Introduction to Data Science

Dr. Irfan Yousuf
Department of Computer Science (New Campus)
UET, Lahore

(Week 15; April 29 – May 03, 2024)

Outline

- Similarity Measures
- Model Building in Machine Learning
- Overfitting vs. Underfitting

Similarity Measure

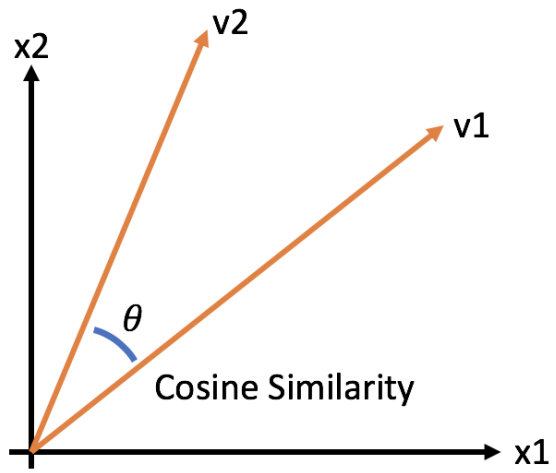
- In statistics, a similarity measure or similarity function is a real-valued function that **quantifies the similarity** between two objects.

- Usually, such measures are in some sense **the inverse of distance metrics**: they take on **large values for similar objects** and either zero or a **negative value for very dissimilar** objects.

- A similarity measure is a data mining or machine learning context **is a distance with dimensions** representing features of the objects.

Cosine Similarity

- Cosine similarity measures the **similarity between two vectors** of an inner product space.
- It is measured by the cosine of the angle between two vectors and determines whether two vectors are **pointing in roughly the same direction**.
- It is often used to measure document similarity in text analysis



Cosine Similarity

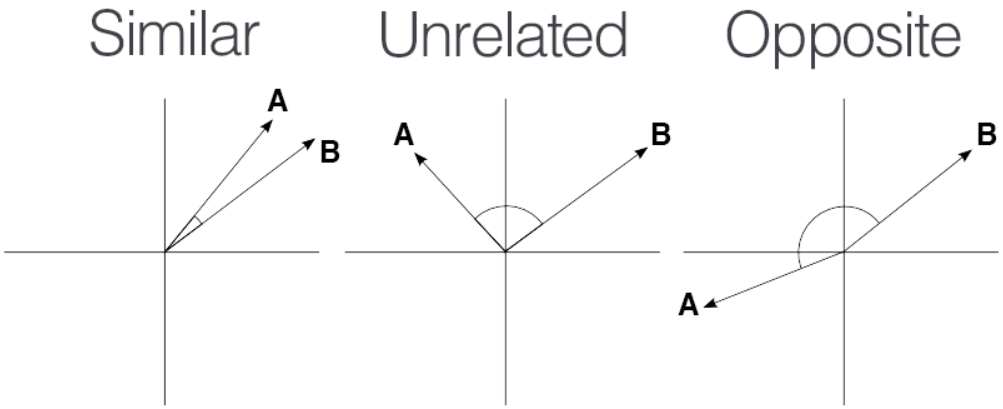
$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}},$$

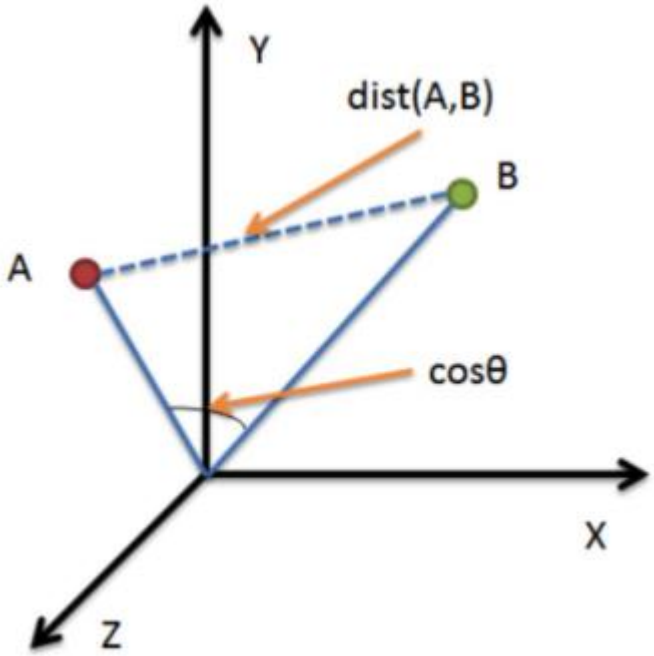
$$similarity(A, B) = cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|}$$

where

- θ is the angle between the vectors,
- $A \cdot B$ is dot product between A and B and calculated as $A \cdot B = A^T B = \sum_{i=1}^{n} A_i B_i = A_1 B_1 + A_2 B_2 + \ldots + A_n B_n$,
- $\|A\|$ represents the L2 norm or magnitude of the vector which is calculated as $\|A\| = \sqrt{A_1^2 + A_1^2 \ldots A_1^n}$.

Cosine Similarity





Cosine similarity measures the cosine of the angle between two multi-dimensional vectors. The smaller the angle, the higher the cosine similarity. Unlike measuring Euclidean distance, **cosine similarity captures the orientation of the documents and not the magnitude**. Cosine similarity pays more attention to the difference in the direction of two vectors than the distance or length.

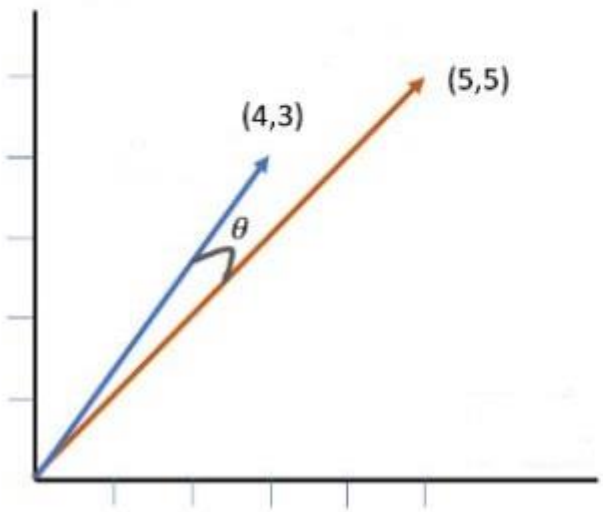
Cosine Similarity

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}},$$

$$similarity = \cos\theta = \frac{b.c}{\|b\|\|c\|}$$

$b.c \Rightarrow$ *Is the Dot product of the two vectors*

$\|b\|\|c\| \Rightarrow$ *Is the product of each vector's magnitude*



Calculating:

