CS 419M Introduction to Machine Learning **Spring 2021-22**

Lecture 6: Overfitting and Regularization

Lecturer: Abir De *Scribe: Group 1*

6.1 Recap

In the modified linear regression problem we were solving,

$$w = \left[\lambda\mathbb{I} + \sum_{i\in D}(x_i x_i^T)\right]^{-1} \sum_{i\in D} x_i y_i$$

However, this w can be ill-conditioned for $\lambda \to 0$ due to the determinant possibly approaching 0. Also, this w is also not the optimal value of the mean squared loss minimization.

$$w \neq w^* = argmin_w \sum_{i\in D}(y_i - w^T x_i)^2$$

It is the optimal value for the following problem-

$$argmin_w \sum_{i\in D}(y_i - w^T x_i)^2 + \lambda\|w\|^2$$

6.2 Overfitting

Suppose we have data D with an extracted feature $x \in \mathbb{R}$ and target variable $y \in \mathbb{R}$. We formulate two problems.
Case-1:
feature variable- x

$$l_1 = min_w \sum_{i\in D}(y_i - wx_i)^2$$

Case-2:
feature variable- $\overrightarrow{x} = [x,\ x^2,\ x^3,\ \ldots,\ x^d]$

$$l_2 = min_w \sum_{i\in D}(y_i - w^T\overrightarrow{x_i})^2$$

Q. Which loss is lower?

$$l_2 < l_1$$

This is because the space of the optimization problem-2 is a superset of the space of problem-1. That is, $l_1 = min_{w:w[2:d]=0} \sum_{i\in D}(y_i - w^T\overrightarrow{x_i})^2$

However, to minimize loss, the solution produced by l_2 tries to fit all the points of the data along with their noise, as shown in the figure below. This results in a non linear function suited only to the training data. Therefore, the loss on the trainset is very low but the loss on a testset is very high. This is called **overfitting**.

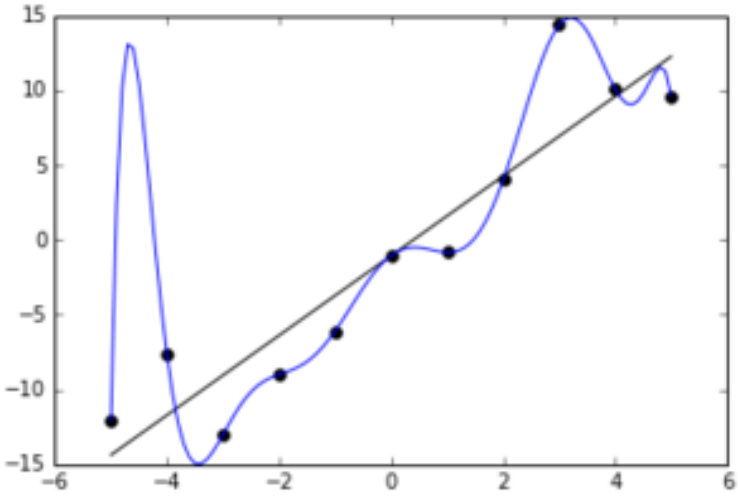

Figure 6.1: Overfitted function

Problem: We need to find a way to make the function depend on features which actually impact the target variable only, so that it does not try to overfit to noise using the remaining features. That is, the coefficients of features x_m that are not required should go to 0 in $h(\overrightarrow{x}) = w^T\overrightarrow{x}$.
Possible Solution: Penalise a norm of w to decrease the number of non-zero values. So,

$$loss = min_w \sum_{i \in D} (y_i - w^T \overrightarrow{x_i})^2 + \lambda \|w\|$$

Q. Which norm, $\|w\|_0, \|w\|_1, \|w\|_2$?
If we want to encourage sparsity (few non-zero values in the vector w), ideally we should try to minimize $\|w\|_0$ which is the number of non-zero values in w. But it is not differentiable. So, it is not used.
Therefore we approximate it with $\|w\|_1$ because it encourages sparsity. If the true w^* has only k non-zero values where $k/dim(w) << 1$ then L1 regularization ($\|w\|_1$ minimization) will recover it almost surely (with probability 1).

6.3 Probabilistic Interpretation

Now the question is what is the Probabilistic Interpretation of the regularized loss?

$$\min_w \sum_i (y_i - W^T x_i)^2 + reg$$

Here reg is the regularization of from L1 norm or L2 norm.

Or we can also ask what is the equivalent MLE?

Say D is data and W is used to generate data.

$P(D|W)$ is the likelihood that data D is generated by W or the distribution of the data given a W

$P(W)$ is the likelihood that the W we are using is correct W or the Prior distribution of W

We want to maximize the likelihood of the distribution to recover the loss. i.e.

$$\max_W P(D|W)P(W)$$

Normally we used to assume that the W is uniformly distributed and thus we use to only maximize the likelihood of the data given a W.

Now we assume that W is a Gaussian distribution with parameter λ.

