Module 1: Introduction to Machine Learning

By
Ms. Sarika Dharangaonkar
Assistant Professor,
KJSCE

Contents: Module1-Introduction to Machine Learning

1.1 Introduction, Types of Machine Learning, Process of Machine learning

1.2 Introduction to terminologies – Weight space, Curse of Dimensionality

1.3 Testing Machine Learning Algorithms

1.4 Minimizing Risk and The Naïve Bayes’ Classifier

1.5 Bias-Variance Trade Off

Machine Learning

- Arthur Samuel, an early American leader in the field of computer gaming and artificial intelligence, coined the term “Machine Learning ” in 1959 while at IBM.
- He defined machine learning as “the field of study that gives computers the ability to learn without being explicitly programmed “.
- The field of study known as machine learning is concerned with the question of how to construct computer programs that automatically improve with experience.
- Machine learning is a subfield of artificial intelligence that involves the development of algorithms and statistical models that enable computers to improve their performance in tasks through experience.

Learning

- “A computer program is said to learn from experience E with respect to some task **T** and some performance measure **P**, if its performance on **T**, as measured by **P**, improves with experience **E**.”
- **Examples**
- Handwriting recognition learning problem
  - Task T : Recognizing and classifying handwritten words within images
  - Performance P : Percent of words correctly classified
  - Training experience E : A dataset of handwritten words with given classifications
- A robot driving learning problem
  - Task T : Driving on highways using vision sensors
  - Performance P : Average distance traveled before an error
  - Training experience E : A sequence of images and steering commands recorded while observing a human driver

Classification of Machine Learning

- **Supervised learning**
- **Unsupervised learning**
- **Reinforcement learning**

Supervised learning

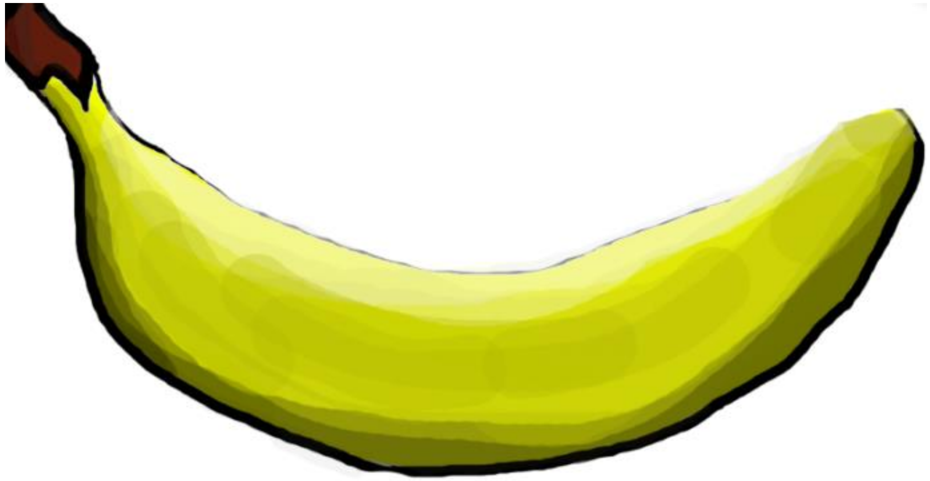
- Supervised learning is the machine learning task of learning a function that maps an input to an output based on example input-output pairs.
- The given data is labeled.
- Basically supervised learning is when we teach or train the machine using data that is well-labelled. Which means some data is already tagged with the correct answer.
- Both *classification* and *regression* problems are supervised learning problems.

Example: Supervised Learning

- suppose you are given a basket filled with different kinds of fruits. Now the first step is to train the machine with all the different fruits one by one like this:

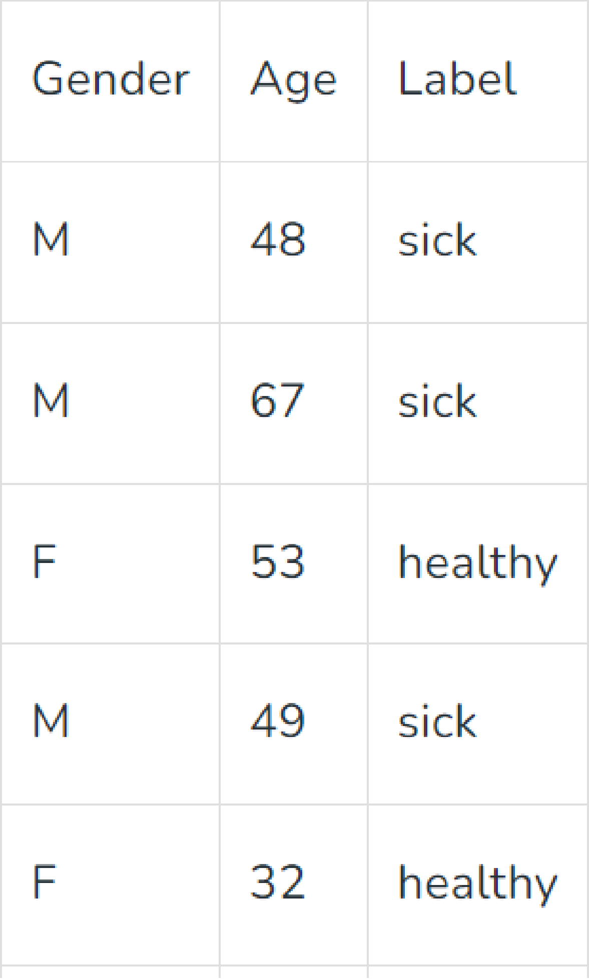
- If the shape of the object is rounded and has a depression at the top, is red in color, then it will be labeled as –**Apple**.
- If the shape of the object is a long curving cylinder having Green-Yellow color, then it will be labeled as –**Banana**.

