Vasilev Vadim gr. 5130203/20102

The Problem of Binary Classification

Binary classification is a fundamental problem in supervised learning, where the objective is to categorize instances into one of two predefined classes, typically labeled as +1 and -1. In machine learning, the goal is to learn a function, called a classifier, that maps inputs from a given input space X to labels in a label space $Y = \{-1, +1\}$. This function is derived from a set of training examples consisting of pairs $(X_1, Y_1), \ldots, (X_i, Y_i)$, where each X_i s an instance, and each Y_1 is the corresponding label.

Key Concepts in Binary Classification:

- Training Data: The examples used to learn the classifier are assumed to come from an unknown joint probability distribution $P(X, Y)$. The examples are drawn independently from this distribution, a condition referred to as independent and identically distributed.
- Classifier: The classifier $f : X \to Y$ aims to minimize classification errors on both seen (training) and unseen (test) data. The challenge is to generalize from the training data to make accurate predictions on new data.
- Label Noise and Overlapping Classes: In real-world scenarios, the labels may not always be deterministic. Noise in the data, such as human errors in labeling, or natural overlaps between classes (e.g., predicting gender based on height), complicates classification.
- Conditional Probability: The key quantity of interest in binary classification is the conditional probability $\eta(x) = P(Y = 1 \,|\, X = x)$, which represents the likelihood that an input x belongs to class +1. If $\eta(x) \geq 0.5$, the classifier should assign label +1 to x ; otherwise, it should assign -1.
- Bayes Classifier: The optimal classifier under the assumption that the probability distribution is known is the Bayes classifier, which minimizes the risk (expected loss). However, since the probability distribution is unknown in practice, constructing the Bayes classifier directly is impossible.

SLT Framework for Solving Binary Classification

Statistical Learning Theory (SLT) provides a mathematical foundation for solving the problem of binary classification. SLT operates under several assumptions:

- Unknown Distribution: The underlying distribution $P(X, Y)$ is unknown and must be inferred from the training data.
- Risk Minimization: The objective is to minimize the *risk* $R(f)$, which represents the expected classification error across the entire input space. However, without knowing P this risk is approximated using the training data.
- Empirical Risk Minimization (ERM): One of the central strategies in SLT is Empirical Risk Minimization (ERM), where the classifier is chosen to minimize the error on the training data. This, however, raises concerns about overfitting, where the model performs well on training data but poorly on new data.

In summary, SLT forms the mathematical backbone for understanding and addressing the challenges of binary classification. By establishing principles such as ERM, SLT provides the theoretical guarantees necessary for designing algorithms that generalize well to unseen data.