Marchenko Kristina,
5130203/20102

In binary classification, we aim to classify objects into one of two categories. Formally, consider:

- **Input Space** X, which contains instances $x \in X$.
- **Output Space** $Y = \{-1, +1\}$, which contains the two possible class labels.

The task is to learn a function $f : X \to Y$ called the **classifier,** which assigns a label $y \in Y$ to each input $x \in X$. Given a set of training examples $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n) \in X \times Y$, drawn from an unknown probability distribution $P(X, Y)$, the goal is to find f that generalizes well to unseen data, minimizing misclassification errors.

The **loss function** measures the performance of the classifier. For binary classification, the 0-1 loss is commonly used:

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

The objective is to minimize the **expected risk** (or generalization error), defined as the expected loss over the data distribution P:

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

The optimal classifier is the **Bayes classifier** f_{Bayes}, which minimizes the risk:

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

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

How SLT Offers a Mathematical Framework

Statistical Learning Theory (SLT) provides the foundational framework to analyze learning algorithms. Key concepts are:

1. **Agnostic setting**: SLT assumes no prior knowledge of the distribution $P(X, Y)$. Instead, the task is to find a classifier f based on empirical data, without assumptions about the data's distribution.
2. **Generalization error**: 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:

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

   SLT provides tools like VC dimension to bound the difference between the empirical risk and true risk, ensuring that with enough data, minimizing $R_{emp}(f)$ also minimizes $R(f)$.
3. **Capacity control**: SLT emphasizes controlling the complexity of the hypothesis space F, from which the classifier is chosen, to avoid overfitting. The VC dimension of F quantifies its capacity, and SLT provides bounds on the generalization error based on the VC dimension and the number of training samples.

In conclusion, SLT offers a rigorous framework by formalizing the learning problem in probabilistic terms, focusing on generalization, and providing mathematical tools to balance fitting the data and avoiding overfitting.