- Now suppose after training the data, you have given a new separate fruit, say Banana from the basket, and asked to identify it.
- Since the machine has already learned the things from previous data and this time has to use it wisely.



- It will first classify the fruit with its shape and color and would confirm the fruit name as BANANA and put it in the Banana category.
- Thus the machine learns the things from training data(basket containing fruits) and then applies the knowledge to test data(new fruit).

- Example 2 — Consider the following data regarding patients entering a clinic . The data consists of the gender and age of the patients and each patient is labeled as “healthy” or “sick”.



Gender	Age	Label
M	48	sick
M	67	sick
F	53	healthy
M	49	sick
F	32	healthy

Supervised Learning contd..

Supervised learning is classified into two categories of algorithms:

- **Classification**: A classification problem is when the output variable is a category, such as “Red” or “blue” , “disease” or “no disease”.
- **Regression**: A regression problem is when the output variable is a real value, such as “dollars” or “weight”.

Types:-

- Regression
- Logistic Regression
- Classification
- Naive Bayes Classifiers
- K-NN (k nearest neighbors)
- Decision Trees
- Support Vector Machine

Supervised Learning Contd…

Advantages:-

- Supervised learning allows collecting data and produces data output from previous experiences.
- Helps to optimize performance criteria with the help of experience.
- Supervised machine learning helps to solve various types of real-world computation problems.
- It performs classification and regression tasks.
- It allows estimating or mapping the result to a new sample.
- We have complete control over choosing the number of classes we want in the training data.

Supervised Learning Contd…

Disadvantages:-

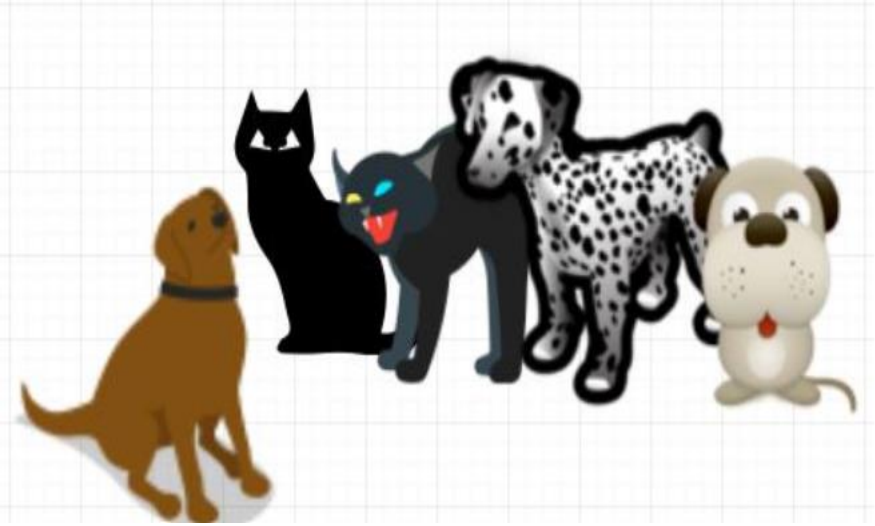
- Classifying big data can be challenging.
- Training for supervised learning needs a lot of computation time. So, it requires a lot of time.
- Supervised learning cannot handle all complex tasks in Machine Learning.
- Computation time is vast for supervised learning.
- It requires a labelled data set.
- It requires a training process.

Unsupervised Learning

- Unsupervised learning is the training of a machine using information that is neither classified nor labeled and allowing the algorithm to act on that information without guidance.
- Here the task of the machine is to group unsorted information according to similarities, patterns, and differences without any prior training of data.
- Unlike supervised learning, no teacher is provided that means no training will be given to the machine. Therefore the machine is restricted to find the hidden structure in unlabeled data by itself.

Unsupervised Learning Contd…

- Example:
- Suppose it is given an image having both dogs and cats which it has never seen.



- Thus the machine has no idea about the features of dogs and cats so we can’t categorize it as ‘dogs and cats ‘.
- But it can categorize them according to their similarities, patterns, and differences, i.e., we can easily categorize the above picture into two parts.
- The first may contain all pics having **dogs** in them and the second part may contain all pics having **cats** in them.
- Here you didn’t learn anything before, which means no training data or examples.

Unsupervised Learning Contd…

- Unsupervised learning is classified into two categories of algorithms:
- **Clustering**: A clustering problem is where you want to discover the inherent groupings in the data, such as grouping customers by purchasing behavior.
- **Association**: An association rule learning problem is where you want to discover rules that describe large portions of your data, such as people that buy X also tend to buy Y.

Unsupervised Learning Contd…

Types of Unsupervised Learning:-

1. Clustering

- Exclusive (partitioning)
- Agglomerative
- Overlapping
- Probabilistic

Clustering Types:-

- Hierarchical clustering
- K-means clustering
- Principal Component Analysis
- Singular Value Decomposition
- Independent Component Analysis

Unsupervised Learning Contd…

Advantages:

- It does not require training data to be labeled.
- Dimensionality reduction can be easily accomplished using unsupervised learning.
- Capable of finding previously unknown patterns in data.
- **Flexibility**: Unsupervised learning is flexible in that it can be applied to a wide variety of problems, including clustering, anomaly detection, and association rule mining.
- **Exploration**: Unsupervised learning allows for the exploration of data and the discovery of novel and potentially useful patterns that may not be apparent from the outset.
- **Low cost**: Unsupervised learning is often less expensive than supervised learning because it doesn't require labeled data, which can be time-consuming and costly to obtain.

