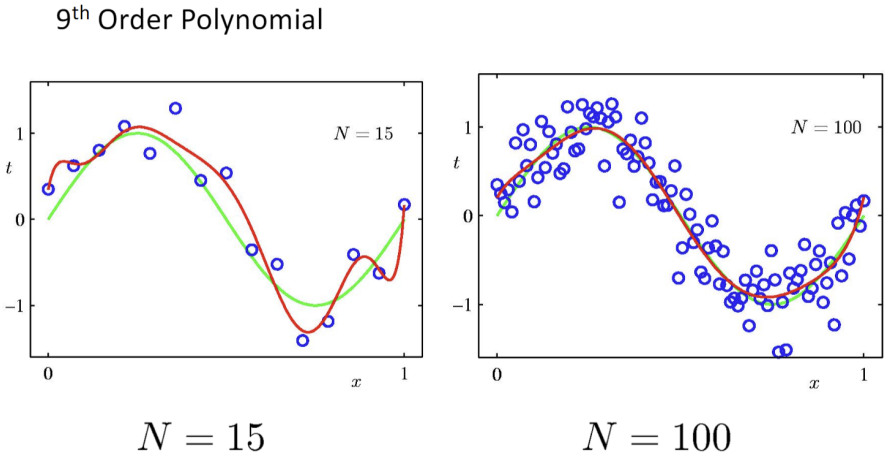
One way is to use both methods and see which one provides optimal results. However, that is an expensive option and often, you cannot perform both. So, it is good to know when each method is the optimal solution. There are some considerations to keep in mind: you want your model to be only as complicated as necessary but not too complicated, and it is important to keep in mind how much data you have. See the Data Set Size section below. It is also important to standardize your data prior to using Lasso and Ridge to compare the weights.

Lasso will help you determine which variables you want to ignore or drop from your model. It helps you to do this by more aggressively pushing your lower weights to 0. So, if you have a model will 10 different weights and you know that they are all useful, then lasso may not be the optimal method. However, if you have a model with 10 weights and you are not sure if they are useful then lasso would help in determining this. Lasso is more aggressive on weights that < 1.

Ridge more aggressively contains the higher weights, especially weights that are < 1. The Ridge regression discourages larger values by adding the penalty term to error. Here, the regularization coefficient, lambda, controls the effective model complexity. Ridge can be a good method to use because it stabilizes the

answers and helps with a collinearity problem. Why is this? If 2 variables are collinear, the determinant of the matrix is small, and the inverse of the matrix is large. This results in your weights varying a lot. When some variables are collinear, the lambda is going to lessen the uncertainty that is associated with the determinant of the matrix being small and the inverse therefore being very large. This results in a more stable answer and contains less variance.

Data Set Size



Regularization will help the N=15 chart more because it doesn't have a lot of data. It's harder to fit a model with limited size data sets and it can be more likely to overfit. Regularization allows complex models to be trained with small data sets without severely over-fitting because it adds a penalty to the model. The penalty introduces bias into model to combat overfitting and therefore reduces the variance in the test set. In contrast, regularization does not help the N=100 chart very much because it has a good amount of data and is less likely to need the penalty to combat overfitting.

Another note on data set size is that usually it's best practice to have at least as many data points as you have parameters. However, ridge regression can remove the requirement and allows you to have more parameters that data points. This is very useful in the practical world where producing very large data sets can be challenging.

More Resources

Ridge & Lasso Regularization Info: https://towardsdatascience.com/ridge-and-lasso-regression-a-complete-guide-with-python-scikit-learn-e20e34bcbf0b

Regularization Process & Benefits: https://www.microsoft.com/en-us/research/uploads/prod/2006/01/Bishop-Pattern-Recognition-and-Machine-

Learning-2006.pdf

Video on Regularization using Ridge Regression: https://www.youtube.com/watch?v=Q81RR3yKn30

Video on Regularization Difference between Ridge and Lasso Regressions: https://www.youtube.com/watch?v=NGf0voTMlcs&t=288s