$b.c = \sum_{i=1}^{n} b_i c_i = (4 \times 5) + (3 \times 5) = 35$

$\|b\| = \sqrt{4^2 + 3^2} = 5$

$\|c\| = \sqrt{5^2 + 5^2} = 5\sqrt{2}$

$similarity = \frac{35}{5 \times 5\sqrt{2}} \sim 0.989$

Cosine Similarity for Text

Document 1: Deep Learning can be hard

Document 2: Deep Learning can be simple

Step 1: First we obtain a vectorised representation of the texts

Vectorised Representation

Word	Document 1	Document 2
Deep	1	1
Learning	1	1
Can	1	1
Be	1	1
Hard	1	0
Simple	0	1

Cosine Similarity for Text

Document 1: [1, 1, 1, 1, 1, 0] let's refer to this as A

Document 2: [1, 1, 1, 1, 0, 1] let's refer to this as B

Above we have two vectors (A and B) that are in a 6 dimension vector space

Step 2: Find the cosine similarity

cosine similarity (CS) = (A . B) / (||A|| ||B||)

- Calculate the dot product between A and B: 1.1 + 1.1 + 1.1 + 1.1 + 1.0 + 0.1 = **4**
- Calculate the magnitude of the vector A: $\sqrt{1^2 + 1^2 + 1^2 + 1^2 + 1^2 + 0^2}$ = **2.2360679775**
- Calculate the magnitude of the vector B: $\sqrt{1^2 + 1^2 + 1^2 + 1^2 + 0^2 + 1^2}$ = **2.2360679775**
- Calculate the cosine similarity: (4) / (2.2360679775*2.2360679775) = **0.80 (80% similarity between the sentences in both document)**

Cosine Similarity for Text

Doc1: Data is the oil of the digital economy
Doc2: Data is a new oil

	data	digital	economy	is	new	of	oil	the
doc_1	1	1	1	1	0	1	1	2
doc_2	1	0	0	1	1	0	1	0

$$A \cdot B = \sum_{i=1}^{n} A_i B_i$$

$$= (1 * 1) + (1 * 0) + (1 * 0) + (1 * 1) + (0 * 1) + (1 * 0) + (1 * 1) + (2 * 0)$$

$$= 3$$

$$\sqrt{\sum_{i=1}^{n} A_i^2} = \sqrt{1 + 1 + 1 + 1 + 0 + 1 + 1 + 4} = \sqrt{10}$$

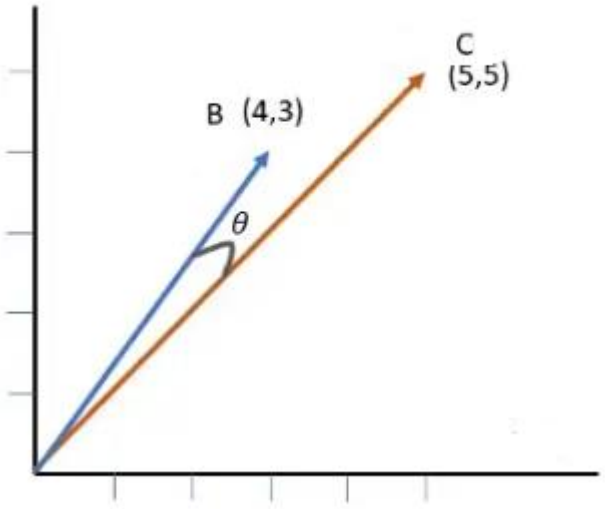
$$\sqrt{\sum_{i=1}^{n} B_i^2} = \sqrt{1 + 0 + 0 + 1 + 1 + 0 + 1 + 0} = \sqrt{4}$$

$$cosine\ similarity = cos\theta = \frac{A \cdot B}{|A||B|} = \frac{3}{\sqrt{10} * \sqrt{4}} = 0.4743$$

Cosine Similarity for Movie Ratings

User	Movie1	Movie2
B	4	3
C	5	5

$$\vec{b} = \begin{bmatrix} 4 \\ 3 \end{bmatrix} \qquad \vec{c} = \begin{bmatrix} 5 \\ 5 \end{bmatrix}$$



Calculating:

$$b.c = \sum_{i=1}^{n} b_i c_i = (4 \times 5) + (3 \times 5) = 35$$

$$\|b\| = \sqrt{4^2 + 3^2} = 5$$

$$\|c\| = \sqrt{5^2 + 5^2} = 5\sqrt{2}$$

$$similarity = \frac{35}{5 \times 5\sqrt{2}} \sim 0.989$$

Cosine Similarity for Movie Ratings

User	Movie1	Movie2	Movie3
B	4	3	5
C	5	5	1

$$\vec{b} = \begin{bmatrix} 4 \\ 3 \\ 5 \end{bmatrix} \qquad \vec{c} = \begin{bmatrix} 5 \\ 5 \\ 1 \end{bmatrix}$$

