MACHINE LEARNING
CS60050
Instructor: Aritra Hazra

Scribes By:
Abhishek Gandhi — 19CS10031
Abhilash Datta — 19CS30001

March 4, 2021

Contents

1 Recapitulation **1**

2 Regularization **2**

3 Augmented Error **4**

4 Weight Decay **4**

5 Weighted Regression **5**

6 Why do we use $\leq$ C rather than $\geq$ C? **5**

7 Optimal λ **6**

1 Recapitulation

Overfitting:

- **Stochastic Noise**: Data with Noise fitting
- **Deterministic Noise**: Higher Order Data fitting

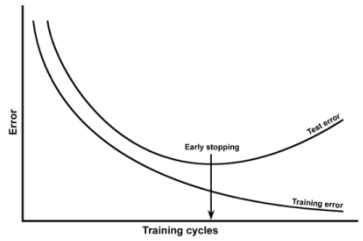

Figure 1: Variation of error with number of training examples

- Increase in Training data results in reduction in Overfitting.
- Increase in Noise levels results in increase in Overfitting.
- Increase in Target complexity results in increased in Overfitting.

- **Relation with VC-dimension:** Allows the generalization bound but it's so pessimistic that effective E_{out} (test error) will be much lesser than the given tolerance.
- **Remedies:**
  1. Regularization
  2. Validation

2 Regularization

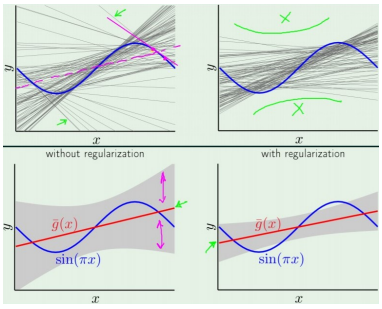

Figure 2: With and without regularization

Instead of abruptly allowing every possible fits we restrict them. Hence reducing **variance** and constraining our solution space.

Legendre Polynomials: Orthogonal in nature and less correlation between the terms. Examples: x, $(3x^2\text{-}1)/2$, etc.

We'll consider our Hypothesis space to be as a linear combination of legendre polynomials.

$$H_Q = \sum_{q=0}^{Q} W_q L_q(x)$$

$$\text{Z} = [\ 1, L_1(x), L_2(x), ..., L_n(x)\ ]$$

We'll transform the training set with respect to these polynomials so as to transform the training points into Z-space. $((z_1, y_1),(z_2, y_2), ...,(z_n, y_n))$. Depending upon the constraints, we can now make some W's as zeros and others as ones. The ones with the zeros are no longer considered in our hypothesis space, thus we can control complexity. Now Linear Regression will minimize the apparent error.

$$E_{in}(W) = \sum_{n=1}^{N} (W^T Z_n - y_n)/N$$

$$= (1/N)(Z \cdot W - Y)^T (Z \cdot W - Y)$$

$$W_{lin} = (Z^T Z)^{-1} Z^T Y$$

[Unconstrained Fitting of N points]

If we convert H_{10} to H_2, we are making $W_q = 0$, for all q greater than 2. This is a **Hard Constraint**.

Rather we can impose a **soft constraint** by

$$\sum_{q=0}^{Q} W_q^2 <= C$$

Now we have to minimize

$$E_{in}(W) = \sum_{n=1}^{N} (W^T Z_n - y_n)/N$$

subject to

$$W^T W <= C$$

This will give us a regularized solution

$$W_{reg}$$

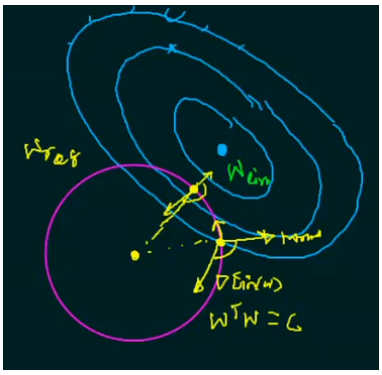

Figure 3: Constrains between Wreg and Win

$$\nabla E_{in}(W_{reg})\alpha - W_{reg} \tag{1}$$

$$\nabla E_{in}(W_{reg}) = -(2\lambda/N)W_{reg} \tag{2}$$

$$\nabla E_{in}(W_{reg}) + (2\lambda/N)W_{reg} = 0 \tag{3}$$

which is the equation we get by the minimization of

$$E_{in}(W) + (\lambda/N)W^T W \tag{4}$$

This bring to the conclusion that the subjected minimization and the latest equation are duals of each other.
C and λ are inversely proportional.
We have converted unconditional to conditional, We can apply VC dimension in unconditional, but unconditional will land up value of W in infinite space so it is easier to analyse VC Dimension in conditional and easier to solve in conditional hence they are both used conversely.
From this we can conclude that

$$If\ C \downarrow \Rightarrow \lambda \uparrow \Rightarrow We\ get\ E_{in}\ as\ solution$$

$$If\ C \uparrow \Rightarrow \lambda \downarrow \Rightarrow WW^T\ dominates$$

Hence no option, provision for **reducing the error**

3 Augmented Error

We are trying to minimize

$$E_{aug}(W) = E_{in}(W) + (\frac{\lambda}{N})W^TW = \frac{1}{N}[(ZW-Y)^T(ZW-Y) + \lambda W^TW] \tag{5}$$

