Introduction

Binary classification is a central problem in machine learning, where the goal is to assign an instance $X \in \mathcal{X}$ to one of two possible classes, $Y \in \{-1, +1\}$. Given a set of training data $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$, where each pair (X_i, Y_i) is sampled independently from an unknown joint probability distribution $P(X, Y)$, the objective is to construct a classifier $f : \mathcal{X} \to \{-1, +1\}$ that minimizes classification error on unseen data. The framework for addressing this problem is provided by Statistical Learning Theory (SLT), which formalizes the goal of minimizing a classifier's risk and provides guarantees on learning performance.

Formal Definition of Binary Classification

In binary classification, we seek a function $f : \mathcal{X} \to \{-1, +1\}$ that minimizes the *expected risk*, defined as:

$$R(f) = E_{(X,Y)\sim P}[\ell(f(X), Y)],$$

where $\ell(f(X), Y)$ is a loss function that quantifies the cost of predicting $f(X)$ when the true label is Y. A common choice of loss is the *0-1 loss*, which is defined as:

$$\ell(f(X), Y) = \begin{cases} 1 & \text{if } f(X) \neq Y, \\ 0 & \text{if } f(X) = Y. \end{cases}$$

The goal is to find a function f that minimizes this expected risk.

Bayes Classifier and Risk Minimization

The *Bayes classifier*, denoted f_{Bayes}, is the optimal classifier that minimizes the risk:

$$f_{\text{Bayes}}(X) = \begin{cases} +1 & \text{if } P(Y = +1 \mid X = x) \geq 0.5, \\ -1 & \text{otherwise.} \end{cases}$$

The risk of the Bayes classifier, known as the *Bayes risk*, represents the theoretical lower bound on classification error, but it cannot be directly computed since the true distribution $P(X, Y)$ is unknown.

SLT Framework for Binary Classification

SLT addresses the challenge of learning a classifier f without knowing $P(X, Y)$. Instead of minimizing the true risk $R(f)$, we minimize the *empirical risk* $R_n(f)$ based on the observed training data:

$$R_n(f) = \frac{1}{n} \sum_{i=1}^{n} \ell(f(X_i), Y_i).$$

However, simply minimizing empirical risk leads to overfitting. SLT introduces the concept of *generalization* to ensure that the classifier performs well on unseen data. This is achieved through the use of *capacity control* methods, such as restricting the hypothesis space $\mathcal{F}$ to functions that are not too complex, measured by concepts like VC-dimension.

Conclusion

SLT provides a rigorous mathematical framework for binary classification by defining the risk minimization problem, introducing the notion of the Bayes classifier, and addressing the practical challenges of learning from finite data through empirical risk minimization and generalization guarantees. This framework underlies many modern machine learning algorithms and ensures that the learned classifier is not only accurate on the training data but also generalizes well to new examples.