Binary classification

In binary classification, we deal with an input space (space of instances) X and and output space (label space) Y. We identify the label space with the set $\{-1, +1\}$. The task involves assigning each object from space of instances to one of these two classes. The issue of learning can be simplified to estimating a functional relationship represented as $f : X \to Y$. This type of mapping f is referred to as a classifier. In order to do this, we get access to some training points $(X_1, Y_1), ..., (X_n, Y_n) \in X \times Y$, drawn from an unknown probability distribution $P(X, Y)$, the goal is to find function f that generalizes well to unseen data, minimizing misclassification errors. For this purpose, loss functions are used, for example the simplest loss function in classification is the 0-1 -loss:

$$l(X, Y, f(X)) = \begin{cases} 1 \; if \; f(X) \neq Y \\ 0 \; otherwise \end{cases}$$

The risk of a function is the average loss over data points generated according to the underlying distribution $P : R(f) := E(l(X, Y, (X)))$. In other words, the risk of a classifier f is the expected loss of the function f across all points $X \in X$. This risk measures the number of elements in the instance space X that are misclassified by the function f. Of course, a function f is a better classifier than another function g if its risk is smaller, that is if $R(f) < R(g)$. To find a good classifier f we need to find one for which $R(f)$ is as small as possible. The best classifier is the one with the smallest risk value $R(f)$. We can formally write down what the optimal classifier should be:

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

Statistical learning theory

Statistical learning theory (SLT) offers a mathematical foundation essential for tackling binary classification problems in machine learning. In this context, SLT defines a collection of functions that convert input features into output classes, such as +1 or -1. This framework aids in evaluating the performance of a hypothesis from this function space based on training data. SLT introduces the loss function concept to measure how a function fails, with the objective of minimizing the expected loss across the data distribution, thereby enhancing generalization capabilities.

The empirical risk minimization principle advocates for reducing empirical risk, which is the average loss on training data, as a method to identify effective hypotheses. This approach balances fitting the training data well with the capacity to handle new, unseen data. Furthermore, SLT provides theoretical bounds, ensuring that a well-learned hypothesis performs effectively on unknown data.