$$b.c = (4 \times 5) + (3 \times 5) + (5 \times 1) = 40$$

$$\|b\| = \sqrt{4^2 + 3^2 + 5^2} = 5\sqrt{2}$$

$$\|c\| = \sqrt{5^2 + 5^2 + 1^2} = \sqrt{51}$$

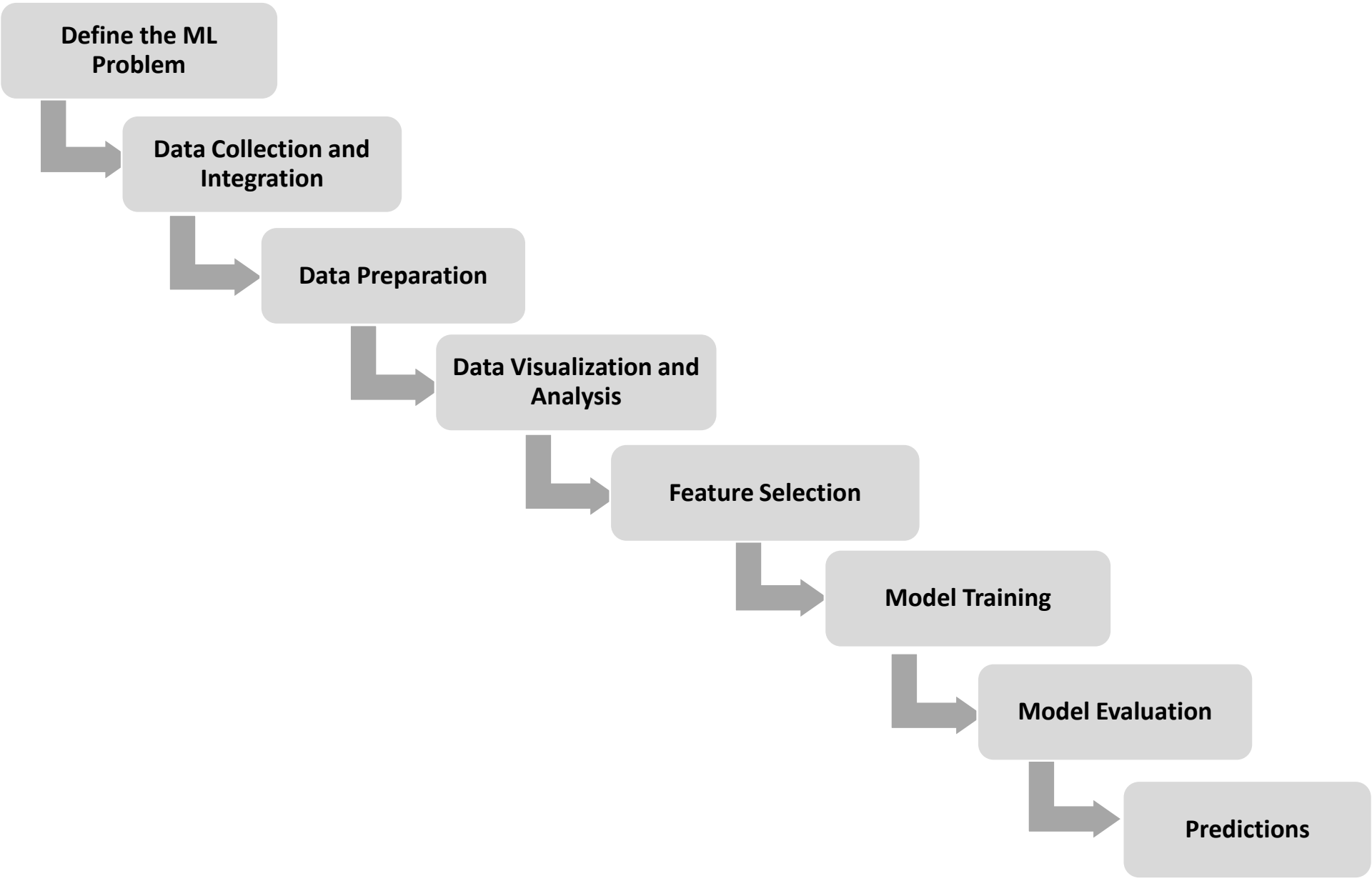
$$similarity = \frac{40}{5\sqrt{2} \times \sqrt{51}} \sim 0.792$$

Model Building in Machine Learning

Model Building in Machine Learning

- A machine learning model is built by learning and generalizing from training data, then applying that acquired knowledge to new data it has never seen before to make predictions and fulfill its purpose.

Model Building in Machine Learning



Step 1: Define the Problem

- We need to define our objectives and evaluate the problem that we are facing.
- Generally, our predictions are divided into three different main categories depending on the ML problem we need to answer;
  - Classification
  - Clustering
  - Regression

Step 2: Data Collection and Integration

- Data is everywhere so it can be collected from multiple sources like internet, databases or other types of storage.
- Chances are, that some of the data you collect going to be noisy - your data is possibly incomplete even irrelevant.
- So, wherever it comes from, it will need to be compiled - get integrated.
- The quality and quantity of data that you gather will directly determine how accurate your model can be.

Step 3: Data Preparation

- Data you collected is
  - Noisy
  - Incomplete
  - Irrelevant
- You must clean it.
- You also need to normalize the data.

Step 4: Data Visualization and Analysis

- Perform Exploratory Data Analysis (EDA)
- It's a technique that helps you understand the relationships within your dataset.
- This leads to better features, better models.
- When you can see the data in a chart or plotted out, you can help unveil previously unseen patterns.
- It reveals corrupt data or outliers that you don't want, properties that could be very significant in your analysis

