The Minimum-Description-Length (MDL) Principle

Kiran Bagale

June 2025

Outline

What is the MDL Principle?

- A method for **model selection** in statistical modeling.
- Pioneered by **Jorma Rissanen** in 1978.
- Inspired by **Kolmogorov Complexity Theory**.
- Goal: Find the model that best explains data by balancing **fit** and **complexity**.

Kolmogorov Complexity

Definition

The **algorithmic complexity** of a data sequence is the length of the shortest binary computer program that prints the sequence and halts.

- Focuses on **data compression** rather than probability distributions.
- Provides a foundation for identifying **regularity** in data.

Key Insights of MDL

1. **Learning as Data Compression**: Learning involves finding regularity in data.
2. **Regularity and Compression**: Regularity is identified with the ability to compress data.

MDL Objective

Given a set of hypotheses h and data sequence d, find the hypothesis that maximizes compression of d.

Two-Part Code MDL Principle

- Simplest and most well-known version of MDL.
- For a model class m with probability density functions $p \in m$, minimize:

$$L_{12}(p, d) = L_1(p) + L_2(d|p)$$

- $L_1(p)$: Description length of the hypothesis p.
- $L_2(d|p)$: Description length of data d encoded with p.

Model-Order Selection Problem

- Goal: Identify the **best model** from a family of linear regression models $m(1), m(2), \ldots, m(k)$, where k is the model order.
- Models have parameter vectors $w(k)$ with increasing dimensionality.
- Use training sample $\{x_i, d_i\}_{i=1}^{N}$, where x_i is the stimulus and d_i is the response.

MDL for Model-Order Selection

Objective

Minimize the composite description length:

$$\min_k \left(\sum_{i=1}^{N} - \log p(d_i | w(k)) \phi(w(k)) + \frac{k}{2} \log(N) + O(k) \right)$$

- **Error Term**: $-\log p(d_i | w(k)) \phi(w(k))$, measures model-data fit.
- **Complexity Term**: $\frac{k}{2} \log(N) + O(k)$, measures model complexity.
- For large N, the $O(k)$ term is often ignored for simplicity.

Key Attributes of MDL

1. **Occam's Razor**: Prefers the *simplest* model that fits the data well.
2. **Consistency**: Converges to the true model order as sample size N increases.

Practical Performance

MDL rarely produces anomalous results and is effective for linear regression models.

Applying MDL in Practice

- **Simplification**: The $O(k)$ term is often ignored, but this can lead to mixed results.
- **Tie-Breaking**: If multiple models minimize $L_{12}(p, d)$, choose the one with the smallest complexity term.
- **Efficiency**: For linear regression, the complexity term can be computed explicitly, improving performance.

Summary

- The MDL principle is a powerful tool for **model selection**.
- Balances **model fit** and **complexity** using data compression.
- Rooted in **Kolmogorov Complexity** and implements **Occam's Razor**.
- Consistent and reliable, especially for linear regression models.

References

- Rissanen, J. (1978). Modeling by shortest data description.
- Kolmogorov, A. N. (1965). Three approaches to the definition of the concept “quantity of information”.
- Grünwald, P. D. (2007). The Minimum Description Length Principle.
- Li, M., & Vitányi, P. (1993). An Introduction to Kolmogorov Complexity and Its Applications.
- Cover, T. M., & Thomas, J. A. (2006). Elements of Information Theory.