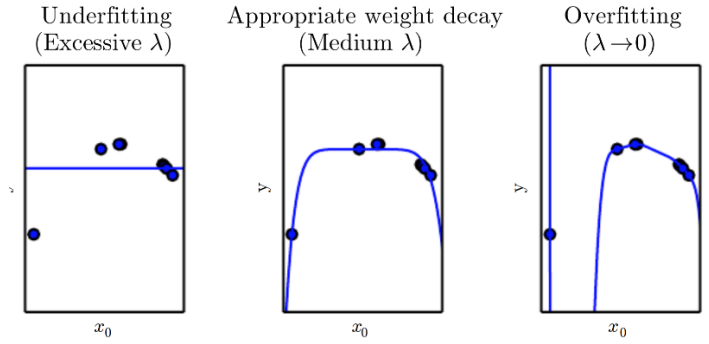
Weight Decay

Weight decay is a regularization technique that is used to regularize the size of the weights of certain parameters in machine learning models. Weight decay is most widely used regularization technique for parametric machine learning models. Weight decay is also known as L2 regularization, because it penalizes weights according to their L2 norm. In weight decay technique, the objective function of minimizing the prediction loss on the training data is replaced with the new objective function, minimizing the sum of the prediction loss and the penalty term. It involves adding a term to the objective function that is proportional to the sum of the squares of the weights. This is how the new loss function looks like using weight decay technique:

$$L(\mathbf{w}, b) + \frac{\lambda}{2}\|\mathbf{w}\|^2$$

In the above equation, L(w, b) represents the original loss function before adding the regularization L2 norm (weight decay) term. For the value of lambda as 0, the original loss function comes into picture. For the value of lambda greater than 0, the size of weight vector is selected appropriately to minimize the overall loss. If the weight vector grows too large, the learning algorithm focuses on minimizing the L2 weight norm vs. minimizing the training error.

Weight decay is typically used with stochastic gradient descent and can be applied to any differentiable loss function. Weight decay has been shown to improve the generalization performance of many types of machine learning models including deep neural networks. Weight decay is usually implemented by modifying the update rule for the weights such that the gradient is not only based on the training data but also on the weight decay term.



Some benefits of using weight decay include:

- Improved generalization performance: Weight decay is used to reduce the generalization error of a machine learning model by preventing overfitting.
- Reduced complexity: weight decay can reduce the complexity of a machine learning model by encouraging the learned representation to be sparser.
- Reduced overfitting: Weight decay helps reduce overfitting by penalizing large weights. This encourages the model to learn simpler functions that are less likely to overfit the training data.
- Improved convergence: weight decay can help improve the convergence of an optimization algorithm by encouraging smaller weights.
- Simpler functions are learned
- Deep neural networks perform better with weight decay
- Efficient training: Weight decay can improve the efficiency of training by reducing the amount of time required to converge on a solution.