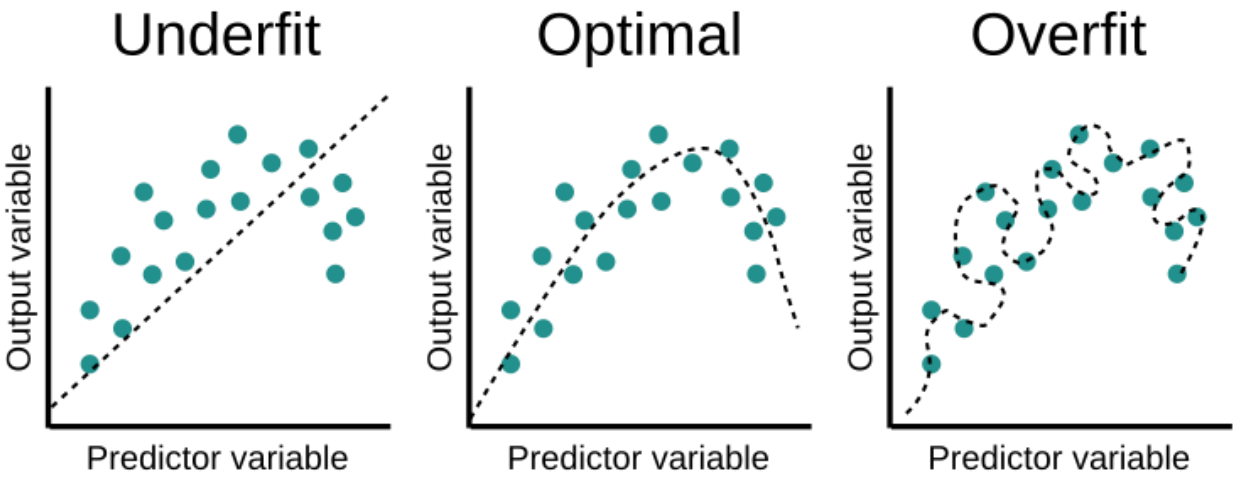
Step 5: Feature Selection

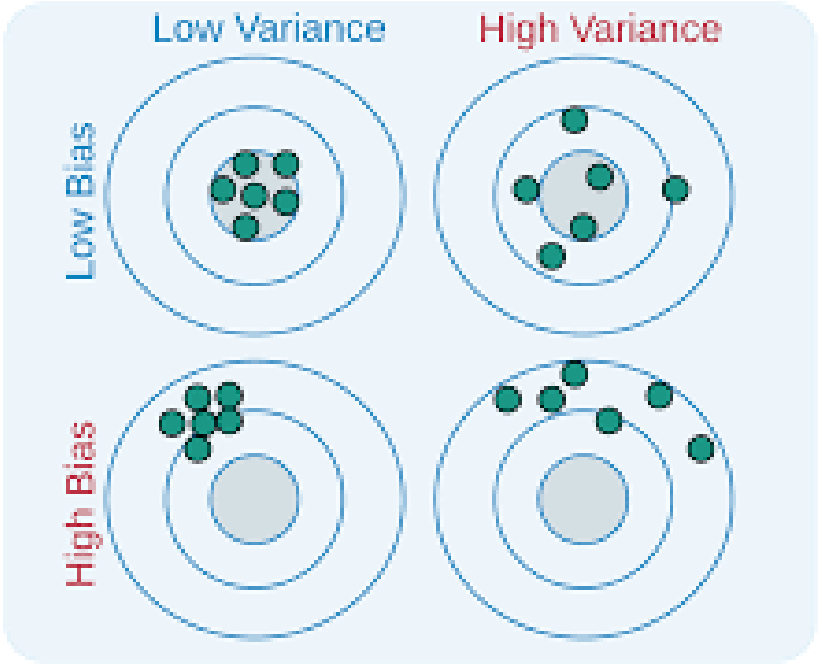
- A feature is a characteristic that might help when solving the problem.
- We will look for features that correlate to our desired output.
- A crucial part in this step is Feature Engineering – the process of manipulating the original data into new and potentially a lot of more useful features.

Step 6: Training

- The training often considered the main part of machine learning.
- **Overfitting:** the model learns the particulars of dataset too well.
- **Underfitting:** will not have enough features to model the data properly
- **Bias:** which is the gap between predicted value and actual value.
- **Variance:** how dispersed your predicted values are.

Step 6: Training





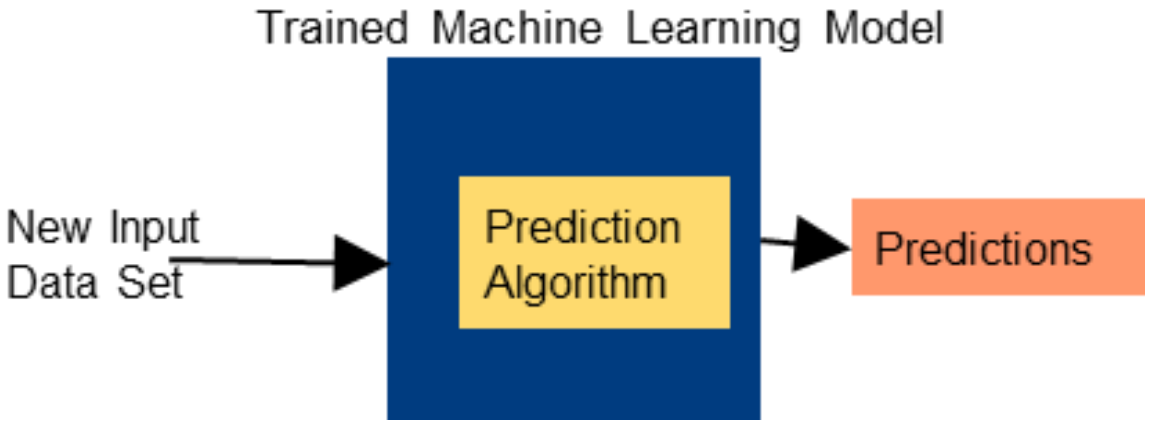
Step 7: Model Evaluation

- One of the most effective ways to evaluate your model's accuracy, precision, and ability to recall involves looking at something called a confusion matrix.

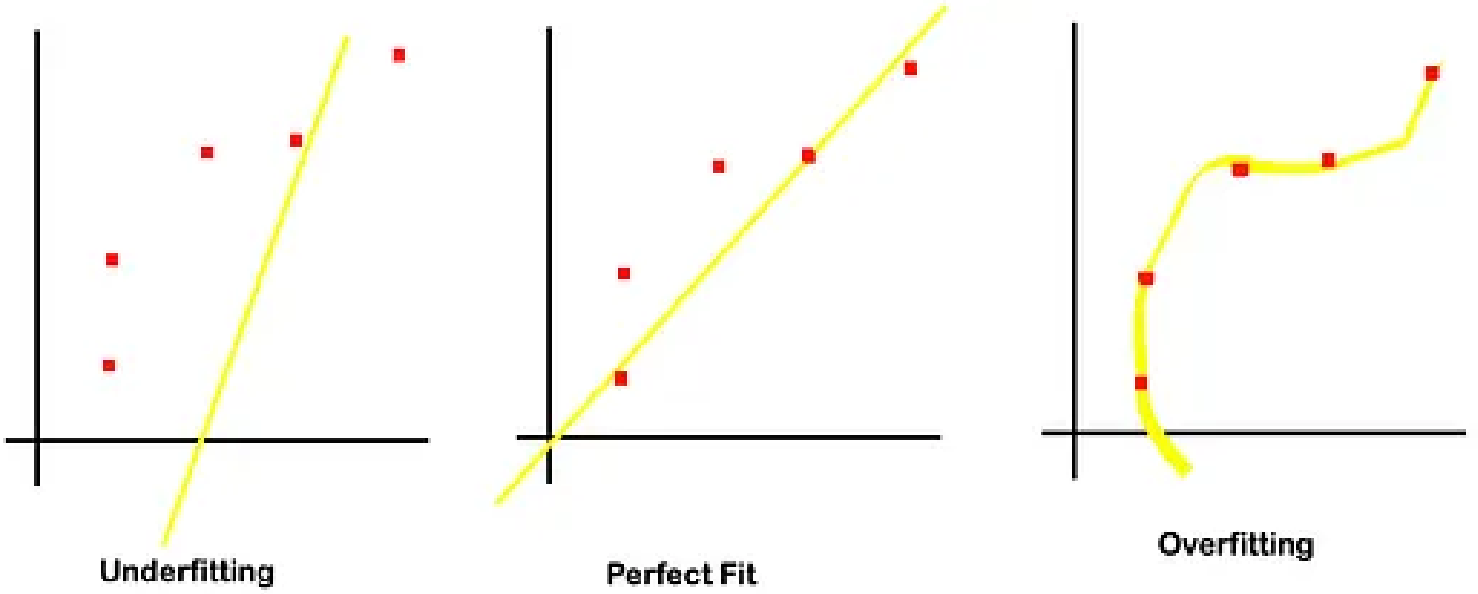
- The confusion matrix analyzes the model and shows how many of the data points were predicted correctly and incorrectly.

