Binary Classification and Statistical Learning Theory

Binary classification is a fundamental problem in machine learning, where the goal is to learn a function $f : X \to Y$, mapping input instances from space X to binary output labels from space Y, where Y is defined as $\{-1, +1\}$. The challenge lies in minimizing the classification error, defined by a loss function ℓ, which quantifies the cost of misclassifying an instance X as a label $f(X)$. A common loss function in binary classification is the 0-1 loss:

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

The performance of a classifier is assessed through the risk $R(f)$, which is the expected loss across the distribution P of the input space:

$$R(f) = E[\ell(X, Y, f(X))]$$

The optimal classifier, known as the Bayes classifier, aims to minimize the risk and is defined as:

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

In practice, the underlying distribution P is unknown, making it challenging to compute the Bayes classifier directly. Here, Statistical Learning Theory (SLT) provides a robust mathematical framework to address the challenges of binary classification.

SLT operates under several key assumptions, including that the training data is sampled independently and identically distributed (iid) from an unknown distribution P. It establishes a foundation for learning by providing insights into how well a classifier can generalize from training data to unseen instances. Importantly, SLT formulates the problem of binary classification by introducing concepts such as the loss function and risk.

SLT develops tools to analyze the performance of learning algorithms, providing bounds on the generalization error, which is the difference between the empirical risk (calculated from the training data) and the true risk (expected

over the distribution P). This analysis allows practitioners to select appropriate learning algorithms and determine the quantity of training data required to achieve a desired level of performance.