S5: Minimum Description Length (MDL) and Real Life Applications

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.*

Scribe: 5

Student: Pranjal Umesh Kalekar

CS 7840: Foundations and Applications of Information Theory (fa24)

https://northeastern-datalab.github.io/cs7840/fa24/

Lecturers: Wolfgang Gatterbauer, Javeed Aslam

Version Dec 11, 2024

MDL - Minimum Description Length

The MDL principle is a powerful statistical framework for inductive inference, with roots in information theory. It aims to find **regularities and patterns** in data by identifying the **most concise representation** of the information.

At its core, MDL views learning as a form of data compression. The best explanation for a set of data is the one that allows us to describe it using the fewest bits. The model that achieves this is considered to have captured the most relevant structure in the data.

Does this ring bell about something? – Kolmogorov complexity!

working of MDL and fit and complexity tradeoff

MDL relies on a fundamental connection between probability distributions and code lengths. A shorter code is assigned to events that are more probable according to a given distribution.

Conversely, less probable events receive longer codes. This is formalized by the Kraft Inequality, which establishes a correspondence between probability mass functions and code length functions.

Given a set of candidate models, MDL selects the one that minimizes the total description length. This length is composed of two parts:

- The description length of the model itself.
- The description length of the data given the model.

The key to MDL's success lies in striking the right balance between fit and complexity. A more complex model might fit the data better but requires more bits to describe. MDL seeks a model that is complex enough to capture the essence of the data but simple enough to avoid overfitting.

Practical implementation Of MDL Refined MDL and Universal Codes

Crude vs. Refined MDL: Early formulations of MDL ("crude MDL") relied on arbitrary two-part codes for describing models and data. This introduced subjectivity and led to suboptimal results in practice. Refined MDL addresses these issues by employing universal codes.

Universal Codes: A universal code can compress any sequence of data almost as well as the best code specifically designed for that sequence. This allows us to compare different models fairly, without giving undue preference to any particular one.

Normalized Maximum Likelihood (NML): One of the most important universal codes is the NML code. It has a desirable property: it treats all distributions within a given model equally, ensuring fair comparison between models.

Entanglement with day-to-day decision making

- Talking in biking terms – Suppose I am planning to decide whether to go on bike ride or no. My way of making the decision involves 2 major thought processes:
  - Process 1:
    - I am thinking about how I am feeling considering nutrition intake, tiredness, previous workouts during the day, sleep and emotional state.
    - Then I check my bike condition – airpressure , batteries for light and other equipment
    - And I check for weather if it's suitable for a bike ride
  - Process 2:
    - I check the weather and then the bike conditions and then the bodily conditions
- It's fairly clear that the thing swhich are in my hand to fix can be 2nd priority and if the weather is bad it's highly possible to make the decision in the 1st step itself!
- MDP works in the same way– Normalizing(NML) gives you what has mostly worked in the past