Step 8: Predictions

- Process unseen data and predict something.



Overfitting and Underfitting



Overfitting

- Overfitting refers to the scenario where a machine learning **model can’t generalize or fit well on unseen dataset**.
- A clear sign of machine learning overfitting is if its error on the testing or validation dataset is much greater than the error on training dataset.
- Overfitting is a term used in statistics that refers to a modeling error that occurs when a function corresponds too closely to a dataset.
- As a result, **overfitting may fail to fit additional data**, and this may affect the accuracy of predicting future observations.

Overfitting

- Overfitting happens when a model **learns the detail and noise in the training dataset** to the extent that it negatively impacts the performance of the model on a new dataset.

- This means that the **noise or random fluctuations in the training dataset is picked up and learned as concepts by the model.**

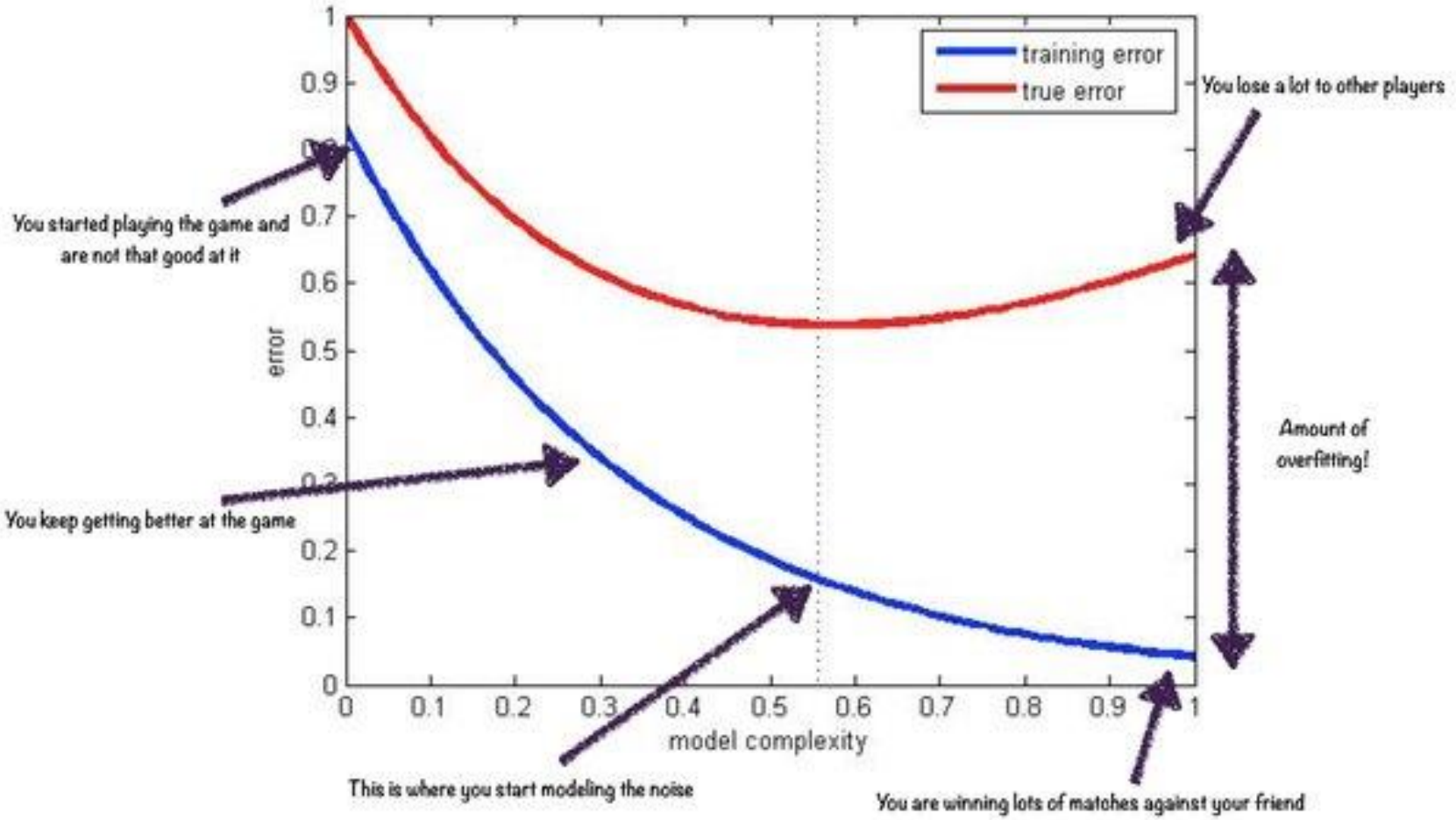
- The problem is that these concepts do not apply to new datasets and negatively impact the model's ability to generalize.

Overfitting

- Say you start playing FIFA with your friend on the Xbox. You lose first. But slowly you learn the tricks of the game and keep getting better at it. You get so good at the game that you start defeating your friend in every other match. You are pretty happy with yourself because you came a long way from knowing nothing.

- One a fine day, your friend is not around and you have to play with someone else. And you lose horribly. You begin to wonder what went wrong. When this phenomenon happens in machine learning, people call it overfitting.

