Advance Machine Learning

01CO1301
4 Credits

Marwadi University

Department of Computer Engineering

Ravikumar R Natarajan

Course Outcomes

- At the end of the course, students will be able to:
- To understand key concepts, tools and approaches for pattern recognition on complex data sets.
- To learn Kernel methods for handling high dimensional and non-linear patterns.
- To implement state-of-the-art algorithms such as Support Vector Machines and Bayesian networks.
- To Solve real-world machine learning tasks: from data to inference.
- To apply theoretical concepts and the motivations behind different learning frameworks.

Marwadi University

Department of Computer Engineering

Key Concepts

Unit #1

Content

- Supervised/Unsupervised Learning
- Loss functions and generalization
- Probability Theory
- Parametric vs Nonparametric methods
- Elements of Computational Learning
- Theory Ensemble Learning
- Bagging
- Boosting
- Random Forest

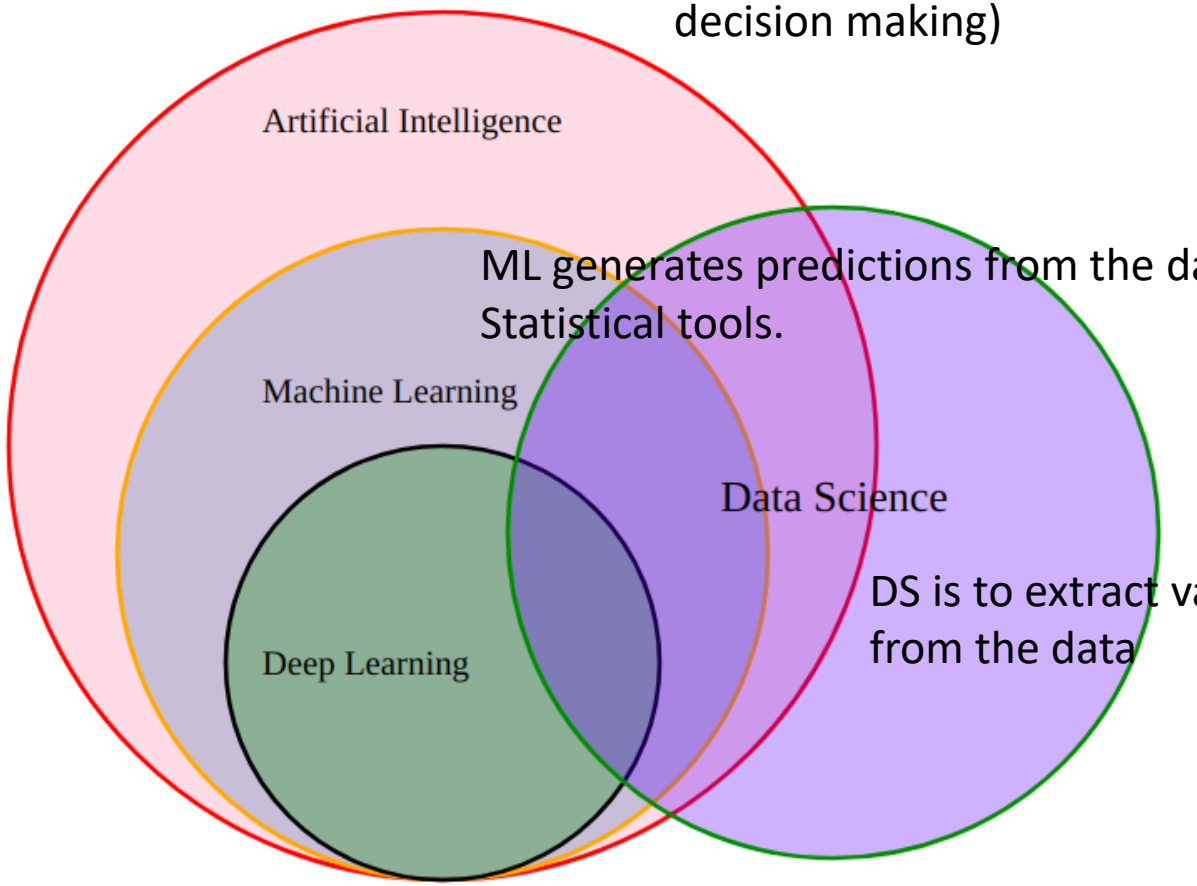
Machine Learning Introduction

ML is an interdisciplinary field:

