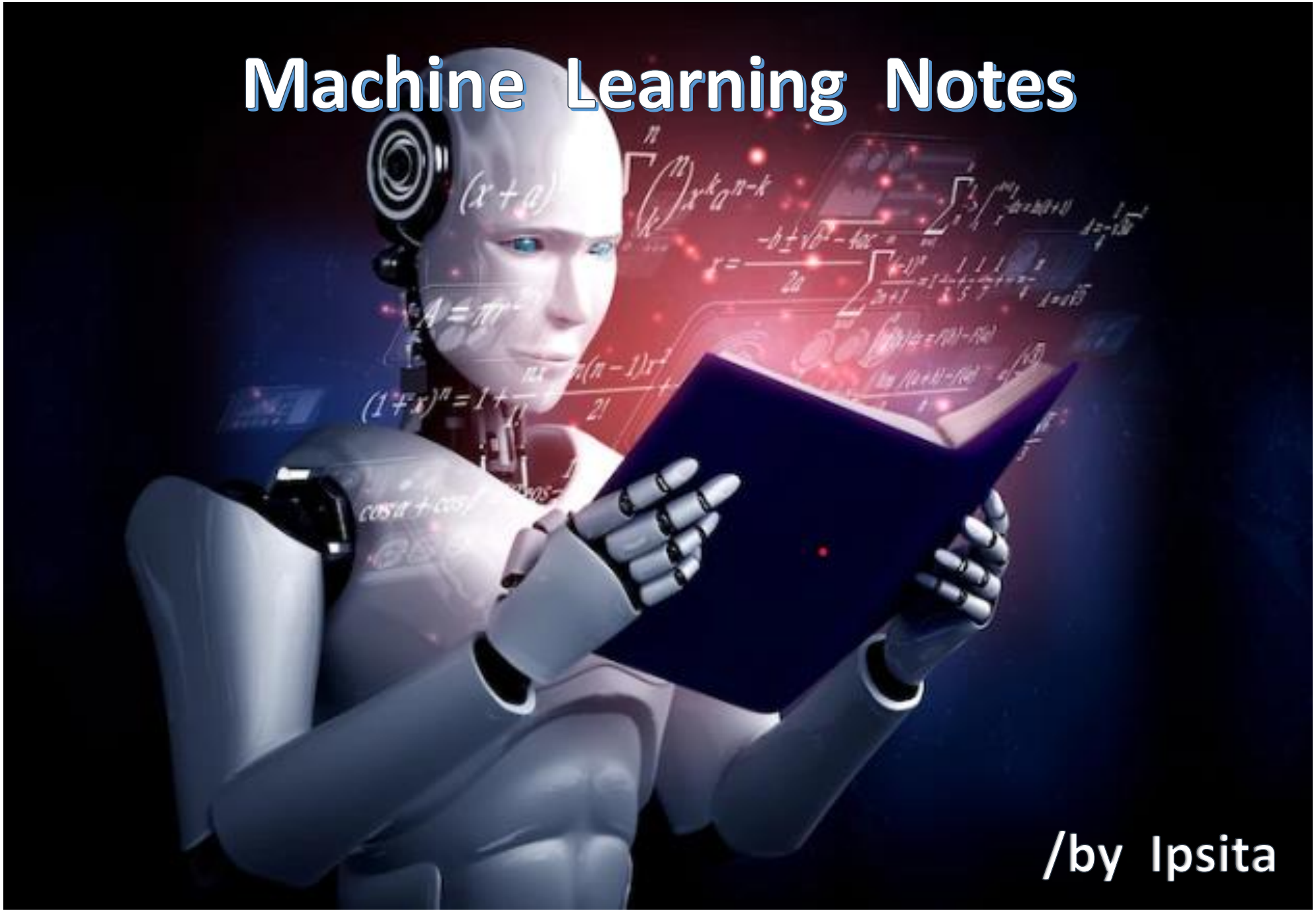
Machine Learning Notes

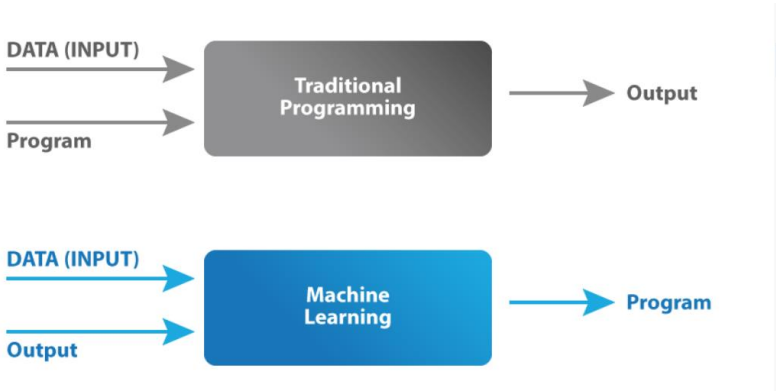


/by Ipsita

Basics of Machine Learning

Module 01

Machine Learning enables a machine to automatically learn from data itself, improve performance efficiently from past experience, and predict things without being explicitly programmed.

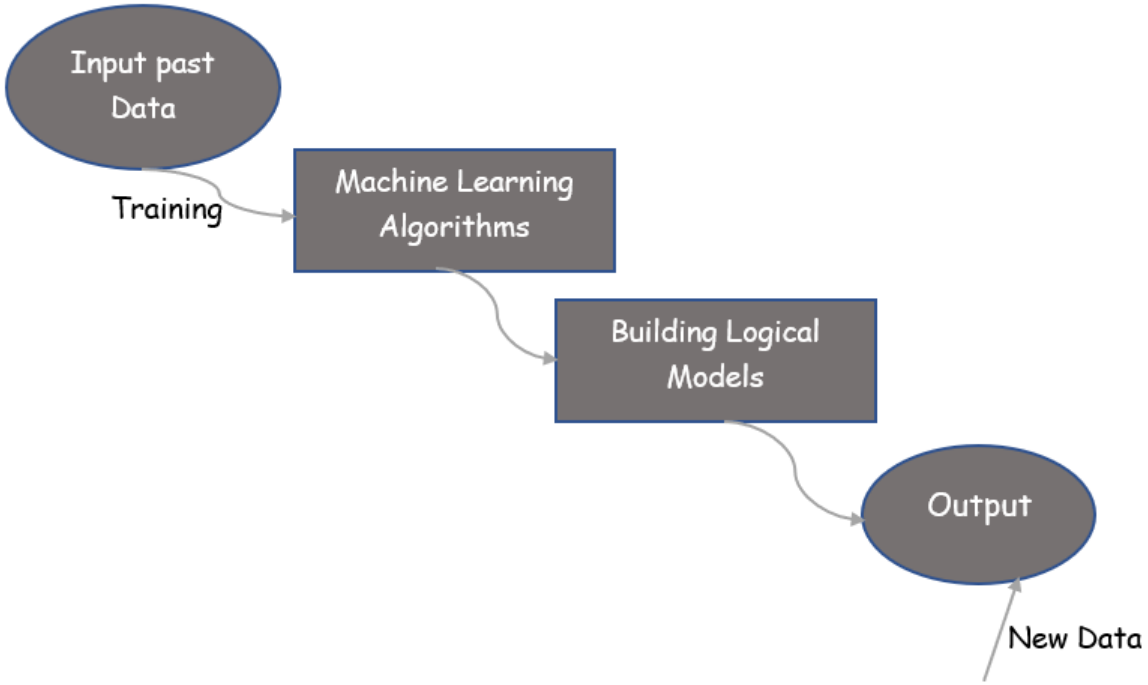


Types of Machine Learning

Mainly machine learning is of **three** types:

- Supervised Machine Learning
- Unsupervised Machine Learning
- Reinforcement Learning

Flow diagram of the learning process:



Importance of Machine Learning

- It can learn from past data & improve automatically.
- **Pattern Matching** & extraction of meaningful information from data.
- Solving **complex real-life problems** which are difficult for human.
- Decision Making in various fields.
- Rapid increment in the production of data.