Unsupervised Learning Contd…

Disadvantages:

- Difficult to measure accuracy or effectiveness due to lack of predefined answers during training.
- The results often have lesser accuracy.
- The user needs to spend time interpreting and label the classes which follow that classification.
- **Lack of guidance**: Unsupervised learning lacks the guidance and feedback provided by labeled data, which can make it difficult to know whether the discovered patterns are relevant or useful.
- **Sensitivity to data quality**: Unsupervised learning can be sensitive to data quality, including missing values, outliers, and noisy data.
- **Scalability**: Unsupervised learning can be computationally expensive, particularly for large datasets or complex algorithms, which can limit its scalability.

Reinforcement Learning

- Reinforcement Learning (RL) is the science of decision making.
- It is about learning the optimal behavior in an environment to obtain maximum reward.
- In RL, the data is accumulated from machine learning systems that use a trial-and-error method. Data is not part of the input that we would find in supervised or unsupervised machine learning.
- Reinforcement learning uses algorithms that learn from outcomes and decide which action to take next.
- After each action, the algorithm receives feedback that helps it determine whether the choice it made was correct, neutral or incorrect.
- It is a good technique to use for automated systems that have to make a lot of small decisions without human guidance.

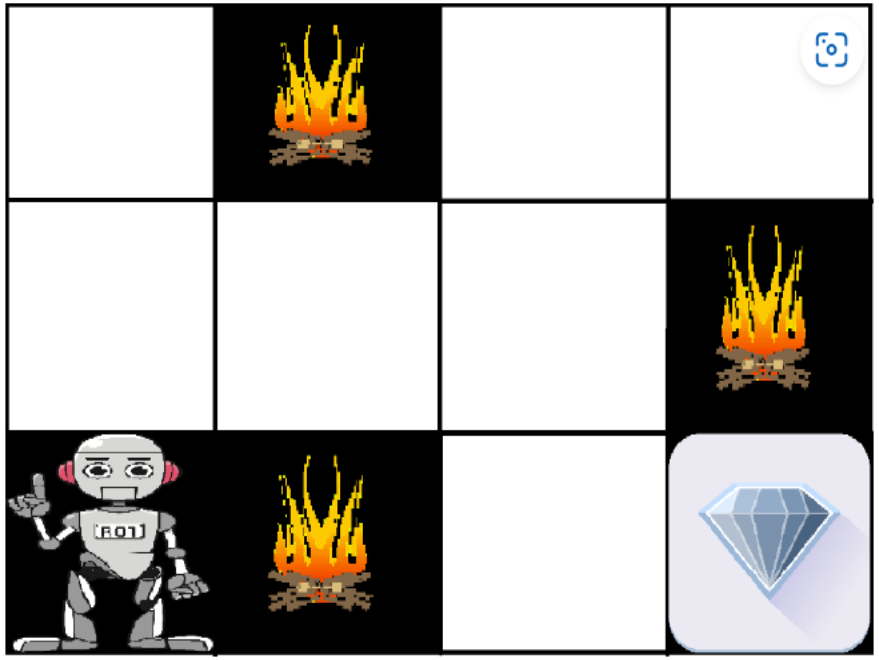
Reinforcement Learning Contd…

- Reinforcement learning is an autonomous, self-teaching system that essentially learns by trial and error.
- It performs actions with the aim of maximizing rewards, or in other words, it is learning by doing in order to achieve the best outcomes.

Reinforcement Learning Contd…

Example:

- The problem is as follows: We have an agent and a reward, with many hurdles in between. The agent is supposed to find the best possible path to reach the reward.



- The goal of the robot is to get the reward that is the diamond and avoid the hurdles that are fired.
- The robot learns by trying all the possible paths and then choosing the path which gives him the reward with the least hurdles.
- Each right step will give the robot a reward and each wrong step will subtract the reward of the robot.
- The total reward will be calculated when it reaches the final reward that is the diamond.

Reinforcement Learning Contd…

Main points in Reinforcement learning –

- Input: The input should be an initial state from which the model will start
- Output: There are many possible outputs as there are a variety of solutions to a particular problem
- Training: The training is based upon the input, The model will return a state and the user will decide to reward or punish the model based on its output.
- The model keeps continues to learn.
- The best solution is decided based on the maximum reward.

Reinforcement Learning Contd…

Types of Reinforcement:

- There are two types of Reinforcement:

1. **Positive:** Positive Reinforcement is defined as when an event, occurs due to a particular behavior, increases the strength and the frequency of the behavior. In other words, it has a positive effect on behavior. Advantages of reinforcement learning are:
   - Maximizes Performance
   - Sustain Change for a long period of time
   - Too much Reinforcement can lead to an overload of states which can diminish the results

Reinforcement Learning Contd…

2. Negative: Negative Reinforcement is defined as strengthening of behavior because a negative condition is stopped or avoided. Advantages of reinforcement learning:

- –Increases Behavior
- –Provide defiance to a minimum standard of performance
- –It Only provides enough to meet up the minimum behavior