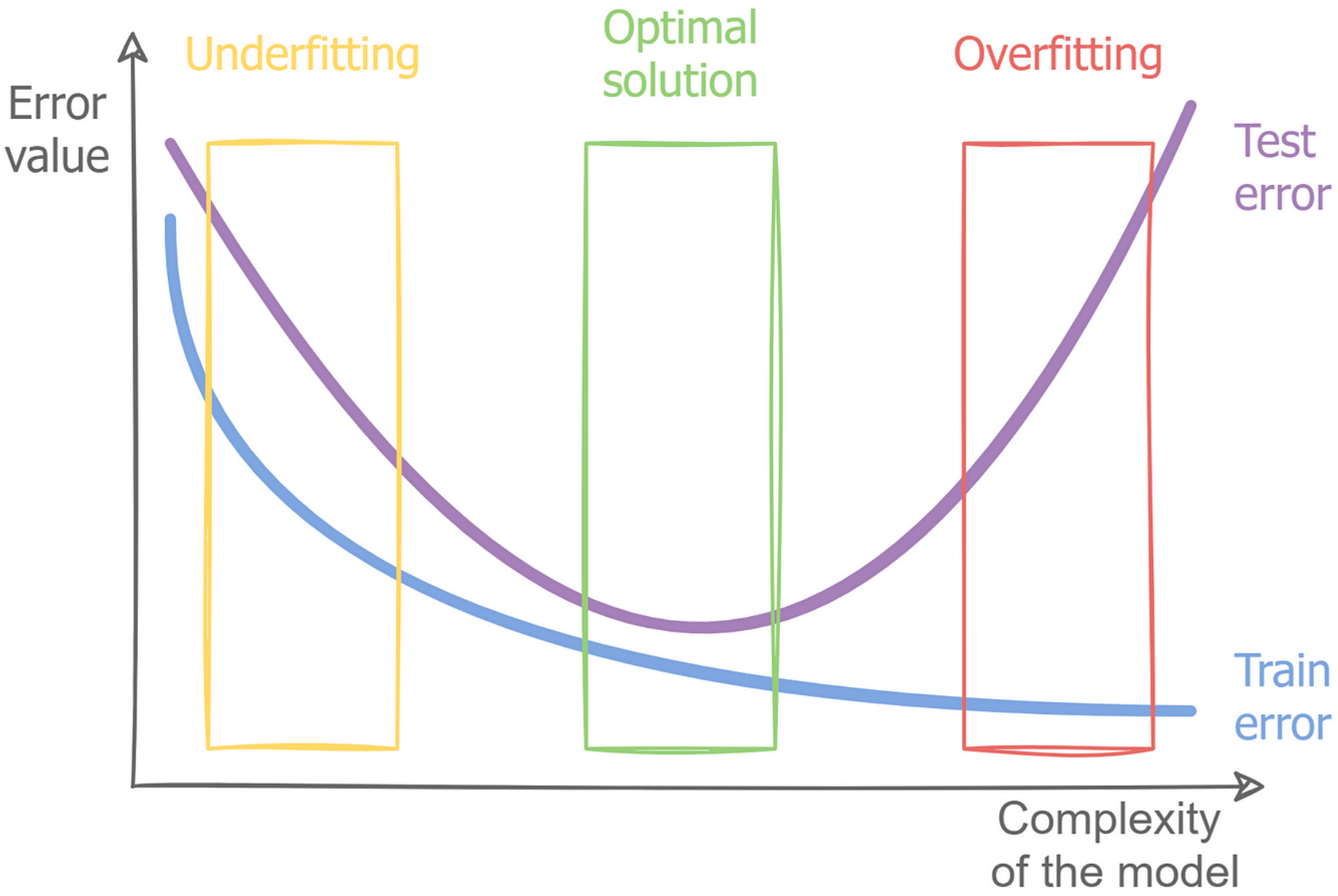
Overfitting



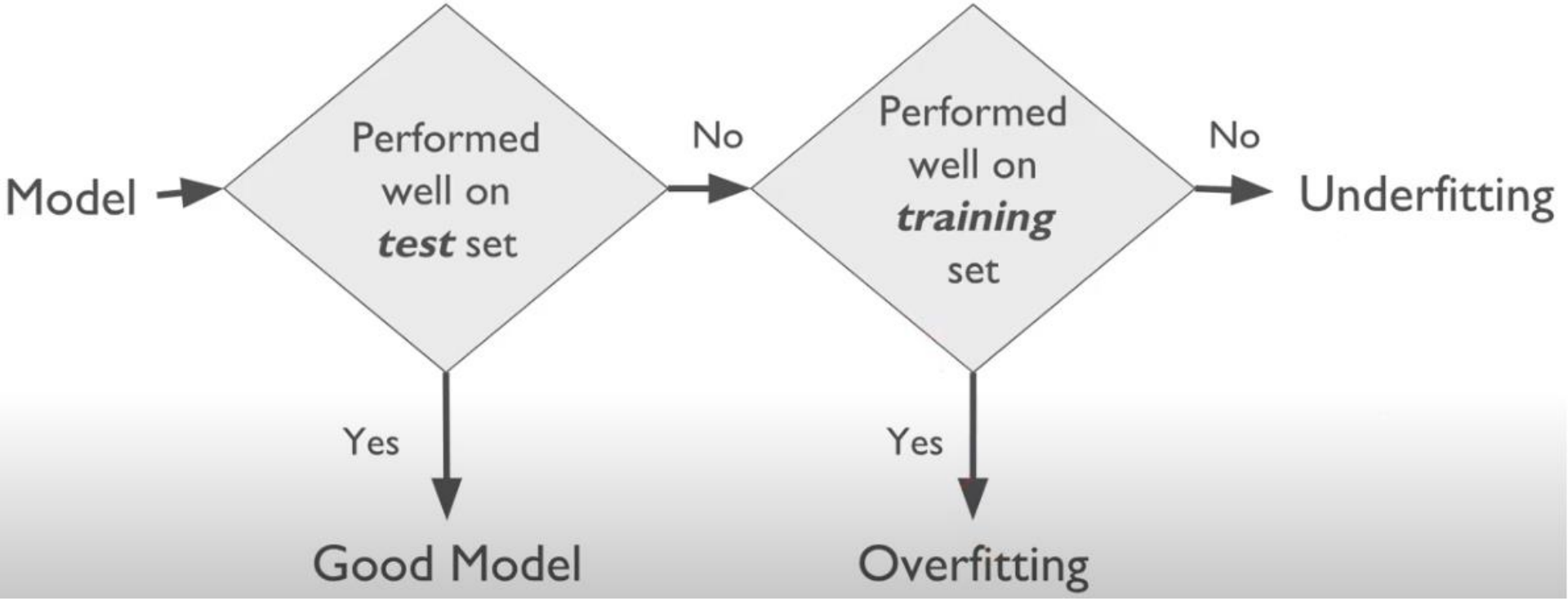
Underfitting

- Underfitting refers to a model that **can neither model the training dataset nor generalize to new dataset**.

