PART-4 (Toy Example: Empirical Risk and Generalization)

Problem

Given a training dataset with features X and labels Y, let $\hat{f}(X)$ be the prediction of a model f, and let $L(\hat{f}(X), Y)$ be the loss function. Suppose we have two models, f_1 and f_2, and the empirical risk for f_1 is lower than that for f_2. We provide a toy example where model f_1 has a lower empirical risk on the training set but may not necessarily generalize better than model f_2.

Toy Example

Consider a small dataset for a regression task:

X	Y
1	1.5
2	2.0
3	2.5
4	3.5
5	5.0

We will now create two models, f_1 and f_2, and analyze their empirical risk and generalization ability.

Model f_1: High Complexity (Overfitting)

Model f_1 is a high-degree polynomial regression, such as a 4th-degree polynomial. It fits the training data perfectly, with an equation that could look like:

$$f_1(X) = 0.05X^4 - 0.3X^3 + 0.7X^2 - 0.4X + 1.5$$

The predicted values of f_1 on the training set are:

$$\hat{Y}_1 = [1.55, 1.94, 2.38, 3.12, 5.01]$$

The empirical risk, computed as the Mean Squared Error (MSE), for f_1 is:

$$MSE(f_1) = 0.033$$

Although the empirical risk is very low, this model is highly complex and likely overfits the noise in the data, which means it may not generalize well to unseen data.

Model f_2: Low Complexity (Better Generalization)

Model f_2 is a simple linear regression model, capturing the overall trend in the data. Its equation could be:

$$f_2(X) = 0.85X + 0.75$$

The predicted values of f_2 on the training set are:

$$\hat{Y}_2 = [1.60, 2.45, 3.30, 4.15, 5.00]$$

The empirical risk (MSE) for f_2 is:

$$MSE(f_2) = 0.255$$

Despite having a higher empirical risk on the training data, f_2 is likely to generalize better because it avoids overfitting and has lower variance.

Testing the Models

To further support our claim, we test both models on a new data point that was not part of the training set, say $X = 6$, where the true value of Y is $Y = 6.0$.

For model f_1:

$$f_1(6) = 0.05(6)^4 - 0.3(6)^3 + 0.7(6)^2 - 0.4(6) + 1.5 = 9.23$$

For model f_2:

$$f_2(6) = 0.85(6) + 0.75 = 5.85$$

We observe that model f_1 predicts 9.23, which is much farther from the true value of 6.0, while model f_2 predicts 5.85, which is closer to the true value.

Thus, the generalization ability of f_2 is better on unseen data, even though its empirical risk on the training set is higher.

Conclusion

In this example, model f_1 has a lower empirical risk on the training set because it overfits the data, but it does not generalize as well as model f_2. Testing on a new data point further shows that model f_2 provides a better prediction. This illustrates that a lower training error does not necessarily imply better performance on unseen data, especially when the model is overly complex and prone to overfitting.