Reinforcement Learning Contd…

Advantages of Reinforcement learning

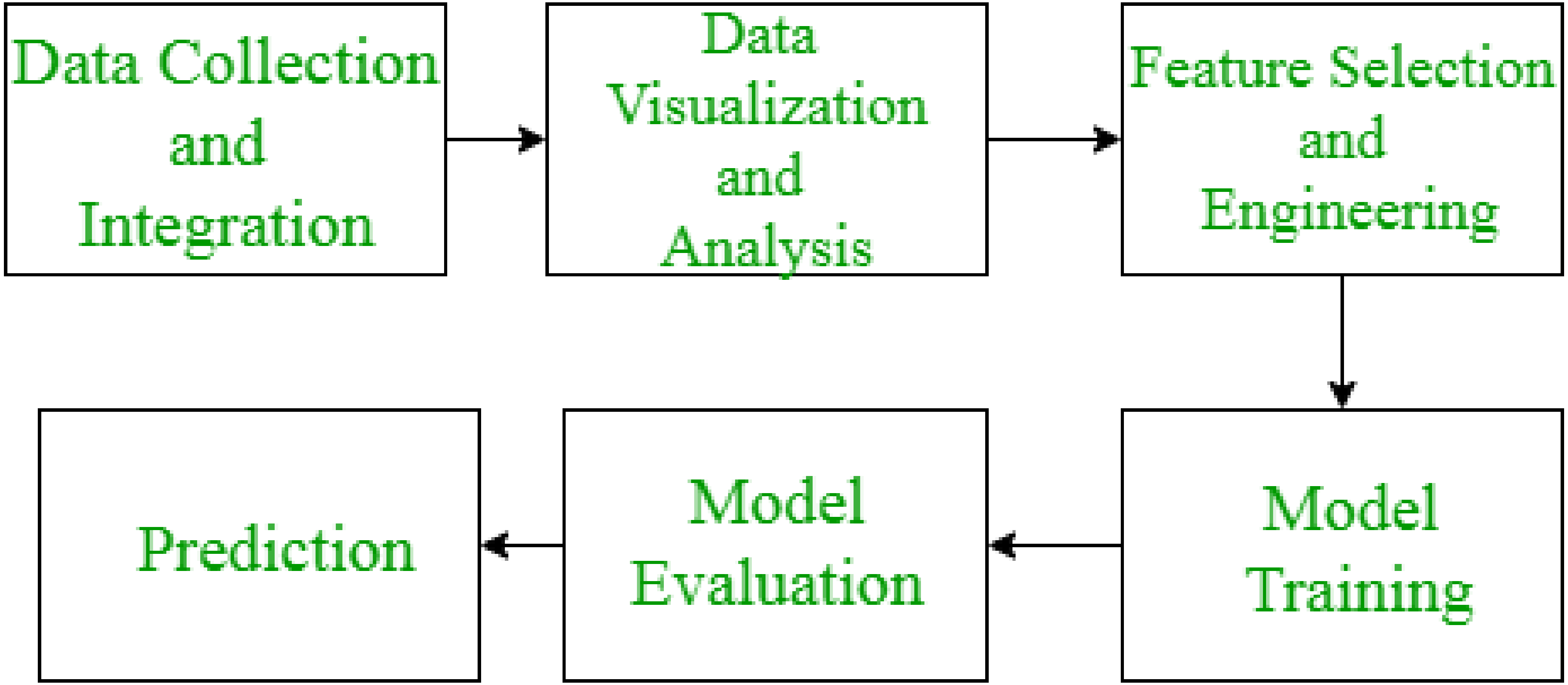
- Reinforcement learning can be used to solve very complex problems that cannot be solved by conventional techniques.
- The model can correct the errors that occurred during the training process.
- In RL, training data is obtained via the direct interaction of the agent with the environment
- Reinforcement learning can handle environments that are non-deterministic, meaning that the outcomes of actions are not always predictable. This is useful in real-world applications where the environment may change over time or is uncertain.
- Reinforcement learning can be used to solve a wide range of problems, including those that involve decision making, control, and optimization.

Reinforcement Learning Contd…

Disadvantages of Reinforcement learning

- Reinforcement learning is not preferable to use for solving simple problems.
- Reinforcement learning needs a lot of data and a lot of computation
- Reinforcement learning is highly dependent on the quality of the reward function. If the reward function is poorly designed, the agent may not learn the desired behavior.
- Reinforcement learning can be difficult to debug and interpret. It is not always clear why the agent is behaving in a certain way, which can make it difficult to diagnose and fix problems.

Process of Machine Learning


Data Collection and Integration
Data Visualization and Analysis
Feature Selection and Engineering
Model Training
Model Evaluation
Prediction

1. Data Collection and integration:

- The first step of the ML process involves the collection of data and integration of data.
- Data collected acts as an input to the model (data preparation phase)
- Inputs are called features.
- The more the data is, more the better our model becomes.
- Once the data is collected we need to integrate and prepare the data.
- Integration of data means placing all related data together.
- Then data preparation phase starts in which we manually and critically explore the data.
- The data preparation phase tells the developer that is the data matching the expectations. Is there enough info to make an accurate prediction? Is the data consistent?

Example:

- Data collected in the case of our considered example involves a lot of data. The collected data should answer the following questions- What is past customer history? What were the past orders? Is the customer a prime member of our bookstore? Does the customer own a kindle? Has the customer made any previous complaints? What was the most number of complaints?

2. Exploratory Data Analysis and Visualization:

- Once the data is prepared developer needs to visualize the data to have a better understanding of relationships within the dataset.
- When we get to see data, we can notice the unseen patterns that we may not have noticed in the first phase.
- It helps developers easily identify missing data and outliers.
- Data visualization can be done by plotting histograms, scatter plots, etc.
- After visualization is done data is analyzed so that developer can decide what ML technique he may use.
- In the considered example case unsupervised learning may be used to analyze customer purchasing habits.

3. Feature Selection and Engineering:

- Feature selection means selecting what features the developer wants to use within the model.
- Features should be selected so that a minimum correlation exists between them and a maximum correlation exists between the selected features and output.
- Feature engineering is the process to manipulate the original data into new and potential data that has a lot many features within it.
- In simple words Feature engineering is converting raw data into useful data or getting the maximum out of the original data.
- Feature engineering is arguably the most crucial and time-consuming step of the ML pipeline.
- Feature selection and engineering answers questions – Are these features going to make any sense in our prediction?
- It deals with the accuracy and precision of data.

4. Model Training:

- After the first three steps are done completely we enter the model training phase.
- It is the first step officially when the developer gets to train the model on basis of data.
- To train the model, data is split into three parts- Training data, validation data, and test data.
- Around 70%-80% of data goes into the training data set which is used in training the model.
- Validation data is also known as development set or dev set and is used to avoid overfitting or underfitting situations i.e. enabling hyperparameter tuning.

