December, 2023

IA703

Algorithmic Information & Artificial Intelligence

Micro-study

teaching.dessalles.fr/FCI

Name: Paul Aristidou, Olivier Lapabe-Goastat

Fine-Tuning Regression Models for Peak Performance

Link to GitHub repository

Abstract

'Fine-Tuning Regression Models for Peak Performance' is a detailed investigation aimed at identifying the 'best model' in terms of complexity. This project systematically assesses various regression models, including polynomial, logarithmic, and exponential types, to determine the optimal balance between model complexity and predictive accuracy.

Problem

Our objective addresses the challenge of identifying the optimal regression model that achieves minimum description length (MDL), essentially a model that encapsulates the underlying data patterns with minimal complexity. Another interesting objective would be to observe whether a model with the smallest MDL also delivers the best performance in terms of mean squared error (MSE). This reflects the quest to find a model that not only succinctly captures the essence of the data but also predicts with high accuracy.

Method

Before delving into our methodology, we address the calculation of the Minimum Description Length (MDL), a crucial metric for model selection. In this framework, we calculate the MDL as a two-part code length:

- The first part, **the complexity of the model** $C(M_j)$, considers the number of parameters and their encoding lengths.

- The second part, **the complexity of the data given the model** $\sum_k C(x_k|M_j)$, is assessed using the Euclidean distance between the observed data points and the model predictions.

The model that minimizes the total description length, as shown below, is considered the most suitable:

$$M = \arg\min_j \left\{ C(M_j) + \sum_k C(x_k|M_j) \right\}$$

Now, to achieve our goal of finding the best regression model, we have established the following methodology :

- We begin by **scaling the feature variables**, a crucial step that normalizes the data and ensures that each variable contributes equally to the analysis, regardless of their original scale.
- Our next phase involves **constructing a variety of regression models**. Each model applies different transformations to the data: we explore polynomial, logarithmic, and exponential models to accommodate various underlying data relationships. Within each model type, we **experiment with a range of polynomial degrees**. This allows us to examine different model complexities, from simple linear relationships to higher-degree equations.
- The **training process** for each model follows, where the models are fitted to the training dataset. This fitting process is where the models 'learn' from the data, adjusting their parameters to best represent the data's structure.
- We then compute for each model the **minimum description length (MDL)**.
- We then compute the **Mean Square Error (MSE)** on a separate test dataset.
- Finally, we engage in a **comparative analysis** where we pit the models against each other to see which strikes the optimal balance. The 'best' model, in our context, is the one that not only predicts with high accuracy (as indicated by a low MSE) but also maintains a lower complexity (lower MDL).

N.B: Within each model type and for each degree of polynomial, we fine-tuned the coefficients to include multiple decimal places. This step ensured that we explored a comprehensive range of model specifications, enhancing the precision of our regression models and potentially improving their performance. What we observe is that it was best to take 0 decimal places.

Results

We will now create different datasets, and we will try different regression models. For each of these, we will calculate the Complexity and the Mean Squared Error. We will then see for which models the Complexity and the Mean Squared Error are minimized.

We will generate different datatsets to try different configurations. Each dataset will be generated around a function, to which some gaussian noise will be added. We will display only 3 datasets in this report, please feel free to see the 3 others in the annex.

Dataset 1: $y = 1 + 2x + 3x^2 - 4x^3 + noise$

In the figure below, you can see the formed dataset with 1000 points.

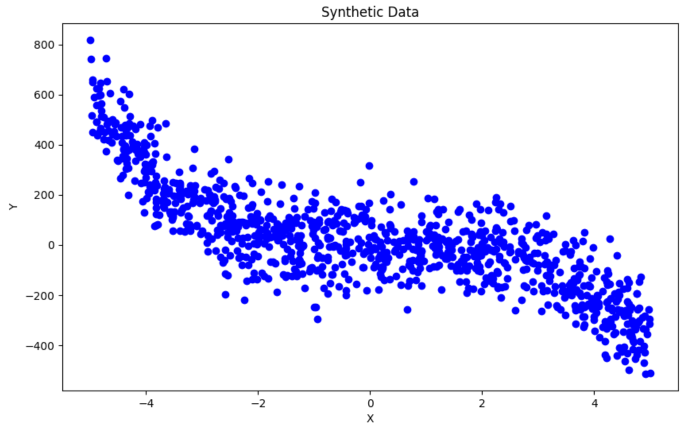

Figure 1: Dataset 1

For each regression model, we calculated the complexity with 0 to 4 decimals. In the figure below, you can see the results for a polynom of degree 2. The model where all parameters are integers is the most simple one, minimizing the complexity.