- An underfit machine learning model is not a suitable model and will be obvious as **it will have poor performance on the training dataset**.


Underfitting
Optimal solution
Overfitting
Error value
Test error
Train error
Complexity of the model

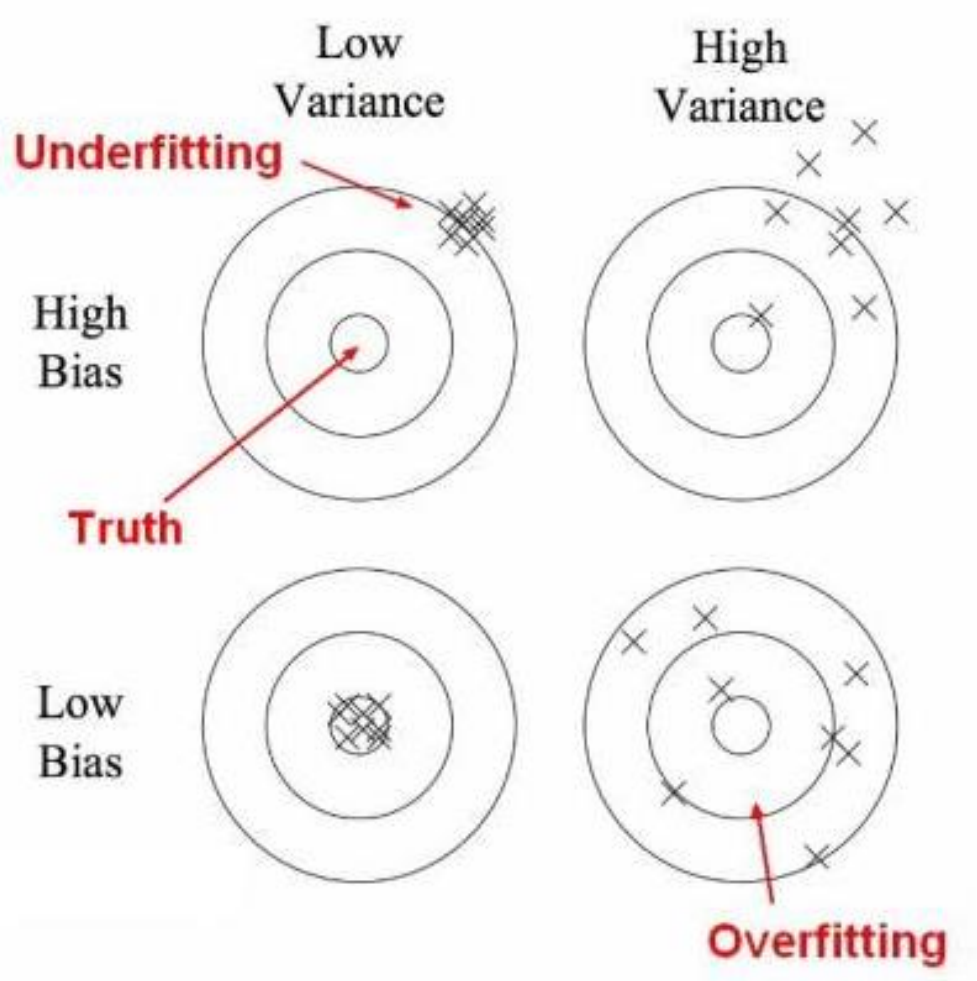
Overfitting and Underfitting



Reasons for Overfitting and Underfitting

- **Overfitting** happens when the size of training data used is not enough, or when our model captures the noise along with the underlying pattern in data.
- It happens when we train our model a lot over noisy dataset. These models have **low bias and high variance**.
- While **Underfitting** happens when a model unable to capture the underlying pattern of the data.
- These models usually have **high bias and low variance**. It happens when we have very less amount of data to build an accurate model or when we try to build a linear model with a nonlinear data.

Reasons for Overfitting and Underfitting



Reasons for Overfitting and Underfitting

- **Specificity:** Ability to perform focused learning on training data such that the model learns the minute details of the data.

- **Generalizability:** Ability to generalize model for unseen test data upon training by seen data.

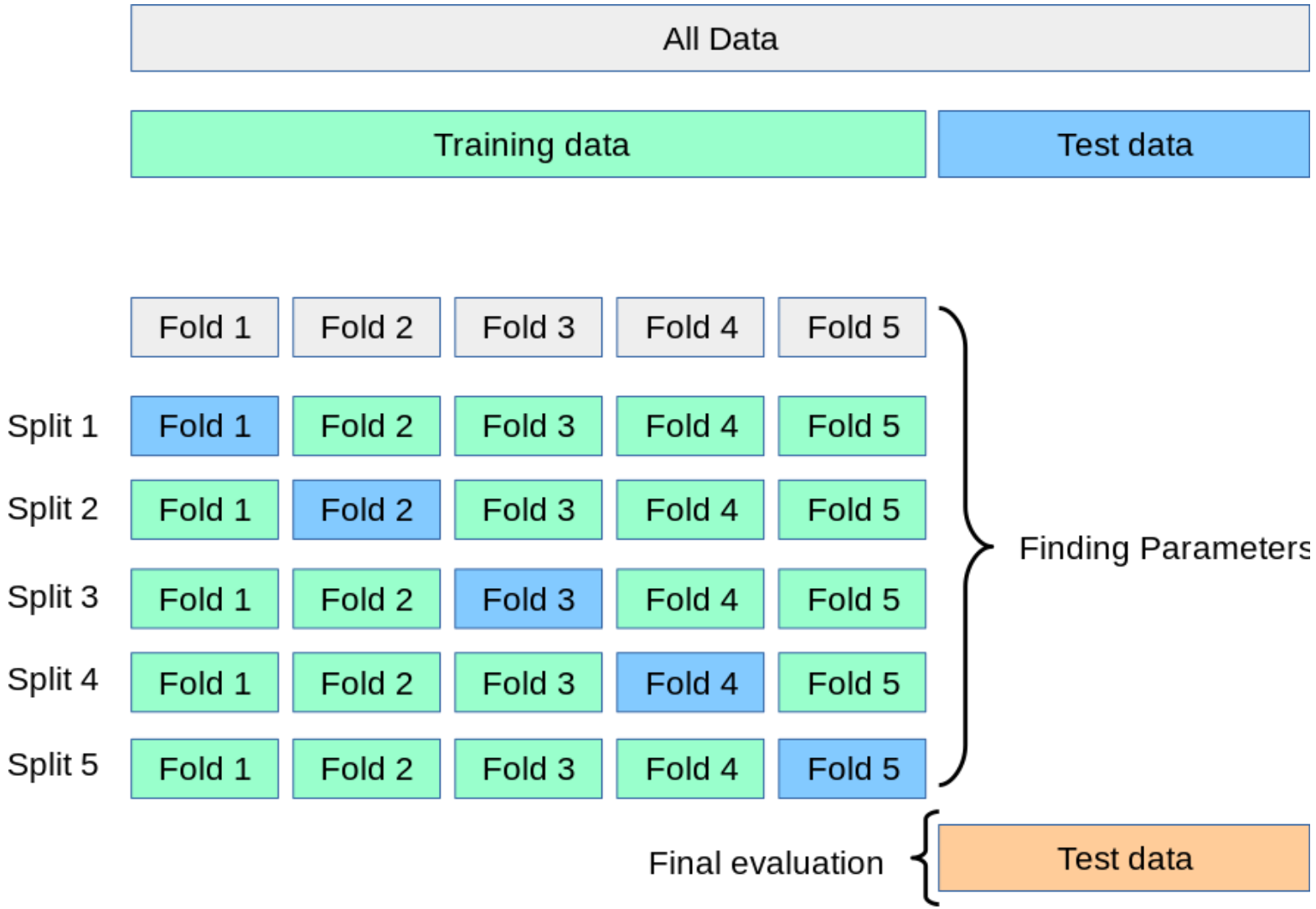
- **OverFitting:** High Specificity, Low Generalizability
- **UnderFitting:** Low Specificity, High Generalizability

