Generalization Bounds

Theoretical Foundations of Deep Learning

Matteo Mazzarelli

December 17, 2024

What is Machine Learning?

- Machine learning is the process of learning from data to make predictions or decisions. [1]

Supervised Learning

- We focus on supervised learning where the data consists of input-output pairs, called features (x_i) and labels (y_i). [1-3]
- The goal is to infer y_i from x_i. [4]

The Learning Problem

- We have a dataset of observations $S = \{(x_1, y_1), ..., (x_m, y_m)\}$. [4]
- We wish to learn how to infer the value of y_i given x_i. [4]

Statistical Model

- We assume the values of (x_i, y_i) in the dataset are a random sample from a larger population. [5]
- The values of x_i and y_i are realizations of two random variables X and Y with probability distributions P_X and P_Y respectively. [5]

Joint Distribution

- There is a relation between the features and the labels. [6]
- The value of Y is conditioned on the value of X. [6]
- This is expressed by the conditional probability $P(Y|X)$. [6]
- We can compress P_X and P_Y into a single joint distribution $P(X, Y) = P(X)P(Y|X)$. [6]

Target Function

- The **target function** is $f(X) = \mathbb{E}[Y|X]$. [7]
- It represents the expected value of the label Y given the features X. [7]
- It becomes the target of the machine learning process. [7]
- The goal is to estimate this function f. [7]

Hypothesis and Hypothesis Space

- A **hypothesis**, denoted by h, attempts to estimate the target function f. [7]
- The **hypothesis space**, denoted by $\mathcal{H}$, is the set of possible functions considered for h. [7, 8]

Empirical Risk

- The **empirical risk**, or training error, $R_{emp}(h)$, is the average loss of a hypothesis h on the training data. [9]
- It can be calculated using a loss function that quantifies the difference between the predicted and actual labels. [9]

Overfitting

- Simply minimizing empirical risk can lead to **overfitting**. [10]
- An overfit model performs well on the training data, but poorly on unseen data. [10, 11]
- This occurs when the model learns the specific details of the training data instead of the underlying patterns. [10]

Generalization Error

- The **generalization error** (risk), $R(h)$, measures how well the hypothesis h performs on unseen data. [12]
- It's the expected value of the loss over the entire joint distribution $P(X, Y)$. [12]

Generalization Gap

- The **generalization gap** is the difference between the training error and the generalization error. [13]
- It quantifies how well the performance on the training data generalizes to unseen data. [13]

Motivation

- **Generalization bounds** provide guarantees that the learned hypothesis will perform well on unseen data. [2, 14, 15]
- They relate the generalization error to quantities we can observe or control, such as empirical risk, hypothesis space complexity, and dataset size. [16]

Hoeffding's Inequality

- For a single hypothesis h, Hoeffding's inequality bounds the difference between the empirical risk and the generalization error. [17, 18]

Limitations of Hoeffding

- Hoeffding's inequality doesn't directly apply to the entire hypothesis space. [8]
- It only considers the boundedness of the functions, not their variance. [19]
- The union bound, used to extend Hoeffding to multiple hypotheses, assumes all hypotheses are independent, which is not generally true. [20, 21]

The Union Bound

- The **union bound** extends the probability bounds to the entire hypothesis space. [22]
- It states that the probability of at least one hypothesis having a large generalization gap is at most the sum of the probabilities of each individual hypothesis having a large gap. [22]

The Growth Function

- The **growth function** quantifies the expressiveness of the hypothesis space. [23]
- It's the maximum number of ways the hypothesis space can label a dataset of a given size. [23]

VC Dimension

- The **VC dimension** is the largest dataset size that the hypothesis space can **shatter**. [24, 25]
- Shattering means the hypothesis space can produce all possible labelings for the dataset. [24]
- It's a measure of the complexity of the hypothesis space. [26]

VC Generalization Bound

- The **VC generalization bound** relates the generalization error to the empirical risk, VC dimension, and dataset size. [27]
- It shows that the generalization error can be bounded by the empirical risk plus a term that depends on the VC dimension and dataset size. [28]

Distribution-Based Bounds

- VC dimension is **distribution-free**, meaning it doesn't consider the data distribution. [29]
- Tighter bounds can be obtained by considering the data distribution. [29, 30]
- **Support Vector Machines (SVMs)** exemplify this by maximizing the margin between classes, which leads to a lower VC dimension and better generalization. [30]

Other Capacity Measures

- **Covering numbers** measure the size of the hypothesis space using a metric based on the difference in predictions on the training data. [31]
- **Rademacher complexity** measures how well the hypothesis space can fit random noise. [32, 33]
- These measures can be used to derive generalization bounds. [32, 34]

The General Form of Generalization Bounds

- The general form of generalization bounds is: $R(h) \leq R_{emp}(h) + C(|\mathcal{H}|, N, \delta)$. [35]
- $C(|\mathcal{H}|, N, \delta)$ represents a complexity term that depends on the hypothesis space complexity, dataset size, and the desired confidence level. [35]
- It highlights the trade-off between minimizing the training error and controlling the model's complexity. [35]

Relevance to Other Topics

Rates of Convergence

- **Rates of convergence** quantify how fast the generalization error decreases with increasing sample size. [36]
- They are closely linked to generalization bounds, as the bounds often provide insights into the rate of convergence. [37]

PAC-Bayes

- **PAC-Bayes** offers a Bayesian approach to deriving generalization bounds. [38]

Key Takeaways

- **Generalization bounds** are crucial for understanding and controlling the performance of machine learning models. [15]
- They guarantee the learned hypothesis will perform well on unseen data. [15]