Statistical Learning Theory: Models, Concepts, and Results

1 Binary Classification Problem

In Machine Learning, binary classification is a fundamental problem that involves assigning a label (± 1) to an input data point $x \in X$. The goal is to find a mapping $f : X \to Y$ that minimizes the risk of misclassifying a new, unseen data point.

1.1 Formal Definition

Let (X, Y) be a random pair of variables, where X is the input space and Y is the label space ($Y = \{-1, 1\}$). The probability distribution P on $X \times Y$ is unknown. We are given a training set of n samples, $\{(x_1, y_1), ..., (x_n, y_n)\}$, drawn independently from P.

1.2 Goal

Find a function $f : X \to Y$ that minimizes the risk:

$$R(f) = E[\ell(x, y, f(x))]$$

where $\ell(x, y, f(x))$ is the loss function, e.g., 0-1 loss or misclassification error.

2 SLT Framework

Statistical Learning Theory (SLT) provides a mathematical framework to solve the binary classification problem. The key concepts are:

1. **Joint Probability Distribution**: P on $X \times Y$, which is unknown.
2. **Training Set**: $\{(x_1, y_1), ..., (x_n, y_n)\}$, drawn independently from P.
3. **Loss Function**: $\ell(x, y, f(x))$, e.g., 0-1 loss or misclassification error.
4. **Risk**: $R(f) = E[\ell(x, y, f(x))]$, which we want to minimize.
5. **Function Space**: F, a set of functions from X to Y.

The SLT framework provides a way to analyze the problem of binary classification and to come up with solutions that can be guaranteed to work well.

3 SLT Solution

SLT offers a mathematical framework to solve the binary classification problem by:

1. **Empirical Risk Minimization**: Find a function $f \in F$ that minimizes the empirical risk:

$$R_{emp}(f) = \frac{1}{n} \sum_{i=1}^{n} \ell(x_i, y_i, f(x_i))$$

2. **Uniform Convergence**: Show that the empirical risk converges uniformly to the true risk:

$$P\left(\sup_{f \in F} |R(f) - R_{emp}(f)| > \epsilon\right) \to 0 \quad \text{as } n \to \infty$$

This ensures that the empirical risk minimizer is a good approximation of the true risk minimizer.