Data-Driven Problem Solving in Mechanical Engineering

Regularization

Masoud Masoumi

ME 364 - Spring 2022

Department of Mechanical Engineering
The Cooper Union for the Advancement of Science and Art

March, 2022

Regularization

Regularization is any modification we make to a learning algorithm that is intended to reduce its generalization error but not its training error.

For example, we can modify the training criterion for linear regression to include **weight decay**.

To perform linear regression with weight decay, we minimize a sum comprising both the mean squared error on the training and a criterion $J(w)$ that expresses a preference for the weights to have smaller values

$$J(w) = MSE_{train} + \lambda w^T w$$

Remember for Linear Regression we had:

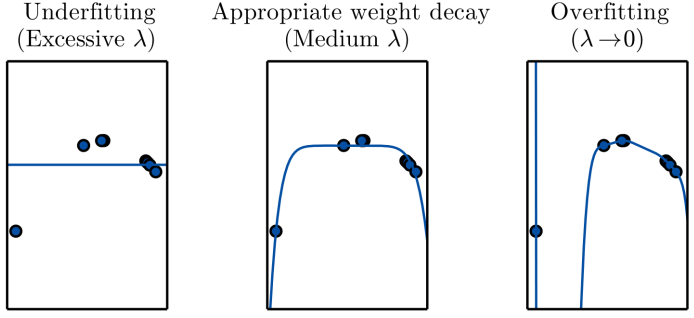
$$J(\mathrm{w}) = \frac{1}{2m} \sum_{i=1}^{m} \left(y^{(i)} - \hat{y}^{(i)} \right)^2$$

where λ is a value chosen ahead of time that controls the strength of our preference for smaller weights.

Regularization

When $\lambda = 0$, we impose no preference. Larger λ forces the weights to become smaller.

Minimizing $J(w)$ results in a choice of weights that make a trade-off between fitting the training data and being small. This gives us solutions that have a smaller slope, or that put weight on fewer of the features.



Ridge and Lasso Regression

In general, we can write the regularization term (penalty term) in the form of

$$\lambda \sum_{j=1}^{n-1} |w_j|^q$$

q = 1 → **lasso**
q = 2 → quadratic regularizer (previous slide). It is known as **ridge**.

If λ is sufficiently large, some of the coefficients w_j are driven to zero, leading to a *sparse* model. We use this property for **feature selection**.