Inductive Bias in Model Selection

Muhammadsharif Nabiev
Department of Intelligent Systems
MIPT
nabiev.mf@phystech.edu

Abstract

This paper tackles the problem of constructing suboptimal models within multitask learning paradigm. Given a set of intrinsically related tasks, the goal is to uncover a shared structure—known as the inductive bias—across all tasks. By applying constrained evolutionary symbolic regression, we show that these models can be decomposed in a way that reveals the tasks' underlying structure. We performed experiments on synthetic data, where the data structure, generation process, and optimal models are known. The results indicate that the models produced by the proposed method are indeed optimal for their respective tasks.

1 Introduction

The concept of inductive bias is a fundamental tenet in the realm of machine learning, encapsulating the core assumptions that underpin the methodology adopted by a particular model in its predictive endeavours, extending beyond the boundaries of explicitly observed data. Understanding and leveraging inductive bias is essential for enhancing model performance, especially in complex environments where data may exhibit diverse characteristics.

In recent years, the field of machine learning has experienced rapid advancements driven by the development of sophisticated algorithms and architectures capable of tackling a wide range of tasks, from natural language processing to computer vision. However, the design and optimization of these models often require significant expertise and resources, leading to the rise of automated machine learning (AutoML) systems. These systems aim to alleviate the burden of manual model selection and tuning, enabling broader accessibility to machine learning techniques.

One notable approach is AutoML-Zero, which autonomously constructs models using a genetic programming framework, assembling them from fundamental mathematical operations [3]. This method represents a significant departure from traditional model selection paradigms, which typically rely on predefined structures or human intuition. By allowing the model architecture to emerge organically from the data, AutoML-Zero minimizes biases introduced by prior knowledge, potentially leading to the discovery of novel architectures better suited to the task at hand.

The concept of inductive bias plays a crucial role in model generalization. Different datasets often possess unique distinguishing features that can be exploited for improved performance. For instance, models trained on image data may benefit from biases related to spatial hierarchies, while those handling sequential data, such as time series or text, may rely on temporal dependencies. Recognizing and systematically integrating these biases into the model selection process can facilitate the identification of suboptimal yet effective architectures.

In this paper, we investigate how inductive biases inherent in the data can inform model selection within the multitask learning framework. We propose an automated approach that leverages evolutionary algorithms in conjunction with symbolic regression to explore a diverse range of model architectures. By allowing models to train independently on various datasets within a specific class, our methodology aims to enhance generalizability and adaptability while reducing the need for extensive manual tuning.

The experiments are focused on a range of datasets that capture different characteristics, enabling a robust evaluation of how well our models generalize across tasks. The goal is to provide insights into the relationship between inductive bias and model architecture. [1]
About inductive bias, multitask, symbolic regression, info bottleneck, and llm interpretability.

2 Problem statement

Let $\mathfrak{T} = \{T_1, T_2, \ldots, T_n\}$ be a set of tasks. Each task T_i has its corresponding dataset $\mathfrak{D}_i = \{(\mathbf{x}_j, \mathbf{y}_j)\}_{j=1}^{N_i}$ with and object $\mathbf{x}_j$ and a label $\mathbf{y}_j$.

A *model* $\mathbf{f}$ is defined as an expression with the following mapping $\mathbb{R}^n \to \mathbb{R}^m$, whose *structure* Γ can be represented via expression tree [4]. In search of such models, we will employ an evolutionary search algorithm to discover a general model — up to constant factors — that can address the tasks $\mathfrak{T}$.

Let $\mathfrak{F}$ be a parametric family of all models which can be decomposed as such $\mathbf{f} = \mathbf{g} \circ \mathbf{h}$, where $\mathbf{g}$ and $\mathbf{h}$ will be referred to as *decoder* and *encoder*. It is also worth noting that any $\mathbf{f}$ can be decomposed in the prescribed way.