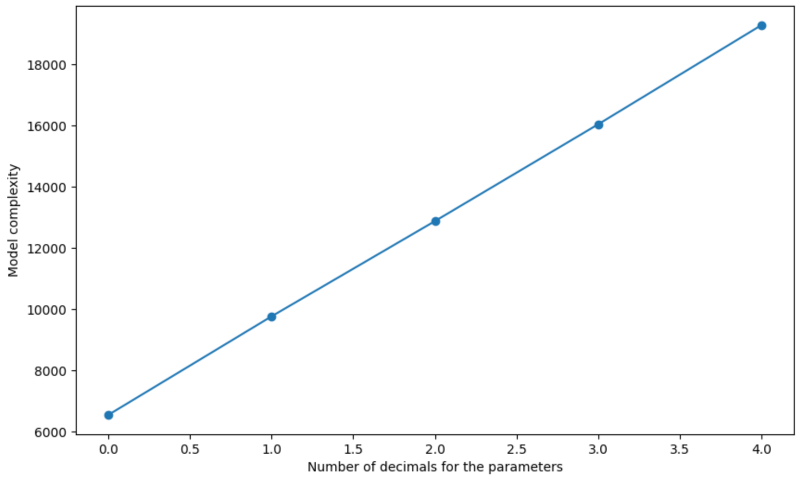

Figure 2: Dataset 1 - Complexity vs. number of decimals for the parameters

We see a similar result for all regression models, and for all other datasets. Therefore we show the results only for Dataset 1, and always keep parameters as integers.

We will now run fit the polynomial, exponential and logarithmic regressions, and let's see which ones minimizes the Complexity.

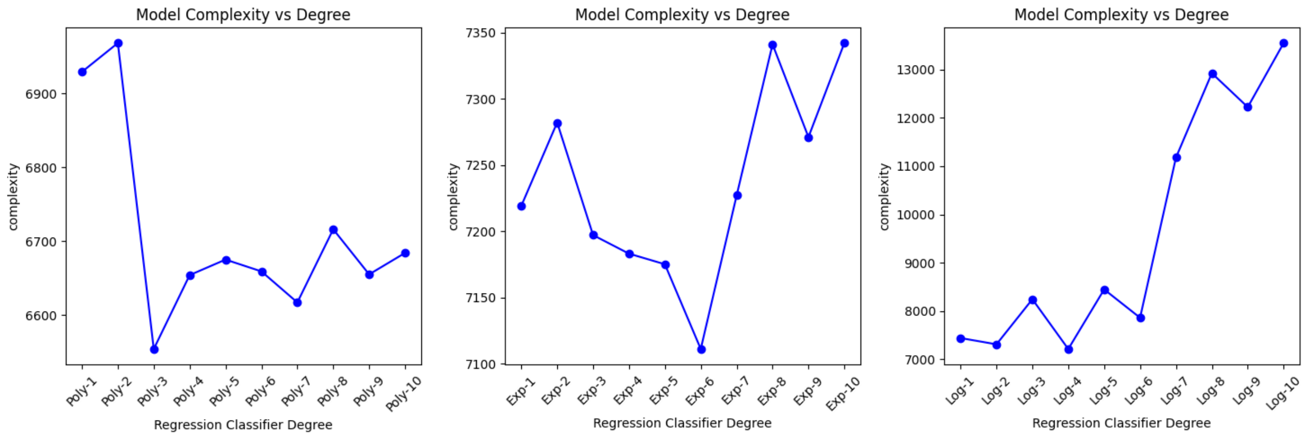

Figure 3: Dataset 1 - Complexity for different regression models

The model which minimizes the Complexity is the polynom of degree 3. This fits perfectly with the function that generated the dataset.

