CSE343: Machine Learning

Assignment 1: Theory

Name: Aditya Aggarwal
Roll Number: 2022028

Solution 1

As we increase the complexity of a regression model, either by adding more features or by increasing the degree of the polynomial for a fixed training set, over-fitting may occur. Over-fitting happens when the model essentially "memorizes" the training set, resulting in poor generalization to unseen data. This is due to the bias-variance trade-off. As the complexity of the model increases, it captures the training data more effectively, reducing its bias. However, at the same time, its variance increases, leading to a situation where the model has low bias and high variance. High variance indicates greater sensitivity to small fluctuations in the data. As a result, the model learns not only the underlying patterns in the training data but also the noise or random fluctuations. Thus, even though the training loss decreases with increasing model complexity, the test loss rises.

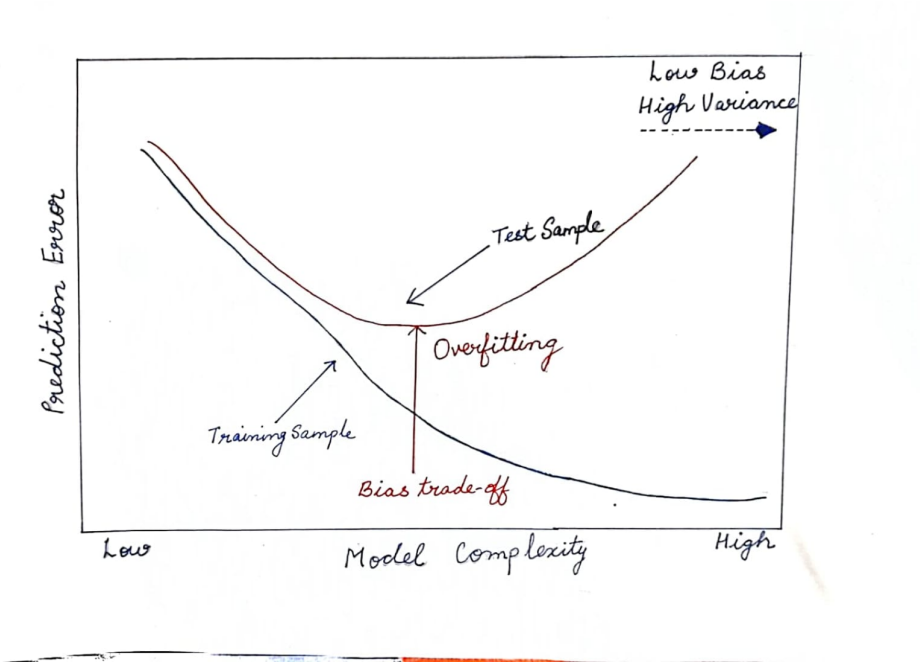

Figure 1: Graphical Representation of Over-fitting

Solution 2

- **True Positive (TP):** Spam emails correctly classified as spam.
- **True Negative (TN):** Legitimate emails correctly classified as legitimate.
- **False Positive (FP):** Legitimate emails incorrectly classified as spam.
- **False Negative (FN):** Spam emails incorrectly classified as legitimate.

From the given problem statement, we can determine the following directly:

- **True Positive (TP)** = 200
- **False Negative (FN)** = 50
- **True Negative (TN)** = 730
- **False Positive (FP)** = 20

Using these values, we can calculate each of the metrics provided by the confusion matrix:

- **Precision:**
$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} = \frac{200}{200 + 20} = 0.9091$$
- **Negative Predictive Value:**
$$\text{Negative Predictive Value} = \frac{\text{TN}}{\text{TN} + \text{FN}} = \frac{730}{730 + 50} = 0.9359$$
- **Recall (Sensitivity):**
$$\text{Recall (Sensitivity)} = \frac{\text{TP}}{\text{TP} + \text{FN}} = \frac{200}{200 + 50} = 0.8000$$
- **Specificity:**
$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} = \frac{730}{730 + 20} = 0.9733$$
- **Accuracy:**
$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} = \frac{200 + 730}{200 + 730 + 20 + 50} = 0.9300$$

The average of these metrics can be calculated as follows:

$$\text{Average classification performance} = \frac{0.9091 + 0.9359 + 0.8000 + 0.9733 + 0.9300}{5} = 0.9097$$

$$\text{Average classification performance} \approx 91\%$$

Thus, the average performance is approximately 91%. Note that the relatively low recall indicates a high number of false negatives. Thus, the number of spam emails classified as legitimate may be higher than desired.