Supervised Machine Learning

It is the type of machine learning technique in which machines are trained using well labelled training dataset, and on the basis of that data machines predict the output. The **labelled data** means some input data is already tagged with the correct output.

- Classification
- Regression

Unsupervised Machine Learning

It is the type of machine learning technique in which models are trained using unlabelled dataset and are allowed to act on that data without any suspension.

- Clustering
- Association

Reinforcement Machine Learning

It is a feedback based Machine learning technique in which an agent learns to behave in an environment by performing the actions and seeing the results of actions. For each **good action**, the agent gets a ***positive feedback*** and for each **bad action**, the agent receives a negative ***feedback or penalty***.

- Positive Reinforcement
- Negative Reinforcement

Types of Data

Categorical (Qualitative)

1. **Nominal:**
   The value of a nominal attribute are just different names i.e., nominal values provide only enough information to distinguish one object from another (= , ≠)
   Example: zip codes, employee Id number, eye color, gender
2. **Ordinal:**
   The value of an ordinal attribute provide just enough information to order objects (< , >)
   Example: hardness of minerals, {good, better, best} grades, street numbers

Numeric (Quantitative)

1. **Interval:**
   For interval attributes, the differences between values are meaningful, i.e., a unit of measurement exists. (+, -)
   Example: Calendar dates, temperature in Celsius or Fahrenheit
2. **Ratio:**
   For ratio variables, both differences and ratios are meaningful (*, /)
   Example: age, mass length, electrical current

Basic Terms

Overfitting & Underfitting:

If our algorithm works well with the training dataset but not well with the test dataset, then such problem is termed as **Overfitting**. And if our algorithm does not perform well even with training dataset, then such problem is **Underfitting**.

Outliers:

Outliers is an observation which contains either very low value or very high value in comparison to other observed values. An outlier may hamper the result so, it should be avoided.

Example: For an event the criteria mentioned is, the people coming for the event must be of age group 20 to 30. The volunteer noticed a group of people of age above 35 buying tickets. So, these people of age above 35 which does not fulfil the age limit mentioned are outliers in this case.

Dependent Variable:

The main factor in Regression analysis which we want to predict or understand is called the **dependent variable**. Also termed as the **target variable.**

Independent Variable:

The factors which affect the dependent variables or which are used to predict the values of the dependent variable are called **independent variables**, also termed as a **predictor**.

Multi-collinearity:

If the independent variables are highly correlated with each other than other variables, then such condition is called Multicollinearity. It should not be present in the dataset, because it creates problem while ranking the most efficient value.

Bias Variance Trade-Off:

Bias: It is a prediction error that is introduced in the model due to oversimplifying the machine learning algorithms. It can be easily defined as the **difference** between the **predicted values** and the **actual values**.

Variance: When a model does not perform as well as it does with the trained dataset, there is a possibility that the model has a variance. " It basically tells how scattered the predicted values are from the actual values."

– a high variance in a dataset means that the model has been trained with a lot of noise and irrelevant data. Thus, causing overfitting in the model.

Signal: It refers to the true underlying pattern of the data that helps the machine learning model to learn from the data.

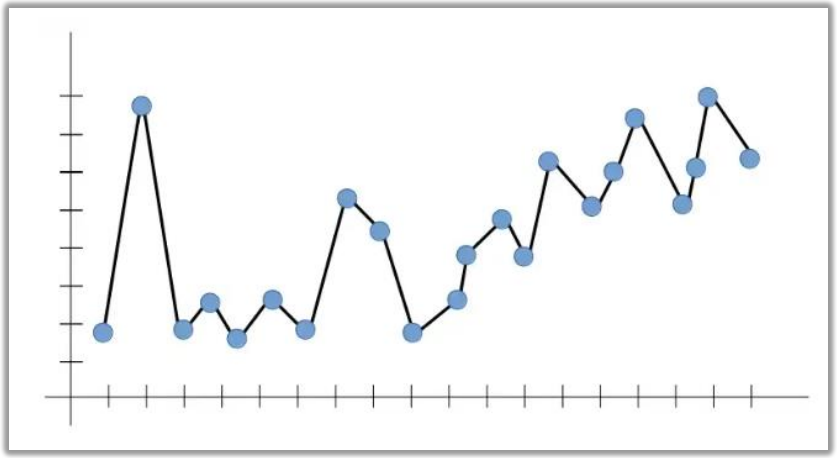
Noise: It is unnecessary and irrelevant data that reduces the performance of the model.

