Binary Classification and the Mathematical Framework of Statistical Learning Theory

Binary classification is a type of classification problem in machine learning where the goal is to assign one of two possible labels to an input based on its features. Formally, consider a dataset: $\mathcal{D} = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$ where $x_i \in \mathbb{R}^d$ is a feature vector and $y_i \in \{0, 1\}$ (or $\{-1, 1\}$) is a label representing one of the two classes. The objective is to learn a function $f : \mathbb{R}^d \to \{0, 1\}$ from a hypothesis space $\mathcal{H}$, which correctly classifies new, unseen data.

Problem Formulation

In binary classification, the task can be formally represented as finding a decision function $h \in \mathcal{H}$, such that for a new input x, we minimize the probability of error:

$$P(h(x) \neq y)$$

where $h(x)$ is the predicted label and y is the true label. The classification model's goal is to minimize this error, ideally finding the optimal hypothesis $h*$ that achieves the smallest error.

Statistical Learning Theory (SLT) and Binary Classification

Statistical Learning Theory (SLT) provides a probabilistic framework for understanding how models generalize to unseen data. The key challenge in binary classification is not just minimizing the error on the training data but ensuring that the model generalizes well to new data, i.e., generalization error is minimized.

Formally, let the empirical risk (training error) be:

$$R_{\text{emp}}(h) = \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}(h(x_i), y_i)$$

where $\mathcal{L}$ is a loss function, such as 0-1 loss:

$$\mathcal{L}(h(x_i), y_i) = \begin{cases} 0 & \text{if } h(x_i) = y_i \\ 1 & \text{if } h(x_i) \neq y_i \end{cases}$$

SLT aims to minimize not only the empirical risk but the true (expected) risk:

$$R(h) = \mathbb{E}_{(x,y)\sim\mathcal{P}}[\mathcal{L}(h(x), y)]$$

The difference between empirical risk and expected risk is captured by generalization error.

How SLT Provides a Mathematical Framework for Binary Classification

SLT provides the mathematical foundation for understanding how well a model learned from a finite sample generalizes to the entire data distribution. This is accomplished using the following key concepts:

1. VC Dimension (Vapnik-Chervonenkis Dimension)

The VC dimension measures the capacity (or complexity) of a model class $\mathcal{H}$. It is defined as the largest set of points that can be shattered (i.e., classified correctly) by the hypothesis class $\mathcal{H}$.

$$VC(\mathcal{H}) = d$$

where d is the maximum number of points that can be labeled in every possible way by some hypothesis in $\mathcal{H}$. A higher VC dimension indicates a more complex model, which can lead to overfitting if too high.

2. Generalization Bound

SLT provides generalization bounds that relate the true risk $R(h)$ and the empirical risk $R_{\text{emp}}(h)$ using the VC dimension:

$$P\left(|R(h) - R_{\text{emp}}(h)| \geq \epsilon\right) \leq 2|\mathcal{H}| \exp(-2n\epsilon^2)$$

This inequality shows that with high probability, the generalization error is bounded by the complexity of the hypothesis space (VC dimension) and the size of the training set n. The goal is to minimize the sum of the empirical risk and the complexity of the hypothesis space.

3. Regularization

Regularization techniques, inspired by SLT, add constraints or penalties to the learning process to control the complexity of the model and prevent overfitting. This balances empirical risk minimization and hypothesis complexity.

Conclusion

In binary classification, the goal is to minimize both the training error and generalization error. SLT offers the mathematical framework that balances this trade-off by introducing the concepts of VC dimension, generalization bounds, and regularization. These tools ensure that machine learning models generalize well to unseen data, avoiding overfitting and underfitting.