Machine Learning

|Main Challenges – Underfitting

Prof. Dr. Mostafa Elhosseini

Professor of Smart Systems Engineering

https://youtube.com/drmelhosseini

Agenda

- Underfitting

Main Challenges

- Since your main task is to select a model and train it on some data, the two things that can go wrong are:
  - bad model and
  - bad data

Underfitting

- Underfitting in machine learning occurs **when a model is too simple** to capture the underlying patterns of the data. This situation typically arises when the **model does not have enough parameters** (or complexity) relative to the complexity of the dataset

Underfitting

Underfitting can manifest in various ways, such as:

- **High Bias**: The model makes strong assumptions about the form of the underlying function that generates the data, leading to systematic errors regardless of how much data is used for training. It's biased towards these assumptions and often fails to capture the true relationships in the data.

Imagine trying to **fit a straight line** (like a simple linear regression model) to **data that actually follows a curved path** (like a **parabola**). The straight line is too simple to capture the curve of the data, so it doesn't perform well in representing the dataset's actual complexity. This oversimplification is what we call "**high bias**."

Underfitting

Underfitting can manifest in various ways, such as:

- **Poor Performance**: An underfitted model **performs poorly not just on unseen data** (as in overfitting) but also on the training data itself. This is a key indicator of underfitting, as the model lacks the capacity to learn from the training data.

Underfitting

Underfitting can manifest in various ways, such as:

- **Poor Performance**: An underfitted model **performs poorly not just on unseen data** (as in overfitting) but also on the training data itself. This is a key indicator of underfitting, as the model lacks the capacity to learn from the training data.

Underfitting

To address underfitting, one might consider the following approaches:

- **Increasing Model Complexity**: Moving to a more complex model with more parameters might help **capture the underlying data structure better**. For instance, switching from a linear model to a polynomial or a more complex non-linear model.
- For example, if you are using a **polynomial regression model**, **increasing the degree** of the polynomial adds more parameters to the model, allowing it to fit more complex datasets.

Underfitting

To address underfitting, one might consider the following approaches:

- **Feature Engineering**: Creating new input features through transformations or combinations of existing features can help increase the model's ability to capture the underlying patterns in the data.

Underfitting

To address underfitting, one might consider the following approaches:

- **Reducing Regularization**: If the model is regularized, reducing the regularization strength can allow the model's complexity to increase, potentially addressing underfitting.

Underfitting

To address underfitting, one might consider the following approaches:

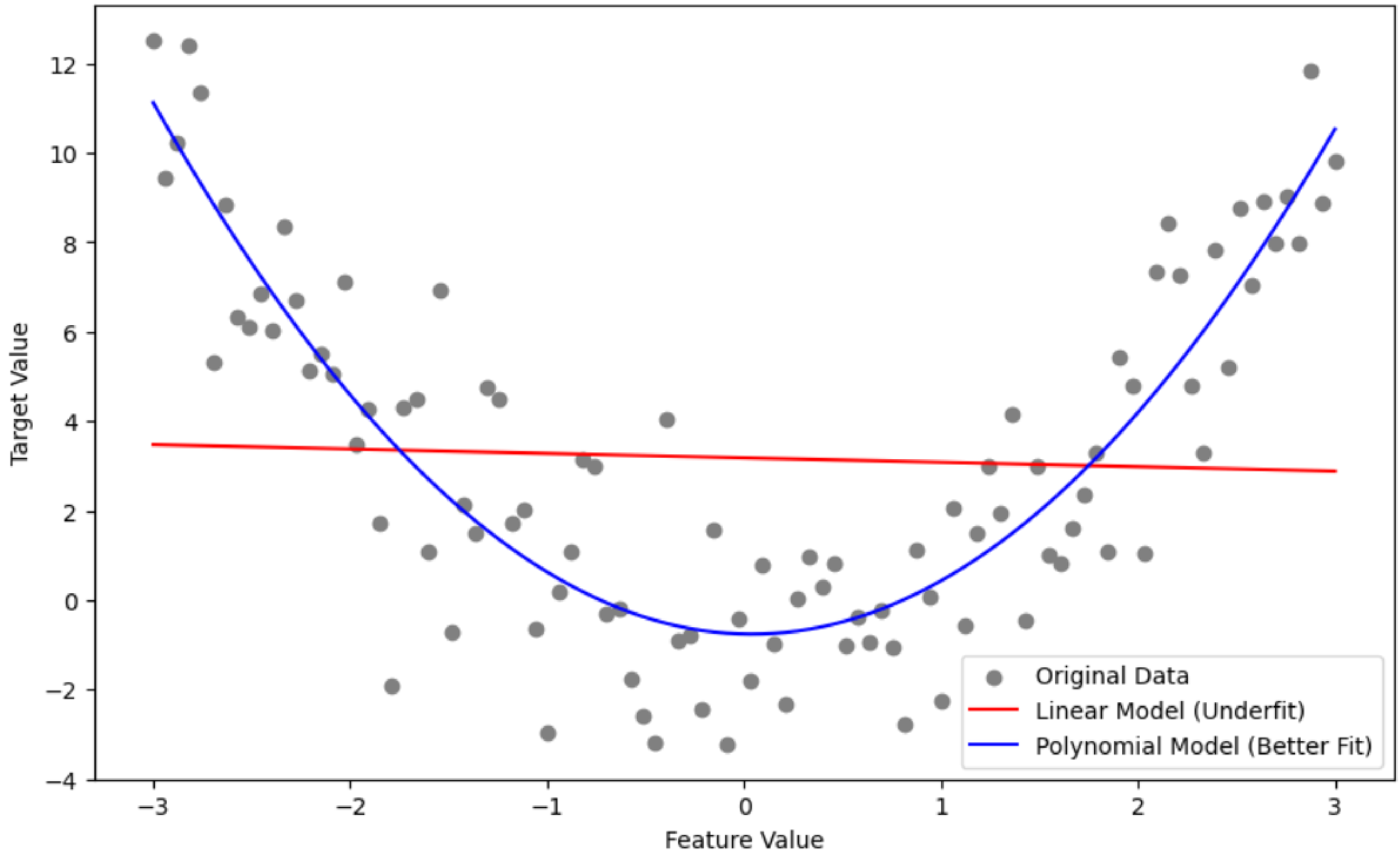
- **Reducing Regularization**: If the model is regularized, reducing the regularization strength can allow the model's complexity to increase, potentially addressing underfitting.

