Generalization Bounds

Theoretical Foundations of Deep Learning

Matteo Mazzarelli

December 17, 2024

Motivation

- **Core Challenge**: How can a model learned from *limited training data* perform well on *unseen data*?
- Generalization lies at the heart of the machine learning process.
- A poorly generalized model risks:
  - **Overfitting**: Performing well on training data but poorly on unseen data.
  - **Underfitting**: Failing to capture the underlying patterns of the data.

The Learning Problem

- **Supervised Learning**:
  - Goal: Learn a function $f : X \to Y$ mapping inputs X to outputs Y based on labeled training data.
- **Key Question**: Can the learned function perform well on unseen data?
- **Generalization**:
  - Ability of a model to extend its learning beyond the training data.
  - **Central Problem** in machine learning: balancing *empirical performance* with *future predictions*.

Why Theory Matters

- **Significance of Theory**:
  - Guides **algorithm design** by providing a foundation for developing new methods.
  - Allows **performance analysis** to identify the strengths and weaknesses of algorithms.
  - Reveals **limitations** of learning systems, helping us understand their boundaries.
- **Theoretical Understanding**:
  - Bridges the gap between empirical performance and guarantees on future behavior.

Simulating Overfitting

- **Objective**:
  - Visualize the impact of model complexity on overfitting in a linear regression model.
- **Dataset**
  - The experiment uses the **Boston Housing dataset**, where the target variable is `medv` (median value of owner-occupied homes), and the features represent housing characteristics.

- **Experimental Setup**:
  - **Model Complexity**:
    - Complexity is defined by the number of features included in the model.
    - Additional random features are generated to simulate increasing complexity beyond the real features in the dataset.
  - **Train-Test Split**:
    - The dataset is split into 70% training and 30% test data.
  - **Range of Complexity**:
    - Models are trained with 1 to 200 features. Beyond the actual features in the dataset, random noise features are added incrementally.

- **Procedure**: For each level of complexity:
  1. A subset of features (real and random) is used to train a linear regression model.
  2. Predictions are made on both the training and test datasets.
  3. Mean Squared Errors (MSE) are calculated for both datasets.
- **Results**:
  - Training error decreases consistently as model complexity increases.
  - Test error initially decreases but then increases, demonstrating the overfitting phenomenon.
- **Visualization**:
  - A line plot shows the relationship between model complexity (number of features) and mean squared error for both the training and test datasets.

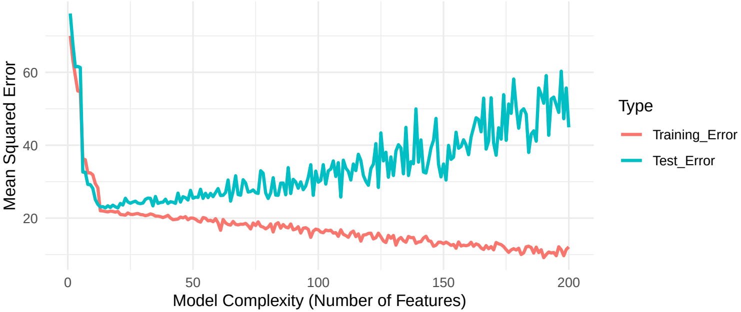

Figure 1: Overfitting Phenomenon in Linear Regression

- **Highlights**:
  - The **bias-variance tradeoff**.
  - The point where overfitting begins, indicated by the divergence of training and test errors.

Key Insights

- Increasing model complexity without consideration of the underlying data structure can lead to overfitting.
- Simple models that focus on the true underlying pattern often generalize better to unseen data.

Double Descent

- However, modern machine learning introduces a fascinating twist: **Double Descent**, where increasing model complexity can sometimes lead to improved generalization after an initial overfitting phase.
- Unlike traditional models where increasing complexity leads to overfitting, further increasing the complexity (e.g., using overparameterized neural networks) can eventually reduce the generalization error after an initial peak.
- This challenges the classical view of overfitting and highlights the complex relationship between model complexity and generalization in modern machine learning.

Key Insights

- While simple models often underfit and overly complex models overfit, the phenomenon of **Double Descent** shows that extremely complex models can sometimes achieve superior generalization, especially in overparameterized regimes.

Introducing Generalization Bounds

- **What Are Generalization Bounds?**
  - Theoretical tools offering guarantees about a model's performance on unseen data.
  - Relate:
    - **Generalization Error**: How well the model performs on unseen data.
    - **Empirical Risk**: Performance observed on training data.
    - **Model Complexity**: How expressive the model is.
    - **Approximation Limits**: What kinds of functions the model class can represent, as explained by Approximation Theory.