- **Data Analyst:** visualize, analyze data, optimization
- **Data Engineers:** build and test scalable/stable/optimal ecosystems for data scientists to run their algorithms
- **Database Administrator:** responsible for the proper functioning of all the databases.
- **Data Scientist:** perform predictive analysis and offer actionable insights.
- **Statistician:** extract and offer valuable insights from the data using statistical theory and tools.

AI vs ML vs DL vs DS

AI enables the machine to think (autonomous decision making)



ML generates predictions from the data using Statistical tools.

DS is to extract valuable information from the data

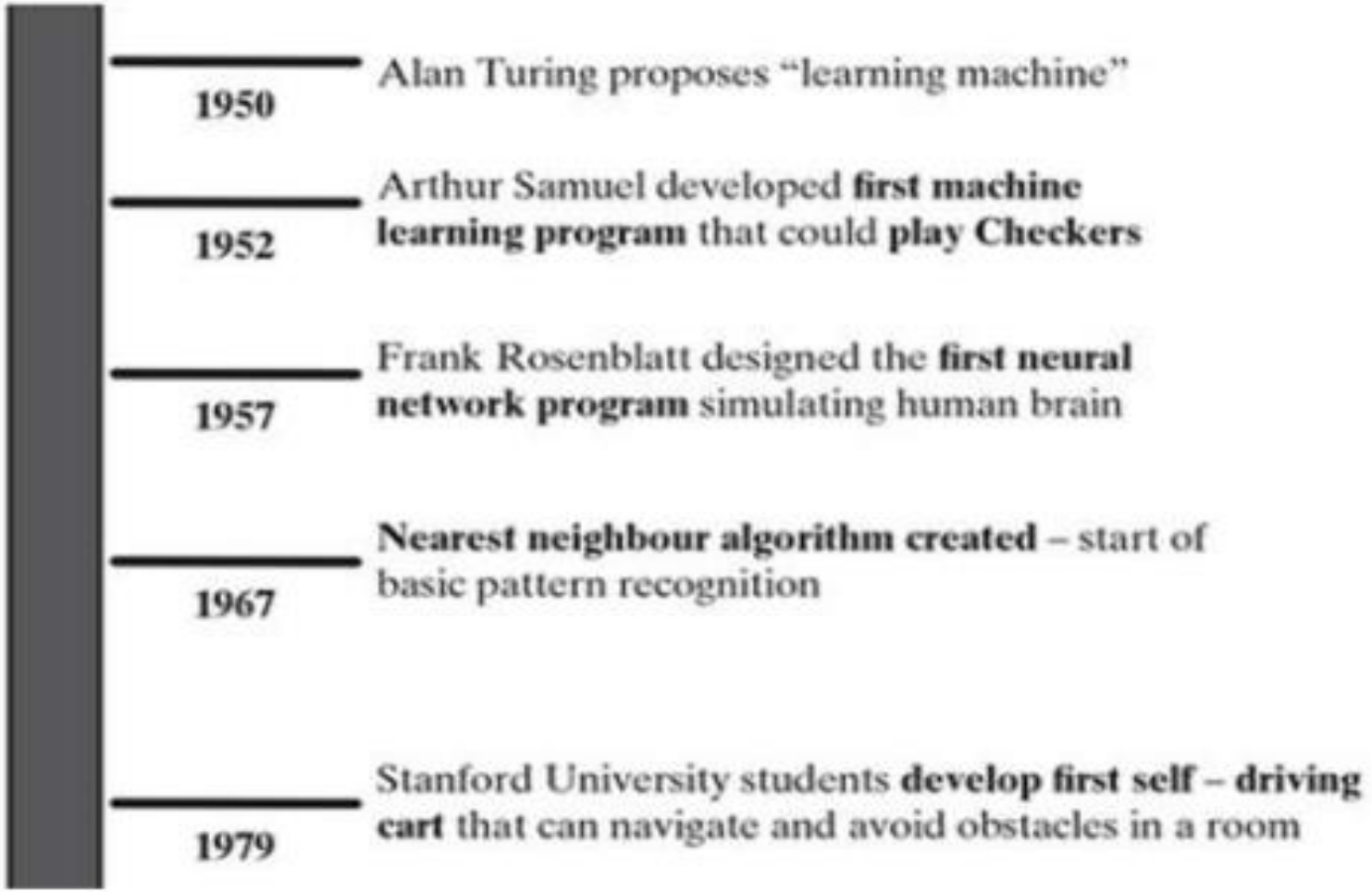
DL is to mimic human brain. MNN – multi neural network architecture

