Katsuba Danila 5130203_20101

Understanding Binary Classification and the Role of Statistical Learning Theory in Machine Learning

Introduction to Binary Classification

Define binary classification mathematically.

Objective: Given data points $x \in R^d$ and corresponding labels $y \in \{0,1\}$, the goal is to find a hypothesis h(x) such that h(x) ≈ y.

Formal Problem Statement

Given a dataset D = {(x1, y1), (x2, y2), ..., (xn, yn)} = drawn from an unknown distribution D(X, Y), the task is to find a function f: X→{0,1} such that the probability of misclassification, i.e., P(f(x)≠y), is minimized.

Loss function: Define the 0-1 loss L(f(x), y) = $1_{\mathrm{f(x)\neq y}}$.

Role of Statistical Learning Theory (SLT)

SLT provides a framework for understanding the generalization of learning algorithms.

Formal objective: minimize expected risk R(f)=$E_{\mathrm{(x,y)\sim D}}$ [L(f(x), y)].

Empirical risk minimization (ERM): Approximate R(f) with the empirical risk R_n(f)=$\frac{1}{n}\sum_{i=1}^{n} L(f(x_i), y_i)$.

SLT Framework for Binary Classification

SLT introduces the concept of the Vapnik-Chervonenkis (VC) dimension to measure the complexity of the hypothesis space H.

The trade-off between model complexity (VC dimension) and generalization is explained using the VC Inequality:

$$R(f) \leq R_n(f) + \sqrt{\frac{h\left(\log\left(\frac{2n}{h}\right) + 1\right) - \log\left(\frac{\delta}{4}\right)}{n}}$$

where h is the VC dimension, and n is the number of samples.

Conclusion

SLT provides theoretical guarantees for binary classification through concepts like VC dimension and generalization bounds.

Empirical risk minimization, combined with SLT principles, enables the development of effective classification algorithms.