4. Model Training…

- Hyperparameter tuning is a technique used to combat overfitting and underfitting.
- Validation data is used during model evaluation.
- Around 10%-15% of data is used as validation data.
- Rest 10%-15% of data goes into the test data set. Test data set is used for testing after the model preparation.

5. Model Evaluation:

- After the model training, validation, or development data is used to evaluate the model.
- To get the most accurate predictions to test data may be used for further model evaluation.
- A confusion matrix is created after model evaluation to calculate accuracy and precision numerically.
- After model evaluation, our model enters the final stage that is prediction.

6. Prediction:

- In the prediction phase developer deploys the model.
- After model deployment, it becomes ready to make predictions.
- Predictions are made on training data and test data to have a better understanding of the build model.

Bias

- The bias is known as the
  - difference between the prediction of the values by the ML model and the correct/ actual value.
  - Being high in biasing gives a large error in training as well as testing data. Its recommended that an algorithm should always be low biased to avoid the problem of underfitting.
  - By high bias, the data predicted is in a straight line format, thus not fitting accurately in the data in the data set.
  - Such fitting is known as **Underfitting of Data**. This happens when the hypothesis is too simple or linear in nature.

Variance

- The variability of model prediction for a given data point which tells us spread of our data is called the variance of the model.
- The model with high variance has a very complex fit to the training data and thus is not able to fit accurately on the data which it hasn’t seen before. As a result, such models perform very well on training data but has high error rates on test data.
- When a model is high on variance, it is then said to as **Overfitting of Data**.
- Overfitting is fitting the training set accurately via complex curve and high order hypothesis but is not the solution as the error with unseen data is high.
  While training a data model variance should be kept low.

Bias Variance Trade-off

- If the algorithm is too simple (hypothesis with linear eq.) then it may be on high bias and low variance condition and thus is error-prone.
- If algorithms fit too complex (hypothesis with high degree eq.) then it may be on high variance and low bias.
- In the latter condition, the new entries will not perform well. Well, there is something between both of these conditions, known as Trade-off or Bias Variance Trade-off.
- This tradeoff in complexity is why there is a tradeoff between bias and variance. An algorithm can’t be more complex and less complex at the same time.

Underfitting

- A statistical model or a machine learning algorithm is said to have underfitting when it cannot capture the underlying trend of the data, i.e., it only performs well on training data but performs poorly on testing data.
- Underfitting destroys the accuracy of our machine learning model. Its occurrence simply means that our model or the algorithm does not fit the data well enough.
- It usually happens when we have fewer data to build an accurate model and also when we try to build a linear model with fewer non-linear data.
- In such cases, the rules of the machine learning model are too easy and flexible to be applied to such minimal data and therefore the model will probably make a lot of wrong predictions.
- Underfitting can be avoided by using more data and the features by feature selection.

Reasons for Underfitting

- High bias and low variance
- The size of the training dataset used is not enough.
- The model is too simple.
- Training data is not cleaned and also contains noise in it.

Techniques to reduce underfitting

- Increase model complexity
- Increase the number of features, performing feature engineering
- Remove noise from the data.
- Increase the number of epochs or increase the duration of training to get better results.

Overfitting

- A statistical model is said to be over fitted when the model does not make accurate predictions on testing data.
- When a model gets trained with so much data, it starts learning from the noise and inaccurate data entries in our data set. And when testing with test data results in High variance.
- Then the model does not categorize the data correctly, because of too many details and noise. The causes of overfitting are the non-parametric and non-linear methods because these types of machine learning algorithms have more freedom in building the model based on the dataset and therefore they can really build unrealistic models.
- A solution to avoid overfitting is using a linear algorithm if we have linear data or using the parameters like the maximal depth if we are using decision trees.

Reasons for Overfitting are as follows:

