Binary Classification and Statistical Learning Theory

Binary classification is one of the fundamental problems in machine learning. It deals with predicting a label from a set of two possible labels, usually represented as -1 and +1. Given a dataset of examples $(X_1, Y_1), \ldots, (X_n, Y_n)$, where X_i represents an input feature vector and Y_i represents its corresponding label, the task is to find a function $f: X \to Y$ that correctly classifies new, unseen data points.

Formally, the goal is to find a classifier f that minimizes the misclassification risk R(f), defined as:

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

where L is a loss function (in binary classification, typically 0-1 loss), X is the input space, Y is the output space, and E is the expectation over the joint distribution P(X, Y).

SLT Framework for Binary Classification

Statistical Learning Theory (SLT) provides a mathematical framework for solving binary classification by considering the problem in terms of probabilistic assumptions. It assumes that the data is generated from an unknown joint probability distribution P(X, Y), and the classifier f aims to minimize the risk (expected loss) with respect to this distribution. The Bayes classifier, f_{Bayes}, represents the optimal classifier:

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

However, P(X, Y) is unknown, so the challenge is to estimate a classifier as close as possible to the

Bayes classifier using only training examples.

SLT defines learning in terms of finding a function from a function space F that approximates the Bayes classifier. By using concepts like Vapnik-Chervonenkis (VC) dimension, SLT establishes generalization bounds, which provide guarantees on how well a classifier will perform on unseen data based on its performance on the training set.

In conclusion, SLT offers a rigorous framework to address the binary classification problem by offering theoretical guarantees on generalization and guiding the selection of classifiers that minimize the expected classification error.