Binary Classification and Statistical Learning Theory

Shmeleva Mariia 5130203/20101

Introduction

Binary classification is a fundamental problem in machine learning, where the goal is to assign one of two possible labels, typically represented as $+1$ and -1, to given input data. Formally, we deal with an input space $\mathcal{X}$ and a label space $\mathcal{Y} = \{-1, +1\}$. The problem is to learn a function $f : \mathcal{X} \to \mathcal{Y}$, known as a classifier, that can accurately predict the label $y \in \mathcal{Y}$ for unseen instances $x \in \mathcal{X}$.

Mathematical Formulation

In the setting of supervised learning, we are given a set of training data $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$, where $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$ are assumed to be independently and identically distributed (i.i.d.) samples from an unknown probability distribution $P(X, Y)$. The objective is to find a function f that minimizes the expected risk, defined as:

$$R(f) = \mathbb{E}_{(X,Y)\sim P}[\ell(X, Y, f(X))], \tag{1}$$

where $\ell : \mathcal{X}\times\mathcal{Y}\times\mathcal{Y} \to \mathbb{R}_{\geq 0}$ is a loss function that quantifies the cost of predicting $f(X)$ when the true label is Y. For binary classification, a common choice is the 0-1 loss:

$$\ell(X, Y, f(X)) = \begin{cases} 1 & \text{if } f(X) \neq Y, \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

Since the underlying distribution P is unknown, the expected risk $R(f)$ cannot be computed directly. Instead, we approximate it using the empirical risk $R_{\text{emp}}(f)$ based on the training data:

$$R_{\text{emp}}(f) = \frac{1}{n}\sum_{i=1}^{n} \ell(x_i, y_i, f(x_i)). \tag{3}$$

Statistical Learning Theory Framework

Statistical Learning Theory (SLT) provides a mathematical framework to analyze the problem of learning a classifier from data. The core idea is to ensure

that the learned function f not only performs well on the training data but also generalizes well to unseen data. SLT introduces the concept of
textituniform convergence, which ensures that, with high probability, the empirical risk $R_{\text{emp}}(f)$ is close to the expected risk $R(f)$ for all functions f in the considered hypothesis class $\mathcal{F}$:

$$\sup_{f \in \mathcal{F}} |R(f) - R_{\text{emp}}(f)| \leq \epsilon, \tag{4}$$

where ϵ is a small positive value that decreases as the number of training samples n increases.

The key result that SLT provides is the
textitVC dimension (Vapnik-Chervonenkis dimension), which is a measure of the capacity or complexity of the hypothesis class $\mathcal{F}$. The VC dimension helps in deriving bounds on the generalization error, which ensures that the learned classifier will perform well on new data, given a sufficient number of training examples.

Conclusion

SLT offers a mathematical framework to address the binary classification problem by providing tools to analyze the generalization ability of classifiers. By leveraging concepts such as empirical risk minimization, uniform convergence, and VC dimension, SLT establishes theoretical guarantees on the performance of learning algorithms, ensuring that they can generalize from finite training data to unseen instances.