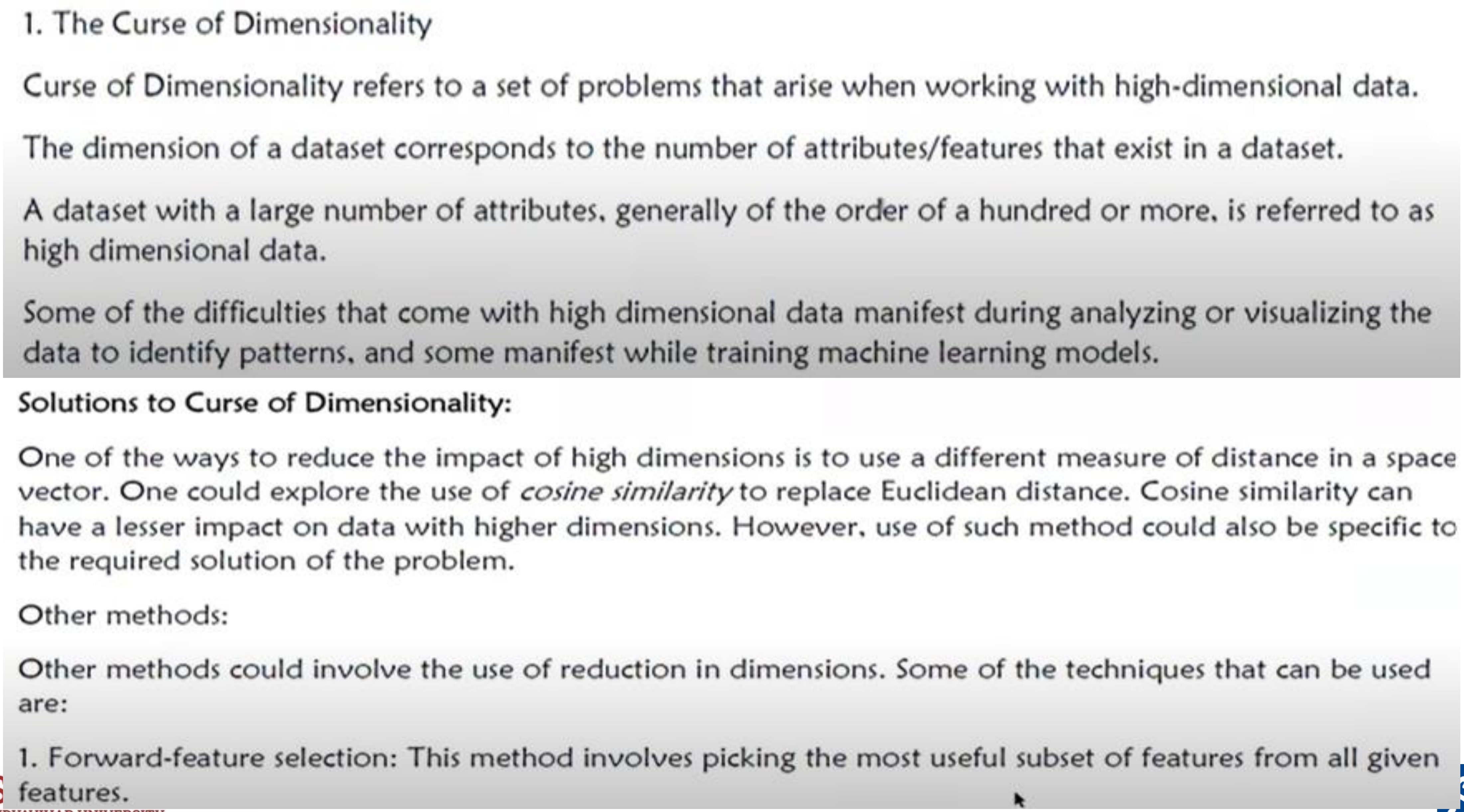
- High variance and low bias
- The model is too complex
- The size of the training data

Techniques to reduce overfitting

- Increase training data.
- Reduce model complexity.
- Early stopping during the training phase
- Have an eye over the loss over the training period as soon as loss begins to increase stop training).
- Use dropout for neural networks to tackle overfitting.

Weight Space

Curse of Dimensionality



1. The Curse of Dimensionality

Curse of Dimensionality refers to a set of problems that arise when working with high-dimensional data.

The dimension of a dataset corresponds to the number of attributes/features that exist in a dataset.

A dataset with a large number of attributes, generally of the order of a hundred or more, is referred to as high dimensional data.

Some of the difficulties that come with high dimensional data manifest during analyzing or visualizing the data to identify patterns, and some manifest while training machine learning models.

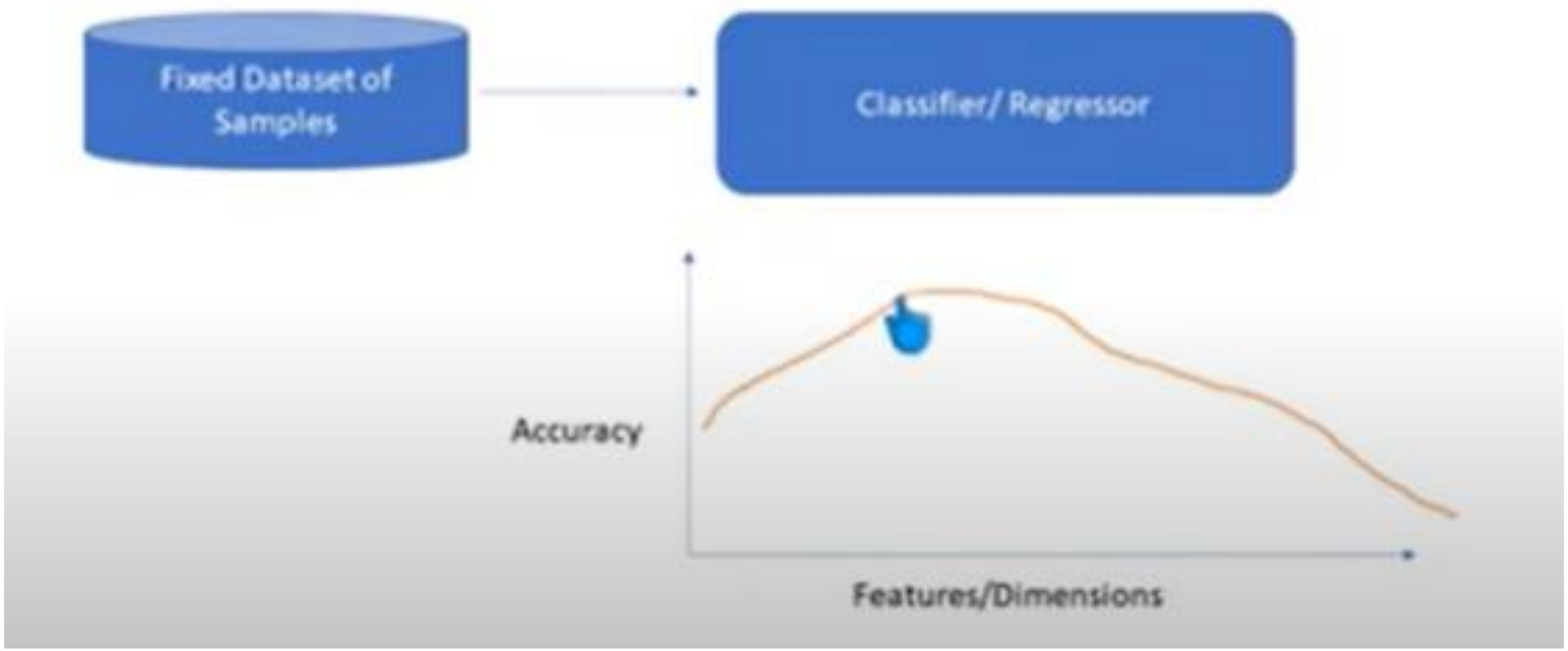
Solutions to Curse of Dimensionality:

One of the ways to reduce the impact of high dimensions is to use a different measure of distance in a space vector. One could explore the use of *cosine similarity* to replace Euclidean distance. Cosine similarity can have a lesser impact on data with higher dimensions. However, use of such method could also be specific to the required solution of the problem.