Theorem 2.1. *For any expression $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^m$ there exists $\mathbf{h} : \mathbb{R}^n \to \mathbb{R}^d$ and $\mathbf{g} : \mathbb{R}^d \to \mathbb{R}^m$ such that $\mathbf{f} = \mathbf{g} \circ \mathbf{h}$.*

Proof. Trivial cases:

- $d = m$. We take $\mathbf{h} = \mathbf{f}$ and $\mathbf{g} = \mathrm{Id}$, hence $\mathbf{f}(\mathbf{x}) = \mathbf{g}(\mathbf{h}(\mathbf{x})) = \mathbf{g}(\mathbf{f}(\mathbf{x})) = \mathbf{f}(\mathbf{x})$.
- $d = n$. We take $\mathbf{h} = \mathrm{Id}$ and $\mathbf{g} = \mathbf{f}$, hence $\mathbf{f}(\mathbf{x}) = \mathbf{g}(\mathbf{h}(\mathbf{x})) = \mathbf{g}(\mathbf{x}) = \mathbf{f}(\mathbf{x})$

For non-trivial cases the decomposition can be obtained using Kolmogorov-Arnold theorem [2]. □

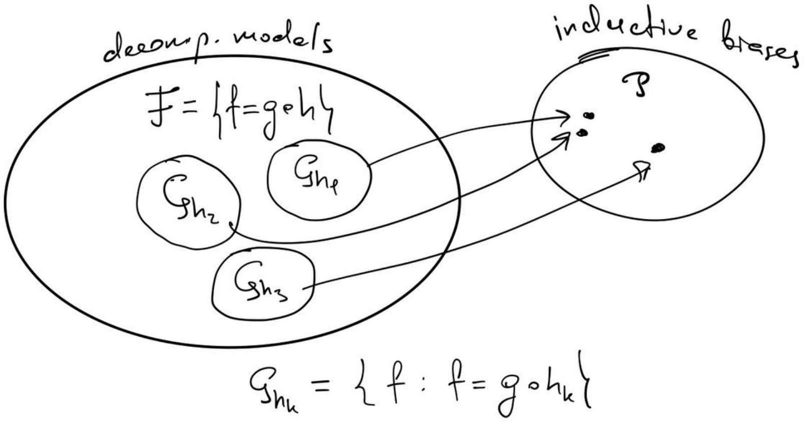

Figure 1: Models coresponding to inductive biases.

By the nature of the given decomposition we have to put restrictions on encoder, otherwise when certain conditions are fulfilled we can fall into degenerate cases for the structure of the decoder.

Theorem 2.2. *Let* $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^m$*. For a latent space* $\mathbb{R}^d$ *such that* $d \geq m$ *we can construct a decomposition* $\mathbf{f} = \mathbf{g} \circ \mathbf{h}$ *such that* $\mathbf{g} = \mathrm{Id}_Z$ *and* $Z \cong \mathbb{R}^m$*.*

Proof. Since $m \leq d$ we can redefine latent space as $Z \cong \mathbb{R}^m$ and the given expression $\mathbf{f}$ by ignoring extra dimensions. If we take $\mathbf{h} = \mathbf{f} : \mathbb{R}^n \to Z$ and $\mathbf{g} = \mathrm{Id}_Z$ we get $\mathbf{f}(\mathbf{x}) = \mathbf{g}(\mathbf{h}(\mathbf{x})) = \mathrm{Id}_Z(\mathbf{f}(\mathbf{x})) = \mathbf{f}(\mathbf{x})$. □

By Theorem 2.2 it follows that we have to restrict the dimensionality of a latent space. In order to do that we incorporate *Information Bottleneck* method.

2.1 Information Bottleneck

A statistic $\mathrm{S}(\mathbf{X})$ is called *sufficient* if the mutual information between $\mathrm{S}(\mathbf{X})$ and $\mathbf{Y}$ is the same as the mutual information between $\mathbf{X}$ and $\mathbf{Y}$, i.e. $\mathrm{I}(\mathbf{X}, \mathbf{Y}) = \mathrm{I}(\mathrm{S}(\mathbf{X}), \mathbf{Y})$. A sufficient statistic is called *minimal* if it minimizes information about input data $\mathbf{X}$:

$$\mathrm{T}(\mathbf{X}) = \underset{\mathrm{S}(X) : \mathrm{I}(\mathrm{S}(\mathbf{X}), \mathbf{Y}) = \mathrm{I}(\mathbf{X}, \mathbf{Y})}{\arg\min} \mathrm{I}(\mathrm{S}(\mathbf{X}), \mathbf{X}).$$

