Student: Tamara Goncharova, 5130203/20101

Report

Statistical Learning Theory

The problem of classification stands alongside clustering and regression as one of the leading techniques in machine learning for dealing with data. The main idea of classification is to categorize given instances into a finite set of labels. All instances form an input space $\mathcal{X}$, while all labels form an output space $\mathcal{Y}$.

If the output space has only two values, this is called binary classification. Typically, $\mathcal{Y}$ is denoted as $\{-1, +1\}$. The process of learning involves finding a function $f: \mathcal{X} \to \mathcal{Y}$ (a classifier) using given examples. The examples are a set of pairs (training points) $(X_1, Y_1), \dots, (X_n, Y_n) \in \mathcal{X} \times \mathcal{Y}$. Since classification is a supervised learning problem, the system knows the correct answers during training and can use them to learn.

Statistical Learning Theory (SLT) typically assumes that data points are sampled independently from an unknown joint distribution $P(X_i, Y_i)$. This assumption holds in many cases (e.g., handwritten digits), but may fail in fields like drug discovery, where the data is hand-selected. In some areas, such as active learning or time-series prediction, this assumption is relaxed.

Labels may not be deterministic due to label noise or overlapping classes. For instance, human error in labeling or inherent overlap (e.g., predicting gender based on height) can introduce uncertainty. The conditional likelihood of a label given an input, $\eta(x) = P(Y = 1 \mid X = x)$, indicates how deterministic the label is. The closer $\eta(x)$ is to 0.5, the harder it becomes to learn due to increased ambiguity.

To decide whether a function f is a good classifier, a loss function is used to quantify the cost of misclassification. A common choice is the 0-1 loss function:

$$\ell\big(X, Y, f(X)\big) = \begin{cases} 1, if\ P(Y = 1|X = x) \geq 0.5, \\ -1, otherwise. \end{cases}$$

The Bayes classifier chooses labels based on the conditional probability $P(Y = 1|X = x)$. However, since this distribution is unknown, we cannot directly compute it. The goal of binary classification is to find a function f that minimizes risk, but without knowing P, this is challenging. Statistical Learning Theory provides a framework to address this problem and guarantees the quality of the solution.

Statistical Learning Theory provides a foundational mathematical framework for solving the binary classification problem in machine learning by formalizing the learning process through key concepts such as risk minimization, loss functions, and probability distributions. SLT employs

Empirical Risk Minimization (ERM) to select classifiers that minimize the average loss on training data, ensuring that the chosen classifier performs well on the observed examples. It introduces generalization bounds using tools like the VC dimension to quantify the difference between empirical risk and true risk, thereby guaranteeing that the classifier will perform reliably on unseen data. Additionally, SLT addresses issues like label noise and overfitting through regularization techniques, which balance model complexity and performance. By analyzing the consistency and convergence of learning algorithms, SLT ensures that as the number of training samples increases, the learned classifier approaches the optimal Bayes classifier. Overall, SLT systematically bridges the gap between theoretical guarantees and practical algorithm performance, enabling robust and effective binary classification in machine learning.

Taking everything into account, SLT provides a rigorous framework for addressing the binary classification problem, ensuring that learned classifiers generalize well from training data to unseen examples through techniques like Empirical Risk Minimization and regularization. By analyzing model consistency and convergence, SLT helps bridge the gap between theoretical guarantees and practical performance in machine learning.