Other methods:

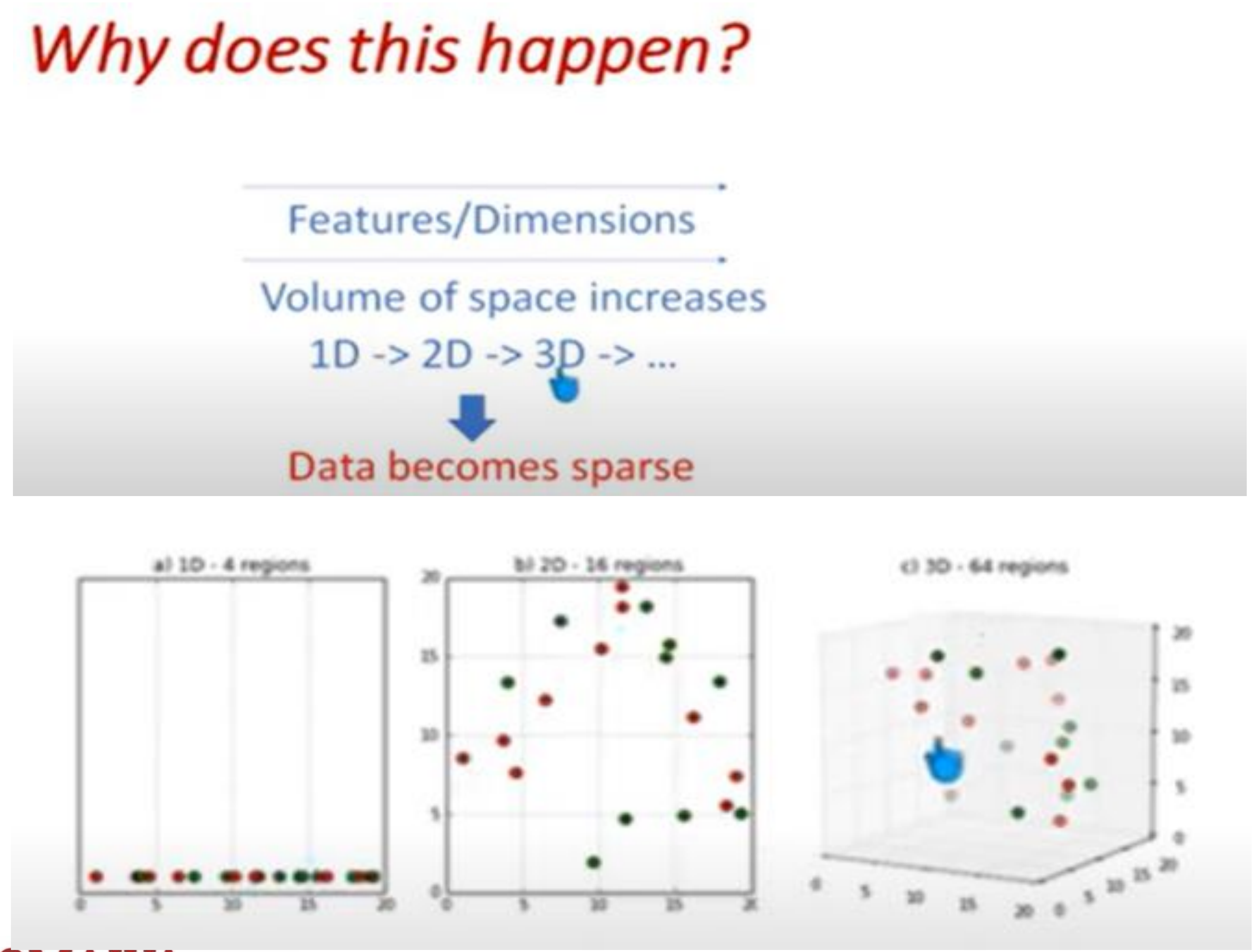
Other methods could involve the use of reduction in dimensions. Some of the techniques that can be used are:

1. Forward-feature selection: This method involves picking the most useful subset of features from all given features.



Fixed Dataset of Samples

Classifier/ Regressor



Why does this happen?

Features/Dimensions

Volume of space increases

1D -> 2D -> 3D -> ...

Data becomes sparse

In Order to obtain the reliable results, the amount of data needed often grows exponentially with the Dimensionality

Naïve Bayes Classifier Algorithm

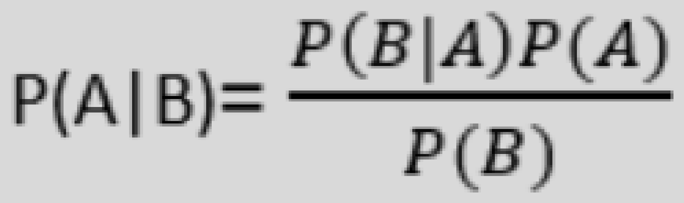
- Naïve Bayes algorithm is a supervised learning algorithm, which is based on **Bayes theorem** and used for solving classification problems.
- It is mainly used in *text classification* that includes a high-dimensional training dataset.
- Naïve Bayes Classifier is one of the simple and most effective Classification algorithms which helps in building the fast machine learning models that can make quick predictions.
- **It is a probabilistic classifier, which means it predicts on the basis of the probability of an object**.
- Some popular examples of Naïve Bayes Algorithm are **spam filtration, Sentimental analysis, and classifying articles**.