Machine Learning Introduction

- AI stands for **Artificial Intelligence**, and is basically the study/process which enables machines to mimic human behaviour through particular algorithm.
- ML stands for **Machine Learning**, and is the study that uses statistical methods enabling machines to improve with experience.
- DL stands for **Deep Learning**, and is the study that makes use of Neural Networks(similar to neurons present in human brain) to imitate functionality just like a human brain.
- **Data science** is the field of applying advanced analytics techniques and scientific principles to extract valuable information from data for business decision-making, strategic planning and other uses.

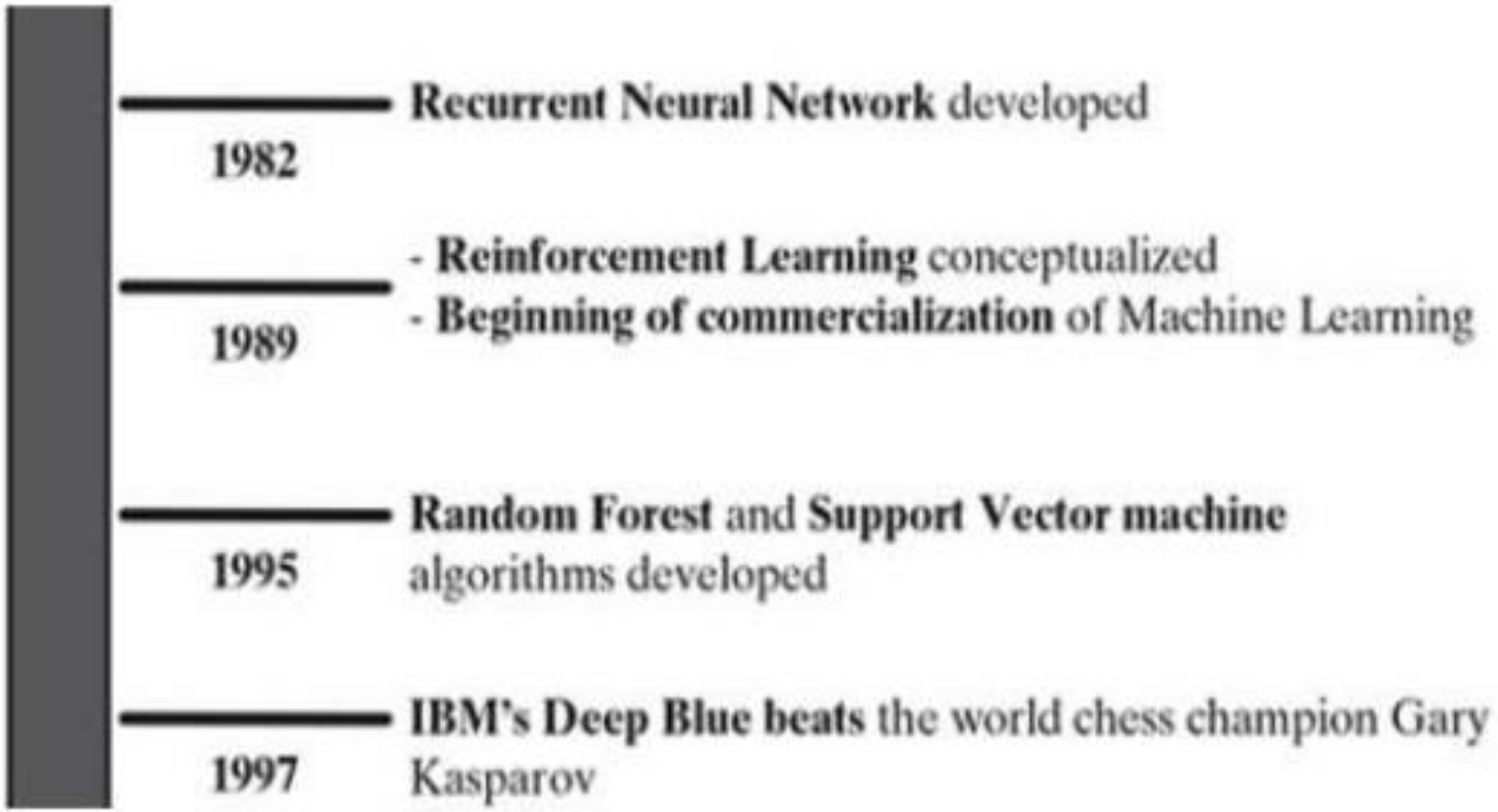
Machine Learning Introduction

Evaluation of Machine Learning



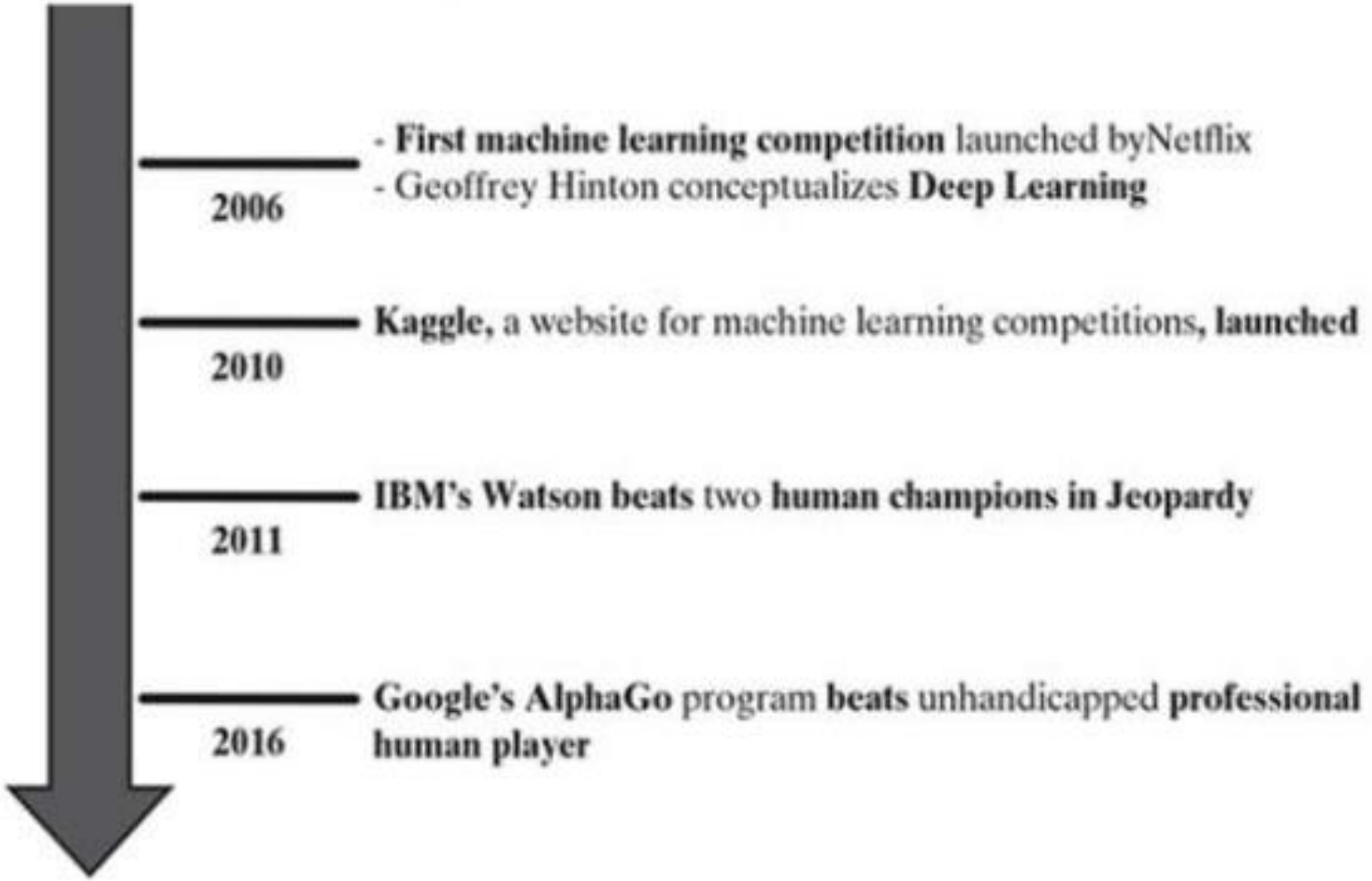
Machine Learning Introduction

Evaluation of Machine Learning



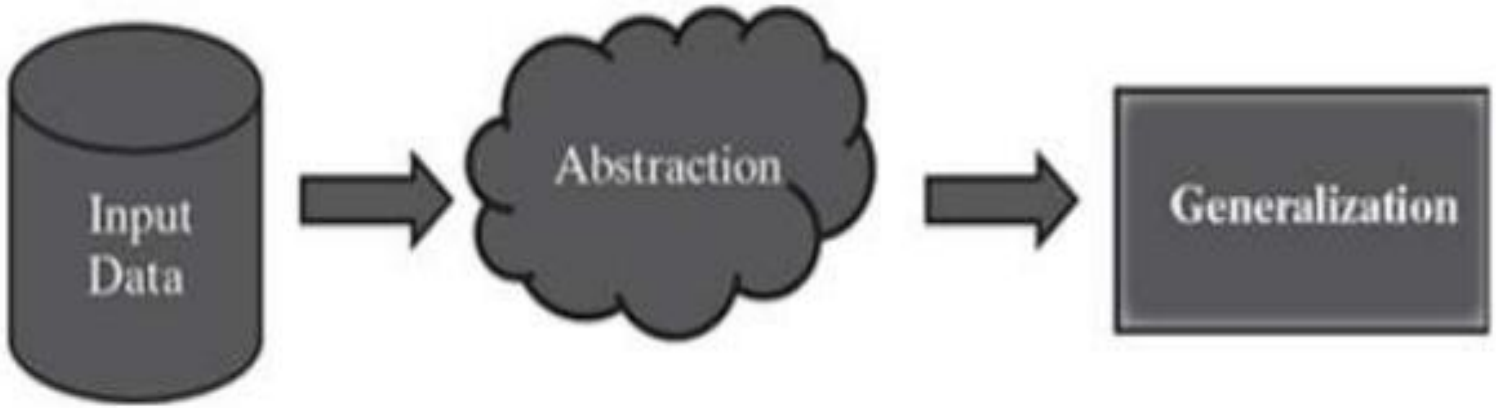
Machine Learning Introduction

Evaluation of Machine Learning



How do machine learn?