We will now look at which regression model minimizes the Mean Squared Error.

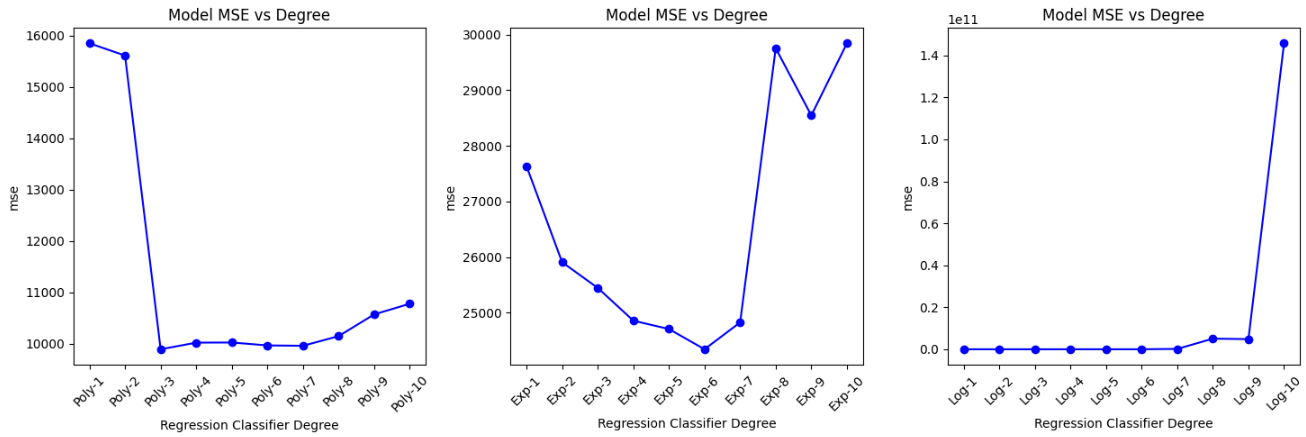

Figure 4: Dataset 1 - MSE for different regression models

The exact same regression model (polynom degree 3) minimizes as well the MSE.

We will finally plot the chosen polynom of degree 3 as "best" model.

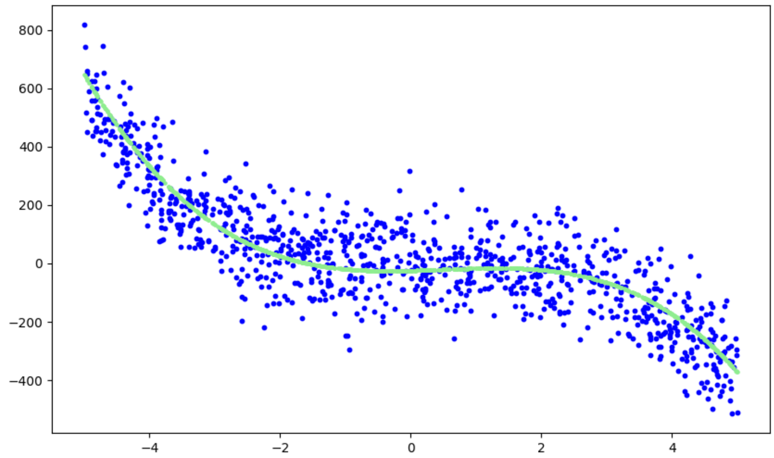

Figure 5: Dataset 1 - "Simplest" regression model minimizing Complexity

As conclusion, both Complexity and MSE approaches reach their minimum for the polynom of degree 3. The "simplest model" (minimizing the Complexity) is also the "closest" model (minimizing the MSE).

Let's now look at two other datasets.

Dataset 2: $y = 1 - exp(x) + 2exp(x)^2 - 4exp(x)^3 + noise$

As previously, the figures below show the formed dataset with 1000 points, as well as the Complexity and the MSE for each regression model.