How to Prevent Overfitting

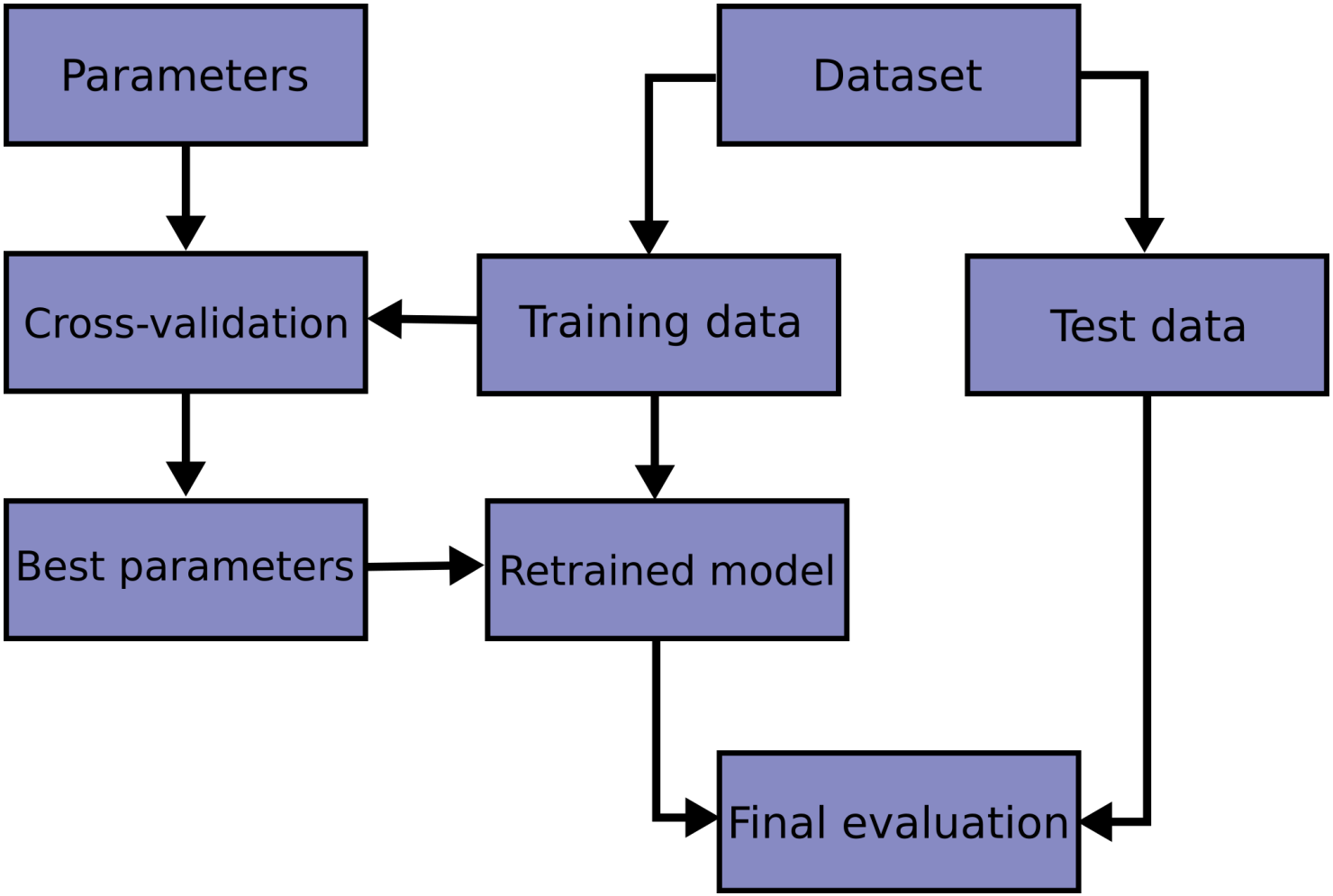
- **Cross-validation** is a powerful preventative measure against overfitting.

- Use your initial training data to generate **multiple mini train-test splits**. Use these splits to tune your model.
- In standard k-fold cross-validation, we partition the data into k subsets, called folds. Then, we iteratively train the algorithm on k-1 folds while using the remaining fold as the test set.

How to Prevent Overfitting



How to Prevent Overfitting



How to Prevent Overfitting

- **Train with more data:** It won't work every time, but training with more data can help algorithms detect the signal better.

- **Feature Selection:** We can improve generalizability of a model by removing irrelevant input features.

How to Prevent Underfitting

- **Increase model complexity:** As model complexity increases, performance on the data used to build the model (training data) improves.
- Remove noise from the data.
- Increase number of features.
- Increase the duration of training

Summary

- Model Building in Machine Learning