Suppose that $W \sim N(0, \sigma^2)$ then, $\lambda \propto \frac{1}{\sigma^2}$

Q. What is the significance of λ?
As $\lambda \to highvalue \implies \sigma \to 0 \implies ||W|| \to 0$ as we are restricting the prior distribution of W.

Q. How to find a good value λ?

1. Find W as a function of λ
2. Test it on the validation set and find error as a function of λ i.e. Error(λ)
3. Now minimize this error function and find the optimal λ i.e. $\lambda^* = \min_\lambda \ Error(\lambda)$

Q. How error(λ) vs λ looks like?
Out of these 3 images which is the best curve for error v/s λ?

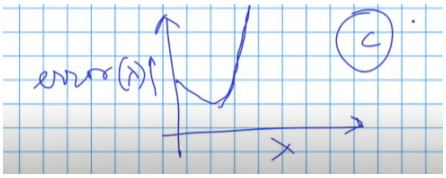

Figure 6.2: Image C: Error(λ) vs λ

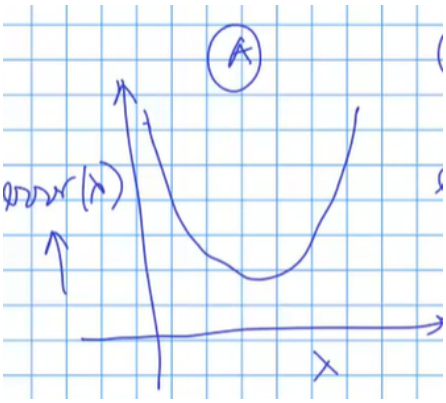

Figure 6.3: Image A: Error(λ) vs λ

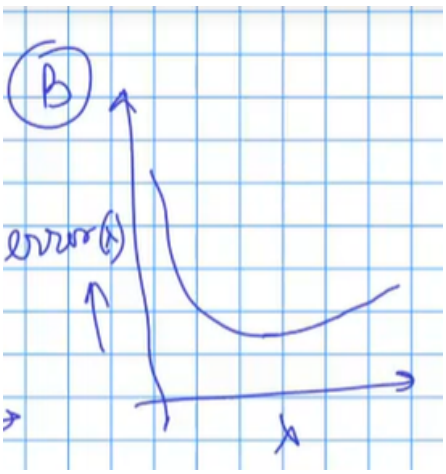

Figure 6.4: Image B: Error(λ) vs λ

Now as $\lambda \to \infty$, $error(\lambda) \to \sum(y_i)^2$ which is a finite value. Thus we rule out the possibility of both A and C. Thus the most suitable graph is graph B.

Now an important point to note is that as $\lambda \to 0$, W matrix becomes more and more ill conditioned and thus $||W|| \to \infty$ and thus $y_{predicted} \to \infty$ and thus, $error(\lambda) \to \infty$
Thus we neither want high value of λ nor want low value of λ. Thus the region where error(λ) reaches its minimum value is the optimal λ or λ^*.

The region where $\lambda \leq \lambda^*$ is overfitting region as we are not restricting W matrix. The region where $\lambda \geq \lambda^*$ is underfitting region as we are putting lot of restricting W matrix.

6.4 Summary

1. Optimal value of mean squared loss minimization is-

$$argmin_w \sum_{i \in D} (y_i - w^T x_i)^2 + \lambda \|w\|^2$$

2. Feature variable- $\overrightarrow{x} = [x,\ x^2,\ x^3,\ \ldots,\ x^d]$

$$l_2 = min_w \sum_{i \in D} (y_i - w^T \overrightarrow{x_i})^2$$

For above problem,loss on the trainset is very low but loss on a testset is very high, this is called **Overfitting**

3. Assuming W is a Gaussian distribution with parameter (λ), and $W \sim N(0, \sigma^2)$ then, $\lambda \propto \frac{1}{\sigma^2}$. Significance of ($\lambda$), As $\lambda \to highvalue \implies \sigma \to 0 \implies ||W|| \to 0$ as we are restricting the prior distribution of W

4. To find a good (λ), We find W as a function of (λ), find error by testing it on validation set and then minimize this error and find optimal (λ).

5. Variation of Error(λ): Error tends to infinity when (λ) tends to 0,it then reaches a minimum value and when (λ) tends to infinity it reaches a finite value of $\sum_i (y_i)^2$

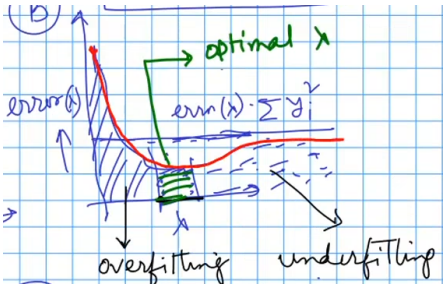

Figure 6.5: error(λ) vs λ

6.5 Group Details

Group	Name	Roll number
Group 1	Adit Akarsh	19D070003
Group 1	Aditya Vijay Jain	190100007
Group 1	Loveneesh Lawaniya	200110064
Group 1	Prateek Jha	200040106