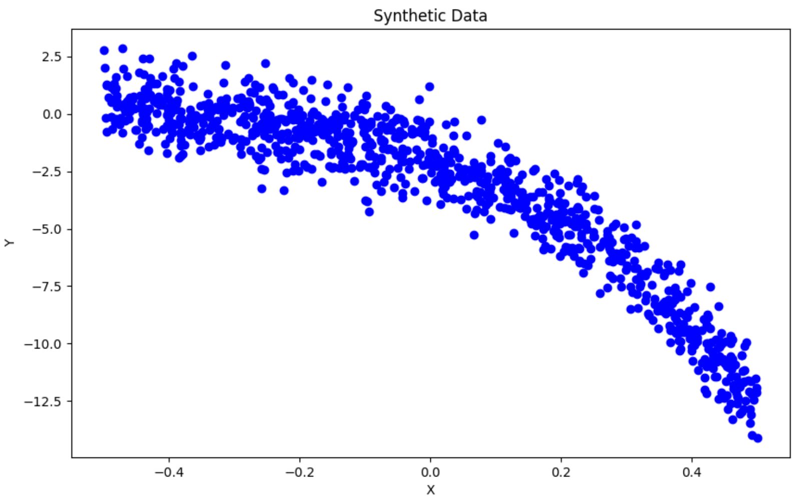

Figure 6: Dataset 2

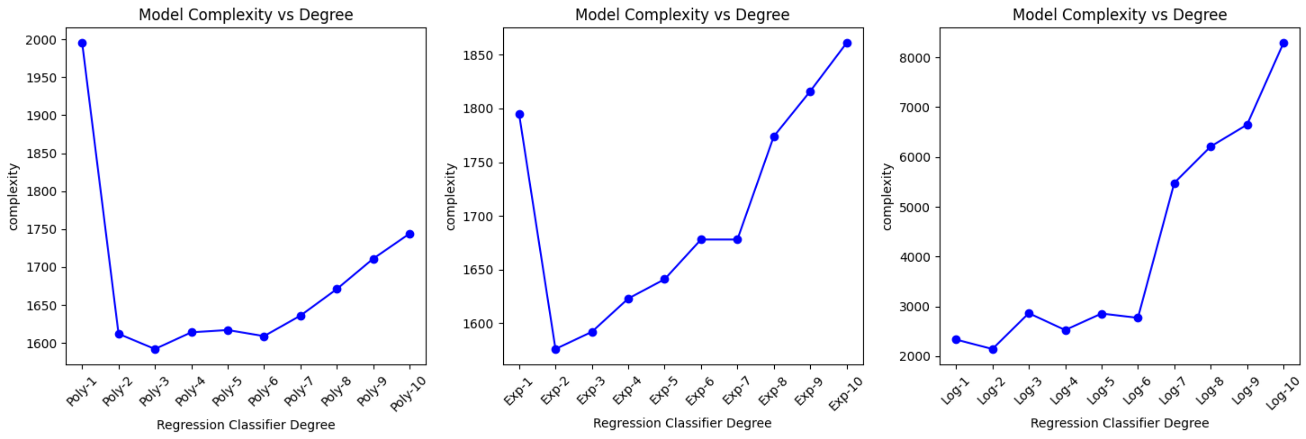

Figure 7: Dataset 2 - Complexity for different regression models

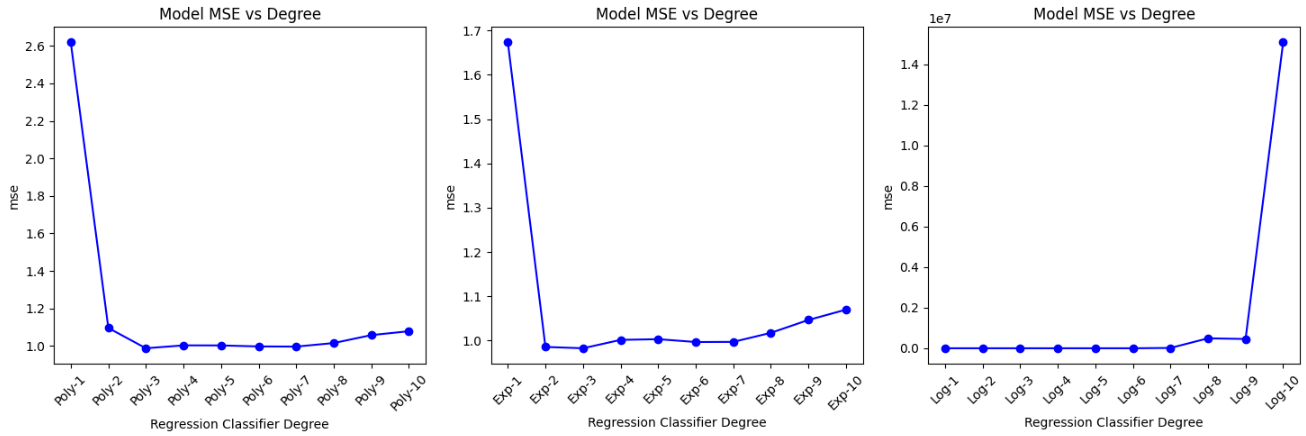

Figure 8: Dataset 2 - MSE for different regression models

For this dataset, the exponential regression model of degree 2 (Exp-2) is minimizing the complexity.
Looking at the MSE, Exp-2 is reaching 0.985 while Exp-3 is reaching 0.982. Both models are very close while the exponential regression model of degree 3 is minimizing the MSE.
This is an interesting result : eventhough Exp-3 is "closest" to the dataset, Exp-2 is the "simplest" one, and still very close to the dataset (almost as much as Exp-3). Therefore, in this case, it might make sense to choose Exp-2, as it simpler than Exp-3 and almost as close than Exp-3 to the dataset.