In equation (5) both the part are quadratic hence we can solve it using **quadratic programming**

$$\begin{gathered}\nabla E_{in}(W) = 0 \Rightarrow Z^T(ZW - Y + \lambda W = 0) \\ W_{req} = (Z^TZ + \lambda I)^{-1}Z^TY \\ Opposed to W_{in} = (Z^T)^{-1}Z^TY\end{gathered} \tag{6}$$

This implies
If λ is large $\Rightarrow$ Constraining Solution that is Making $W_{reg} = 0$
More λ makes the **circle smaller** and **decreases** chances of getting to W_{in}
Makes the **curve flatter(smother)**
If λ is small $\Rightarrow$ Encapsulating $W_{in} inside of it that is making W_{reg} = W_{in}$

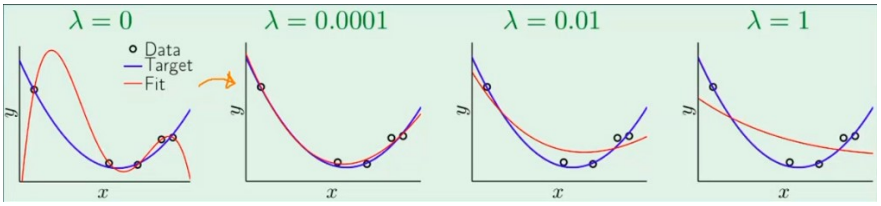

Figure 4: effects of λ values.

$$over - Fitting - - > Reg1 - - > Reg1^{'} ... - -- > ...under - Fitting$$

Choice of λ is very Important
We have to choose λ, **validate** it and choose out which is the best

4 Weight Decay

Instead of one step solution, We can use **gradient decent(Batch or stochastic)**
We try to update the parameter

$$\begin{gathered}W(t+1) = W(t) - \eta\nabla[E_{in}(W(t)) + \frac{2\lambda}{N}W(t)] \\ W(t+1) = W(t)[1 - \frac{2\eta\lambda}{N}] - \eta\nabla E_{in}(W(t))\end{gathered} \tag{7}$$

This decay means that instead of moving towards solution directly. It shrings a little bit and move in that direction so that going towards the target change in such a way that we do not **hover** randomly and **overfit**.
In Neural networks we can use

$$W^TW = \sum\sum\sum(W_{ij}^l)^2 \tag{8}$$

5 Weighted Regression

Instead of only constraining weights to C, we constrain them to

$$\sum \gamma_q W_q^2 \leq C \tag{9}$$

If $\gamma_q = 2^q \Rightarrow$ Fit low order Polynomial, Smooth Curve
If $\gamma_q = 2^{-q} \Rightarrow$ Fit high order Polynomial

6 Why do we use $\leq$ C rather than $\geq$ C?

The main reason is
Stochastic Noise ⇒ High Frequency
Deterministic Noise ⇒ Non Smooth
High order Legendre polynomial are usually non smooth and low level are smooth, In practical we go for low frequency, low order functions as low frequency cancels both high frequency Stochastic noise as well as Deterministic noise. This is called **OCCAM result**

Figure 5: Shows training error vs λ value and Weight Growth vs Lambda

After Regularisation,

$$E_{aug}(h) = E_{in}(h) + \frac{\lambda}{N}\Omega(h) \tag{10}$$

In our VC bound we have found

$$E_{out} \geq E_{in}(h) + \Omega(H) \tag{11}$$

Here 'H' denotes entire space while 'h' means training distribution

Here we can say that from above equations $E_{aug}(h)$ is **closer** to E_{out} than E_{in}. Therefore, we can say that E_{aug} is better than E_{in}.

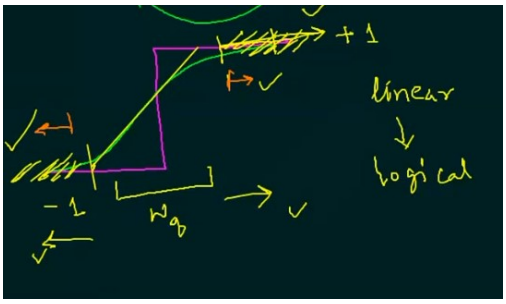

Figure 6: Shows the range in which W resides

Sometimes in VC analysis we compute

$$\Omega(h) = \sum \frac{(W_{ij}^l)^2}{\beta^2 + (W_{ij}^l)^2} \tag{12}$$

Here β is **Smoothing Parameter** and this is known as **self weight elimination**

7 Optimal λ

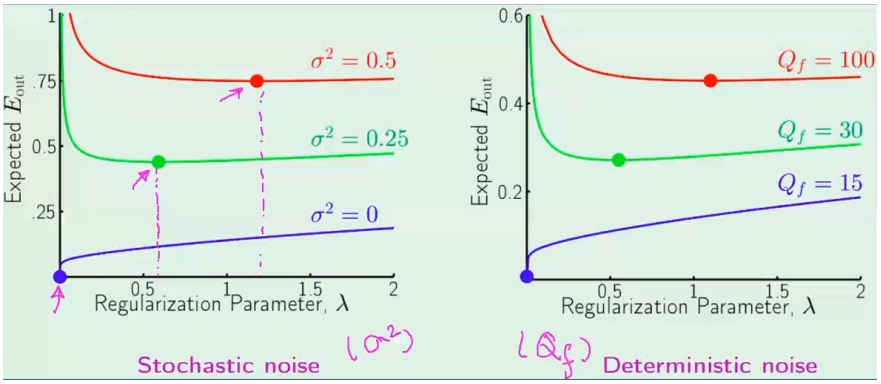

Figure 7: Optimum value of λ for different frequency of noise

These show those two noises are **not very different**