Solution 3

Given the dataset X, the weights W, a linear regression model makes the prediction Y calculated as:

$$Y = XW$$

The weights and bias can be calculated by minimizing the Maximum Likelihood Estimation (MLE) and obtaining the expression:

$$W = (XX^T)^{-1}X^TY$$

For a dataset with only one feature, this simplifies to:

$$W = \begin{bmatrix} w \\ b \end{bmatrix}$$

where w and b are calculated as follows:

$$w = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2}$$
$$b = \bar{Y} - w \cdot \bar{X}$$

Given dataset:

$$X = \{3, 6, 10, 15, 18\}, \quad Y = \{15, 30, 55, 85, 100\}$$

The means of X and Y are:

$$\bar{X} = \frac{3 + 6 + 10 + 15 + 18}{5} = 10.4$$
$$\bar{Y} = \frac{15 + 30 + 55 + 85 + 100}{5} = 57$$

Using these values, we calculate w as:

$$w = \frac{(3 - 10.4)(15 - 57) + (6 - 10.4)(30 - 57) + (10 - 10.4)(55 - 57) + (15 - 10.4)(85 - 57) + (18 - 10.4)(100 - 57)}{(3 - 10.4)^2 + (6 - 10.4)^2 + (10 - 10.4)^2 + (15 - 10.4)^2 + (18 - 10.4)^2}$$

$$w \approx 5.7833$$

Next, we calculate b:

$$b = 57 - 5.7833 \times 10.4 \approx 3.1463$$

Thus, the weight vector W is:

$$W = \begin{bmatrix} 5.7833 \\ 3.1463 \end{bmatrix}$$

For a new data point $x = 12$, we calculate y as:

$$y = w \cdot x + b = 5.7833 \times 12 + 3.1463 \approx 66.253$$

Hence, for $x = 12$, the predicted y value is approximately 66.253.

Solution 4

Lower empirical risk, i.e., training loss, does not imply better generalization. Good generalization requires the model to perform well on unseen test data. It may happen that a model performs well on training data after optimizing its weights according to that data, but it performs poorly on test data for various reasons, like over-fitting.

Let us take a toy example consisting of $\hat{f}_1$ and $\hat{f}_2$. Let $\hat{f}_1$ be a quadratic model and $\hat{f}_2$ be a linear model. It may happen that the training set provided to both the models has a distribution that lies closer to a quadratic curve than a linear curve. In this case, the empirical risk of $\hat{f}_1$ will be lower. However, upon generalization, it may occur that $\hat{f}_1$ fails to perform due to over-fitting (since the complexity of the model might be higher than required for the unseen data distribution).

A visual representation of the described phenomenon is attached below.

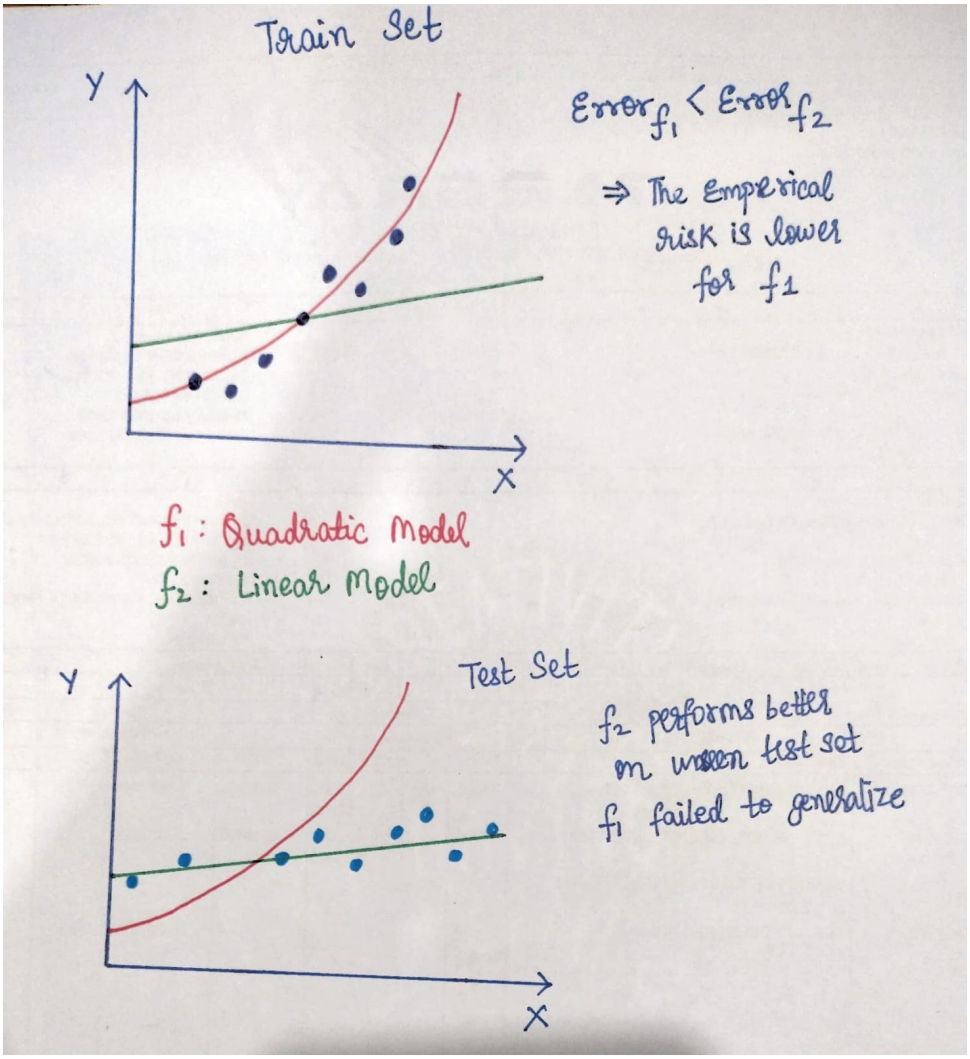

Figure 2: Graphical Representation of Over-fitting