- **Purpose**:
  - Provide insights into the trade-offs between:
    - **Model Accuracy**: How well the model captures the data patterns.
    - **Model Complexity**: The expressiveness of the model and its ability to fit intricate patterns.
    - **Training Data Size**: How much data is required to achieve reliable generalization.
  - Approximation Theory informs this by helping us understand the limits of a model class:
    - Simpler models may not be able to represent complex data (high approximation error).
    - Overly complex models risk overfitting, where generalization error increases despite better fit on training data.

Hoeffding's Inequality: A Starting Point

- **What is Hoeffding's Inequality?**
  - A fundamental result in probability theory used to bound the difference between the **empirical risk** and the **generalization error** for a fixed hypothesis.
  - Provides a way to measure how closely a model's performance on training data reflects its performance on unseen data.

Mathematical Formulation of Hoeffding's Inequality

- **Hoeffding's Inequality**:

$$P(|R(h) - R_{\text{emp}}(h)| > \varepsilon) \leq 2\exp(-2m\varepsilon^2)$$

  - $R(h)$: Generalization error (true performance on unseen data).
  - $R_{\text{emp}}(h)$: Empirical risk (error on training data).
  - ε: A small positive value (tolerance).
  - m: Size of the dataset.

Key Insights

- The probability that the generalization error $R(h)$ deviates significantly from the empirical risk $R_{\text{emp}}(h)$ decreases **exponentially** with:
  - Larger dataset size m.
  - Smaller tolerance ε.

Rates of Convergence

- **What Are Rates of Convergence?**
  - Quantify how quickly the generalization error approaches the empirical risk as the dataset size m grows.
  - In Hoeffding's inequality:

$$P(|R(h) - R_{\text{emp}}(h)| > \varepsilon) \leq 2\exp(-2m\varepsilon^2)$$

    - The **exponential term** $\exp(-2m\varepsilon^2)$ shows that the convergence is faster with larger datasets.
- **Key Factors**:
  - **Dataset Size (m)**: Larger datasets reduce the gap between $R(h)$ and $R_{\text{emp}}(h)$ more quickly.
  - **Tolerance (ε)**: Smaller tolerances require larger datasets for the same level of confidence.
- **Practical Insight**:
  - Rates of convergence provide a guideline for determining the dataset size needed to achieve a desired level of generalization.

Interpretation of Hoeffding's Inequality

- **What Does It Mean?**
  - As the dataset size (m) increases, the empirical risk becomes a more reliable indicator of the generalization error.
  - For a fixed hypothesis, we can be confident that the performance observed on training data is close to what can be expected on unseen data.
  - The **rate of convergence** shows how quickly this reliability improves as m grows.

Key Insights

- Hoeffding's inequality gives a **quantitative guarantee** about the relationship between training performance and unseen data performance.
- Understanding convergence rates helps in planning how much data is needed for robust generalization.

Limitations of Hoeffding's Inequality

- **The Challenge of Multiple Hypotheses**:
  - In practical machine learning, we often choose the best hypothesis from a large hypothesis class $\mathcal{H}$.
  - Hoeffding's inequality applies to a **single fixed hypothesis**, not to the case where multiple hypotheses are considered.
- **Implication**:
  - It doesn't directly address:
    - The **selection bias** introduced by choosing the hypothesis that minimizes the empirical risk.
    - The increased risk of overfitting when evaluating multiple hypotheses.

Limitations of Hoeffding's Inequality

- **Beyond Fixed Hypotheses**:
  - Hoeffding's inequality assumes a single, fixed hypothesis.
  - In practice, machine learning involves selecting the best hypothesis from a **large hypothesis class** $\mathscr{H}$, increasing the risk of overfitting.
- **Need for Complexity-Aware Bounds**:
  - Simple bounds like Hoeffding's do not consider the complexity of the hypothesis class, which influences generalization.

Motivation for Advanced Bounds

- **Advanced bounds** address this by incorporating:
  - **VC Dimension**: A measure of the capacity or expressiveness of a hypothesis class. Higher VC dimensions indicate more complex models, which may require more data to generalize well.
  - **Rademacher Complexity**: A data-dependent measure of how well a hypothesis class can fit random noise in the training data. It captures both the hypothesis class and the specifics of the data distribution.

- **Extending Convergence Rates**:
  - Advanced bounds refine the rates of convergence by linking the generalization error to:
    - The size of the dataset m.
    - The complexity of the hypothesis class (e.g., **VC dimension** or **Rademacher complexity**).
  - For example, the generalization error is often bounded as:

$$R(h) - R_{\text{emp}}(h) \leq \mathcal{O}\left(\sqrt{\frac{\text{Complexity}(\mathcal{H})}{m}}\right)$$

    - Larger datasets m reduce error, but higher complexity increases the required data for a desired level of generalization.
- **Practical Implications**:
  - These bounds provide actionable insights for balancing model complexity and dataset size.