Underfitting

To address underfitting, one might consider the following approaches:

- **Using More Complex Algorithms**: Sometimes, the choice of algorithm is too simple for the task. Using more complex machine learning algorithms that can capture complex patterns might resolve underfitting.
- For instance, **moving from linear regression** (which assumes a linear relationship between the features and the target variable) to a **decision tree or a neural network**. These more complex algorithms do not assume linearity and can capture much more complex patterns in the data.

Underfitting



Note

Variance:

- Variance refers to the error from sensitivity to **small fluctuations in the training set**. **High variance** can cause an algorithm to **model the random noise** in the training data, rather than the intended outputs (**overfitting**). This happens when the model is too complex and learns patterns from the training data that don't generalize to unseen data

Feature	Underfitting	Overfitting
Complexity	Too simple	Too complex
Bias/Variance	High bias, Low variance	Low bias, High variance
Model Assumptions	Too strong, oversimplifies the problem	Too weak, captures noise as patterns
Performance on Training Data	Poor, cannot capture underlying patterns	Excellent, captures patterns and noise
Performance on Unseen Data	Poor, due to oversimplification	Poor, due to noise fitting
Generalization	Poor, due to lack of learning capacity	Poor, due to learning irrelevant patterns
Typical Causes	- Insufficient model complexity - Not enough features	- Excessive model complexity - Too many features without enough data
Solution Approaches	- Increase model complexity - Add more features - Use a more complex algorithm	- Simplify the model - Use regularization techniques - Collect more data - Reduce the number of features

Bias-variance tradeoff

- The **bias-variance tradeoff** is a fundamental principle in machine learning that describes the tension between the error introduced by the model's assumptions about the underlying data structure (bias) and the error from sensitivity to small fluctuations in the training set (variance). Understanding this tradeoff is crucial for creating models that generalize well to new, unseen data.
- The bias-variance tradeoff highlights the problem faced **when trying to minimize both bias and variance**. Ideally, we want a model that accurately captures the underlying patterns in the data (low bias) without capturing noise or fluctuations (low variance).

Bias-variance tradeoff

However, in practice, reducing bias often increases variance, and reducing variance increases bias. This is because:

- Decreasing bias usually involves making the model more complex (e.g., adding more parameters or features), which can lead to a model that starts to capture noise in the training data, increasing variance.
- Conversely, to reduce variance, one might simplify the model, which can increase bias by making the model too simplistic to capture the underlying structure of the data.

Bias-variance tradeoff

- The relationship between bias and variance is inherently a balancing act. A model must be complex enough to capture the true relationships in the data (low bias) but not so complex that it starts to model the random noise (low variance). This balance is crucial for achieving good generalization on unseen data. The goal in model selection and training is to find the sweet spot where both bias and variance are minimized to achieve the lowest possible total error.