We will finally plot the chosen exponential polynom of degree 2 as "best" model.

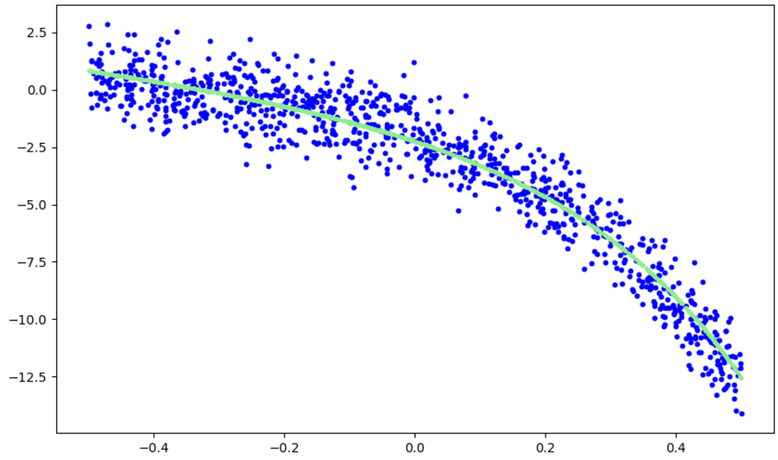

Figure 9: Dataset 2 - "Simplest" regression model minimizing Complexity

As conclusion, Dataset 2 is an interesting example where the "simplest" model (minimizing the Complexity) would be favored over the "closest" model (minimizing the MSE).

Let's now look at the third dataset.

Dataset 3: $y = 1 - 3x^2 - 3exp(x)^2 + noise$

As previously, the figures below show the formed dataset with 1000 points, as well as the Complexity and the MSE for each regression model.

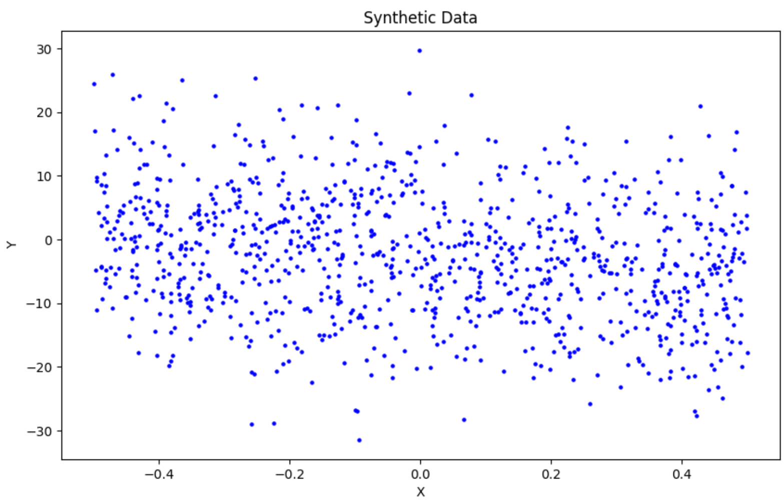

Figure 10: Dataset 3

This time, we have chosen a noise with high variance. As a consequence, the original function from which we generated the dataset is less visible. Let's see if this has an impact on the regression models minimizing the Complexity and the MSE.

