Duro Ilya, 5130203/20102

Binary classification refers to the task of classifying an input $x \in X$ into one of two classes, typically represented as $\{-1, +1\}$. Mathematically, the goal is to find a function, called a classifier, that maps inputs from a given input space X to labels in a label space $Y = \{-1, 1\}$.

Given a dataset $D = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$, where $x_i \in X$ represents the input and $y_i \in \{-1, 1\}$ is the corresponding label, the problem is to find a function f that minimizes the classification error. 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. The challenge is to generalize from the training data to make accurate predictions on new data. For this purpose, loss functions are used, for example the simplest loss function in classification is the 0-1-loss:

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

The risk of a function is the average loss over data points generated according to the underlying distribution P:

$$R(f) := E(\ell(X, Y, f(X))).$$

This risk measures the number of elements in the instance space X that are misclassified by the function f. The optimal classifier under the assumption that the probability distribution is known is the Bayes classifier, which minimizes the risk (expected loss):

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

where $\eta(x) = P(Y = 1 \mid X = x)$ is the conditional probability of the label being +1 given x.

Statistical Learning Theory (SLT) provides a mathematical framework to solve binary classification by defining conditions for learning from finite samples. The key concepts of SLT are:

- Unknown Distribution: The underlying distribution is unknown and must be inferred from the training data.
- Generalization: SLT focuses on minimizing the generalization error R(f), not just fitting the training data. SLT introduces the concept of empirical risk minimization (ERM), where we minimize the average loss over the training set:

$$\mathrm{R}_{\mathrm{emp}}(f) := \frac{1}{n} \sum_{i=1}^{n} \ell(X_i, Y_i, f(X_i)).$$

In conclusion, SLT forms the mathematical backbone for understanding and addressing the challenges of binary classification.