ML Assignment 1

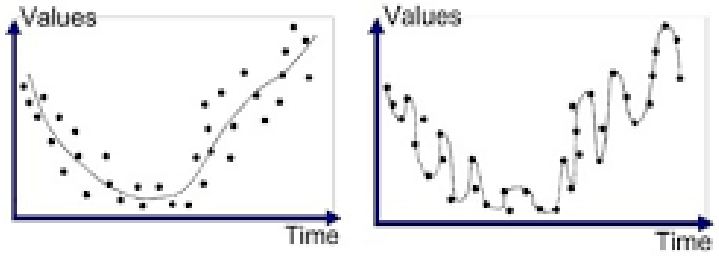
Arush Sharma 16BCE1127

L2 Regularization

Theory

What does Regularization achieve?

A standard least squares model tends to have some variance in it, i.e. this model won't generalize well for a data set different than its training data. *Regularization, significantly reduces the variance of the model, without substantial increase in its bias*. So, the tuning parameter λ, used in the regularization techniques described above, controls the impact on bias and variance. As the value of λ rises, it reduces the value of coefficients and thus reducing the variance. *Till a point, this increase in λ is beneficial as it is only reducing the variance (hence avoiding overfitting), without losing any important properties in the data*. But after certain value, the model starts losing important properties, giving rise to bias in the model and thus underfitting.



Good Fit/Robust Overfitted

We can see in the figure, when the model tries to fit each and every training sample, then overfitting happens. This usually happens when the model is too complex.

A regression model that uses L1 regularization technique is called *Lasso Regression* and model which uses L2 is called *Ridge Regression*.

Ridge regression adds "*squared magnitude*" of coefficient as penalty term to the loss function.

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}\beta_j^2=\mathrm{RSS}+\lambda\sum_{j=1}^{p}\beta_j^2$$

Above image shows ridge regression, where the *RSS is modified by adding the shrinkage quantity*. Now, the coefficients are estimated by minimizing this function. Here, *λ is the tuning parameter that decides how much we want to penalize the flexibility of our model*. The increase in flexibility of a model is represented by increase in its coefficients, and if we want to minimize the above function, then these coefficients need to be small. This is how the Ridge regression technique prevents coefficients from rising too high. Also, notice that we shrink the estimated association of each variable with the response, except the intercept β0, this intercept is a measure of the mean value of the response when xi1 = xi2 = ...= xip = 0.

When λ = 0, the penalty term has no effect, and the estimates produced by ridge regression will be equal to least squares. However, *as λ→∞, the impact of the shrinkage penalty grows, and the ridge regression coefficient estimates will approach zero*. As can be seen, selecting a good value of λ is critical. Cross validation comes in handy for this purpose. The coefficient estimates produced by this method are *also known as the L2 norm*.