Introduction

Binary classification is one of the fundamental tasks in machine learning, where the goal is to assign objects to one of two categories. In this task, each object from the set X is assigned a label $y \in \{0,1\}$. The goal is to find a function (classifier) $f:X \to Y$ that minimizes the classification error on unseen data. Statistical Learning Theory (SLT) provides the mathematical foundation for understanding the learning process from finite data and analyzing the quality of solutions under limited samples.

Binary Classification: Mathematical Formulation

Consider the following essential elements of binary classification:

- $X \subseteq R^d$ is the feature space (each object $x \in X$ is represented as a feature vector).
- $Y=\{0,1\}$ is the label space, where $y=0$ or $y=1$ represent two possible classes.
- $P(X,Y)$ is the probability distribution over $X \times Y$, which is unknown but defines the relationship between objects and labels.

The objective: Find a function $f:X \to Y$ that minimizes the classification risk:

$R(f)=E_{(X,Y)\sim P}[\ell(f(X),Y)]$

where $\ell(f(x),y)$ is the loss function (e.g., 0-1 loss) that measures the discrepancy between the model’s prediction f(x) and the true label y.

SLT’s Contribution to Solving the Problem

SLT focuses on learning from finite samples and provides tools to analyze the generalization ability of models. Key SLT concepts include:

1. **Empirical Risk Minimization (ERM)**: In practice, the true probability distribution P(X,Y) is unknown, so instead of minimizing the true risk R(f), we minimize the empirical risk based on a training sample $S=\{(x_1,y_1),\dots,(x_n,y_n)\}$:

$$R_S(f) = \frac{1}{n}\sum_{i=1}^{n} l(f(x_i), y_i)$$

   The ERM principle suggests selecting a function f that minimizes this empirical risk.

2. **Generalization Boundaries**: A key question in SLT is how well the chosen solution f on the training sample will perform on new data. The theory provides generalization bounds, which limit the difference between the empirical risk $R_S(f)$ and the true risk R(f). These bounds depend on the complexity of the hypothesis space F(e.g., VC dimension) and the sample size n:

$$|R(f) - R_S(f)| \leq O(\sqrt{\frac{VC(F)}{n}})$$

   The more complex the model (i.e., higher VC dimension), the higher the likelihood of overfitting to the training data.

3. **Regularization and Avoiding Overfitting**: To ensure that a model generalizes well to new data, SLT introduces the concept of regularization — controlling the complexity of the hypothesis space F to prevent overfitting. This can be achieved by adding regularizers to the loss function or choosing models with lower complexity.

Conclusion

SLT provides fundamental tools for solving the binary classification problem, offering mathematical guarantees of a model’s ability to generalize. By applying principles of ERM, analyzing VC dimension, and using regularization, it is possible to build models that perform well on unseen data, minimizing the risk of overfitting.