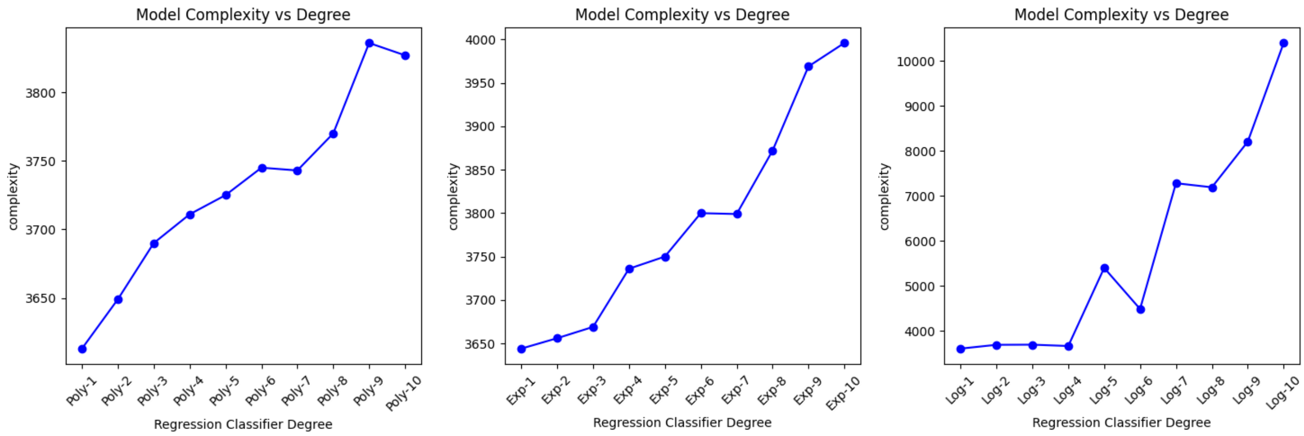

Figure 11: Dataset 3 - Complexity for different regression models

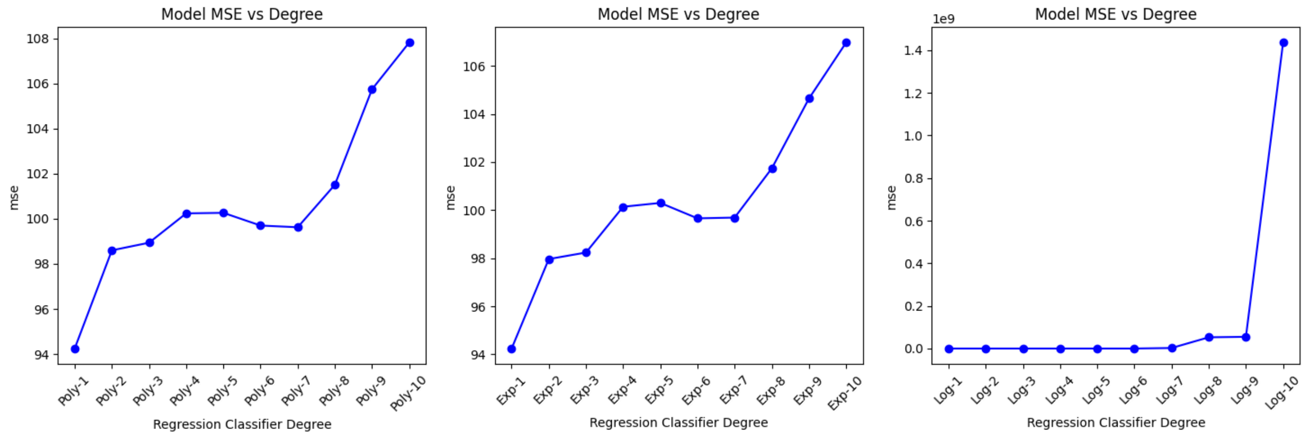

Figure 12: Dataset 3 - MSE for different regression models

We observe that the "best" regression model in the sense of the Complexity and in the sense of the MSE is the same : it's the lineear regression (Poly-1).