Overfitting *(explained in detail)*

It occurs when our machine learning model tries to cover all the data points or more than the desired data points present in the given dataset.

Because of this, the model starts catching noise and inaccurate values present in the dataset, and all these factors **reduce the efficiency and accuracy** of the model. *

* The overfitted model has low bias and high variance. *



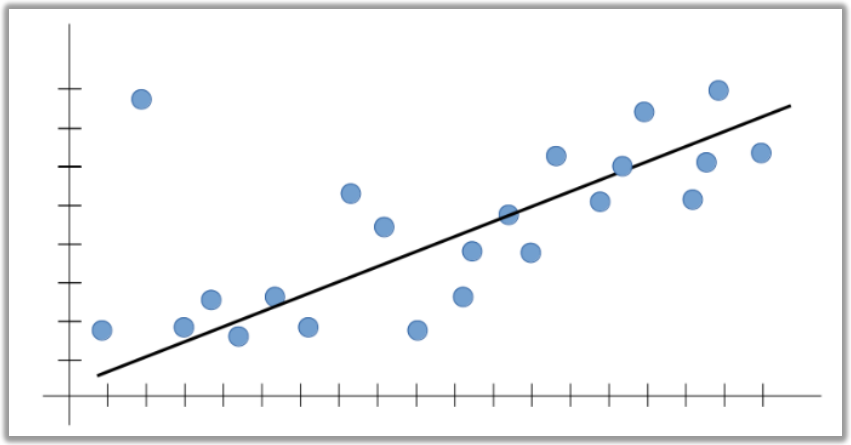
Ways to avoid overfitting in a model:

- Cross-Validation
- Training with more data
- Removing featured
- Regularization
- Early stoppage of training
- Ensembling

Underfitting *(explained in detail)*

In case of underfitting, the model is not able to learn enough from the training data, and hence it reduces the accuracy and produces unreliable prediction.

* An underfitted model has high bias and low variance.

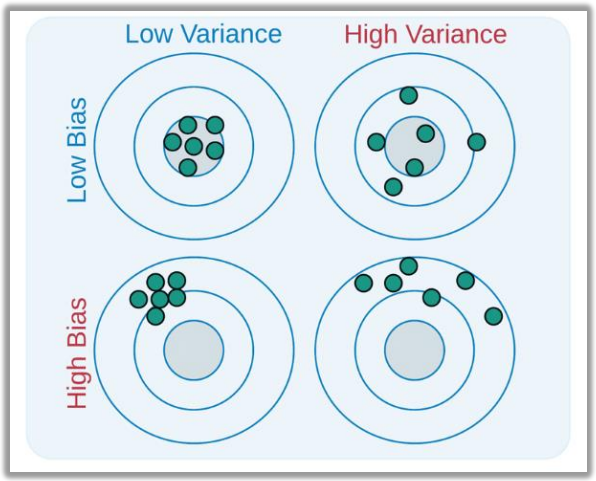


Ways to avoid overfitting in a model:

- By increasing the training time of the model.
- By increasing the number of features.

Bias-Variance Trade-Off:

Finding the right balance between the bias and variance of the model is called the Bias-Variance Trade off. It is basically a way to make sure that the model is neither overfitted or underfitted in any case.



"If the model is too simple and has few parameters, it will suffer from high bias and low variance. On the other hand, if the model has a large number of parameters, it will have low bias & high variance. This trade-off should result in a perfectly balanced relationship between the two. Ideally, low bias & low variance is the target for any machine learning model."

Bias-Variance Trade-off Graph