Information bottleneck principle aims to find an internal representation that maximizes information about $\mathbf{Y}$ while minimizing information about $\mathbf{X}$, i.e.

$$\min_{p(T|\mathfrak{D})} \mathrm{I}(\mathbf{X}, \mathrm{T}(\mathbf{X})) - \beta\, \mathrm{I}(\mathrm{T}(\mathbf{X}), \mathbf{Y}).$$

Suppose $\mathcal{L}_{\text{task}}$ is a loss of the task being solved. It is worth noting that the task-loss could not be differentiable. Taking informational bottleneck approach into account we get the final optimization function:

$$\begin{aligned} \Gamma = \underset{f=g\circ h:\ \Gamma-\text{structure of } h}{\arg\min} \quad & \mathbb{E}_{\mathfrak{D}} \mathcal{L}_{\text{task}}(f(X), Y) \\ \text{s.t. } \beta\, \mathrm{I}(h, Y) - \mathrm{I}(X, h) \to \max & \end{aligned}$$

We define *inductive bias* as the structure of $\mathbf{h}$. The goal is to construct models which general form can approximate the given tasks. In order to favor generalization an informational bottleneck approach was taken.

Having multi-task paradigm is crucial for our decomposition. Single-task setting will degenerate our solution to having $\mathbf{g} = \mathrm{Id}$.

Theorem 2.3. *For single-task paradigm there exists a solution* $\mathbf{f}$ *that has a decomposition* $\mathbf{f} = \mathbf{g} \circ \mathbf{h}$*, where* $\mathbf{g} = \mathrm{Id}$ *and* $\mathbf{h} = \mathbf{f}$*.*

Proof. For single-task paradigm the optimal solution $\mathbf{f}$ can itself represent the encoder h. Hence $\mathbf{h} = \mathbf{f}$ and $\mathbf{g} = \mathrm{Id}$. □

Theorem 2.3 immediately gives us the following corollary.

Corollary 2.3.1. *For multi-task paradigm where all tasks can be generalized by single* $\mathbf{f}$*, there exists a solution with* $\mathbf{g} = \mathrm{Id}$*.*

3 Computational experiment

We used PySR library for to construct model using evolutionary search [1]. The experiments were conducted on binary classification, autoregression and CNN tasks. For each task the meta-learner trained on 1, 2, 5, 10 and 15 datasets. To incorporate our objective we specified custom loss along with template expression to extract the mapping $\mathbf{h}$.

We conduct experiments on three types of data modalities: tabular, sequential and image-based.

3.1 Data

3.2 Experiments

3.2.1 Tabular

We used dataset consisting of circles. Multi-task involved 3 tasks. Binary classification and 2 custom function regressions.

The decision boundary for datasets was chosed to be a circle for its simplicity and practicality. The circle analitically can be represented as a second-order equation.

3.2.2 Sequential

Simple AR(1) was chosed to generated synthetic datasets.

3.2.3 Image-based

We consider a primitive binary image-tasks with 4/10 pixels. There we must find a distinct pattern

3.3 Results of the experiments

4 Conclusion

References

[1] M. Cranmer. Interpretable machine learning for science with pysr and symbolicregression.jl. *ArXiv*, abs/2305.01582, 2023.

[2] Andrey Nikolaevich Kolmogorov. *On the representation of continuous functions of several variables by superpositions of continuous functions of a smaller number of variables.* American Mathematical Society, 1961.

[3] Esteban Real, Chen Liang, David So, and Quoc Le. Automl-zero: Evolving machine learning algorithms from scratch. In *International conference on machine learning*, pages 8007–8019. PMLR, 2020.

[4] G. I. Rudoy and V. V. Strijov. Algorithms for inductive generation of superpositions for approximation of experimental data. *Informatics and Applications*, 7(1):44–53, 2013.