Why is it called Naïve Bayes?

- The Naïve Bayes algorithm is comprised of two words Naïve and Bayes, Which can be described as:
- **Naïve**: It is called Naïve because it assumes that the occurrence of a certain feature is independent of the occurrence of other features.
- Such as if the fruit is identified on the bases of color, shape, and taste, then red, spherical, and sweet fruit is recognized as an apple. Hence each feature individually contributes to identify that it is an apple without depending on each other.
- **Bayes**: It is called Bayes because it depends on the principle of Bayes' Theorem.

Bayes' Theorem:

- Bayes' theorem is also known as **Bayes' Rule** or **Bayes' law**, which is used to determine the probability of a hypothesis with prior knowledge. It depends on the conditional probability.
- The formula for Bayes' theorem is given as:



$$P(A|B)=\frac{P(B|A)P(A)}{P(B)}$$

Where,

P(A|B) is Posterior probability: Probability of hypothesis A on the observed event B.

P(B|A) is Likelihood probability: Probability of the evidence given that the probability of a hypothesis is true.

P(A) is Prior Probability: Probability of hypothesis before observing the evidence.

P(B) is Marginal Probability: Probability of Evidence.

Example:

- Suppose we have a dataset of **weather conditions** and corresponding target variable "**Play**". So using this dataset we need to decide that whether we should play or not on a particular day according to the weather conditions. So to solve this problem, we need to follow the below steps:
  - Convert the given dataset into frequency tables.
  - Generate Likelihood table by finding the probabilities of given features.
  - Now, use Bayes theorem to calculate the posterior probability.
- **Problem**: If the weather is sunny, then the Player should play or not?

	Outlook	Play
0	Rainy	Yes
1	Sunny	Yes
2	Overcast	Yes
3	Overcast	Yes
4	Sunny	No
5	Rainy	Yes
6	Sunny	Yes
7	Overcast	Yes
8	Rainy	No
9	Sunny	No
10	Sunny	Yes
11	Rainy	No
12	Overcast	Yes
13	Overcast	Yes

1. Frequency table for the Weather Conditions:

Weather	Yes	No
Overcast	5	0
Rainy	2	2
Sunny	3	2
Total	10	5

2. Likelihood table weather condition:

Weather	No	Yes	
Overcast	0	5	5/14= 0.35
Rainy	2	2	4/14=0.29
Sunny	2	3	5/14=0.35
All	4/14=0.29	10/14=0.71	

3. Applying Bayes'theorem:

P(Yes|Sunny)= P(Sunny|Yes)*P(Yes)/P(Sunny)

P(Sunny|Yes)= 3/10= 0.3

P(Sunny)= 0.35

P(Yes)=0.71

So P(Yes|Sunny) = 0.3*0.71/0.35= **0.60**

P(No|Sunny)= P(Sunny|No)*P(No)/P(Sunny)

P(Sunny|NO)= 2/4=0.5

P(No)= 0.29

P(Sunny)= 0.35

So P(No|Sunny)= 0.5*0.29/0.35 = **0.41**

So as we can see from the above calculation that **P(Yes|Sunny)>P(No|Sunny)**

Hence on a Sunny day, Player can play the game.

Example of Naïve Bayesian:
Unknown sample---- { Red, SUV, Domestic,**?**}

Example No.	Color	Type	Origin	Stolen?
1	Red	Sports	Domestic	Yes
2	Red	Sports	Domestic	No
3	Red	Sports	Domestic	Yes
4	Yellow	Sports	Domestic	No
5	Yellow	Sports	Imported	Yes
6	Yellow	SUV	Imported	No
7	Yellow	SUV	Imported	Yes
8	Yellow	SUV	Domestic	No
9	Red	SUV	Imported	No
10	Red	Sports	Imported	Yes

Color	
P(Red|Yes)=3/5	P(Red|No)=2/5
P(Yellow|Yes)=2/5	P(Yellow|No)=3/5
Type	
P(SUV|Yes)=1/5	P(SUV|No)=3/5
P(Sports|Yes)=4/5	P(Sports|No)=2/5
Origin	
P(Domestic|Yes)=2/5	P(Domestic|No)=3/5
P(Imported|Yes)=3/5	P(Imported|No)=2/5

P(Yes) = P(Yes) * P(Red | Yes) * P(SUV | Yes) * P(Domestic|Yes)
= 5/10 * 3/5 * 2/5 * 1/5 = 0.024
and for
P(No) =
P(No) * P(Red | No) * P(SUV | No) * P (Domestic | No)
= 5/10 * 2/5 * 3/5 * 3/5 = 0.072

Since 0.072 > 0.024, our example gets classified as ’NO’

- **Advantages of Naïve Bayes Classifier:**
  - Naïve Bayes is one of the fast and easy ML algorithms to predict a class of datasets.
  - It can be used for Binary as well as Multi-class Classifications.
  - It performs well in Multi-class predictions as compared to the other Algorithms.
  - It is the most popular choice for **text classification problems**.
- **Disadvantages of Naïve Bayes Classifier:**
  - Naive Bayes assumes that all features are independent or unrelated, so it cannot learn the relationship between features.

- Applications of Naïve Bayes Classifier:
- It is used for **Credit Scoring**.
- It is used in **medical data classification**.
- It can be used in **real-time predictions** because Naïve Bayes Classifier is an eager learner.
- It is used in Text classification such as **Spam filtering** and **Sentiment analysis**.

Thank You