It might be a surprising result, given the function that generated the dataset, having a member in x^2 and a member in $exp(x)^2$. We would have expected that either the Poly-2 or the Exp-2 would have been a better fit.

This can be explained by the strong noise that we have applied. The noise is strong and the dataset is reduced between -0.5 and 0.5. As a consequence, the function is flatten by the noise and a simple Linear Regression (Poly-1) is the "best" fitting model (see figure below).

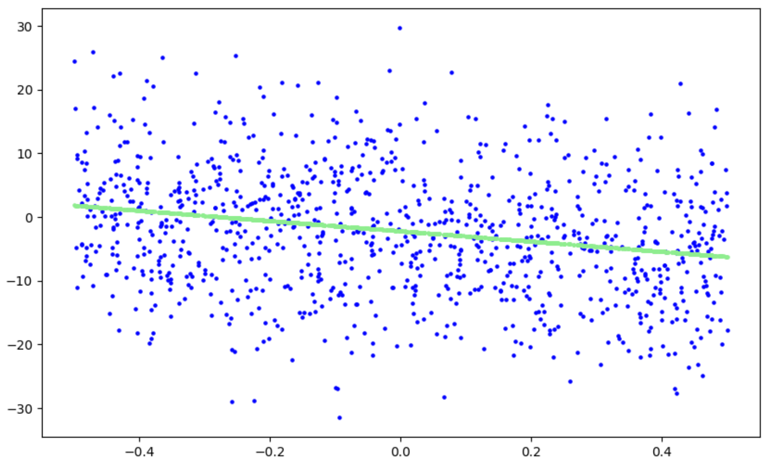

Figure 13: Dataset 3 - "Simplest" regression model minimizing Complexity

Discussion

Comparison with expectations

We have seen that the "best" Regression model in the terms of the MSE is often the same or very close to the "simplest" one in the terms of the Complexity. It confirms the initial hypothesis that minimizing the Description Length / the Complexity is a proper approach to choose the "best" regression model.

However, the MDL approach has a big advantage over the MSE approach. The MDL approach can be used without any test set, because the MDL shows an overfitting by training solely on a train set, without comparing it to a test set. On the contrary, the MSE approach on the train set with always improve with higher model complexity. Without any test set, the MSE approach cannot show an overfitting.

We can then conclude that, in the case when the amount of observations is limited and when we cannot afford to "lose" some data in splitting the dataset into train and test, the MDL is a powerful approach to choose the "best" Regression model without overfitting.

Limits

If the MDL approach has its advantages over the MSE approach, it has also some limits. The MDL approach might be less easy to be interpreted compared to the MSE one, which is more explainable. Also, calculating the Complexity of different models might involve more computational power than "simply" calculating the MSE for each Regression model.

Perspectives

One should consider the specific requirements of the application, the nature of the data, and the consequences of prediction errors. In some cases, the context of the problem may favor one approach over the other.

Also, mixing both approaches MDL and MSE might be something to investigate to get best of both worlds. Different mixing scenarios might be investigated:

- Weighing the 2 indicators and choosing the "best" regression model based on a this mixed weighted indicator.
- Without test set and having only a train set, one could have MSE as main criterion, but keeping a certain threshold of the MDL, stopping when a certain MDL threshold is reached. This would allow a model to train more extensively, but still keeping colmplexity below a certain limit.
- The MDL indicator could be compared to the BIC approach (Bayesian Indicator criterion), the BIC helping to prevent overfitting as well.

Bibliography

IA703 Course: Algorithmic Information and A.I., J-L. Dessalles
Model Selection and the Principle of Minimum Description Length, Mark H Hansen & Bin Yu, 2001
MDL Modeling — An Introduction, Jorma Rissanen, 1992