- Data Input: Past data or information is utilized as a basis for future decision-making.
- Abstraction: The input data is represented in a broader way through the underlying algorithm.
- Generalization: The abstracted representation is generalized to form a framework for making decisions.

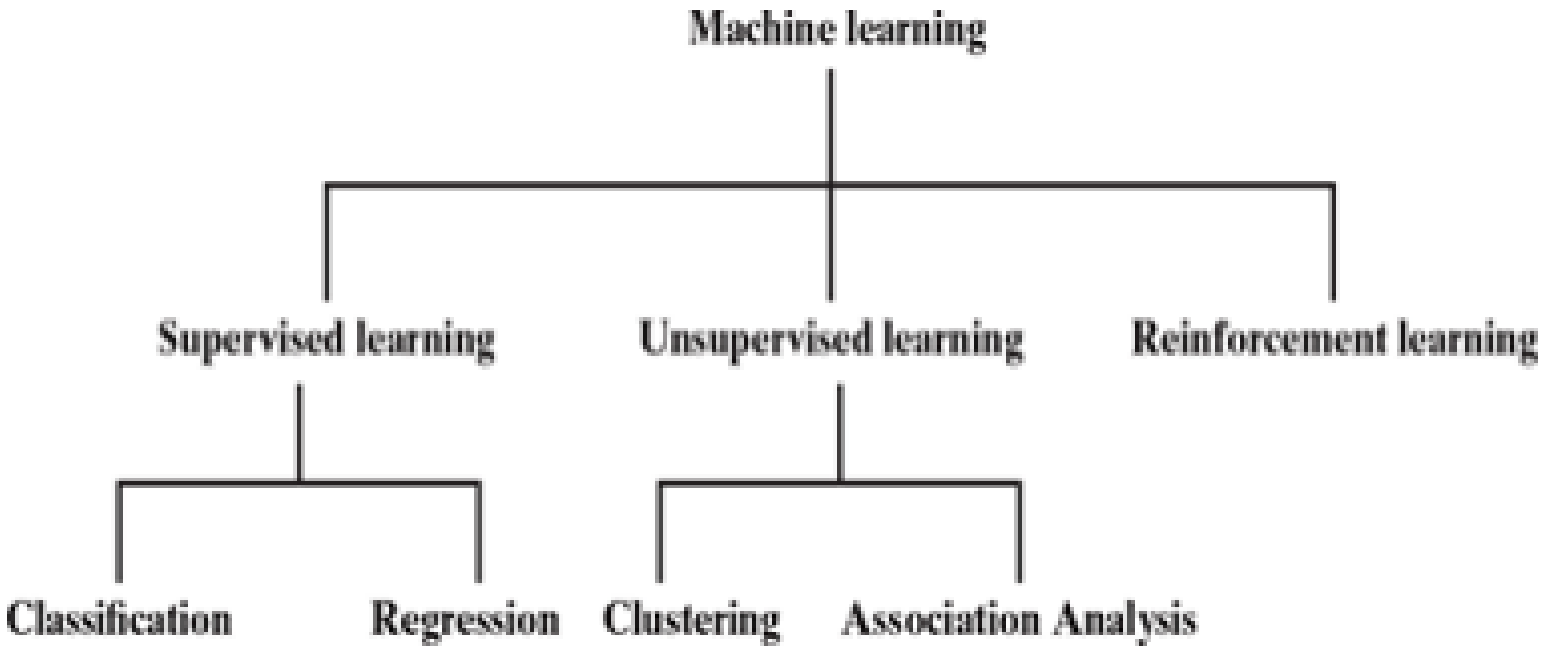


Did You Know?

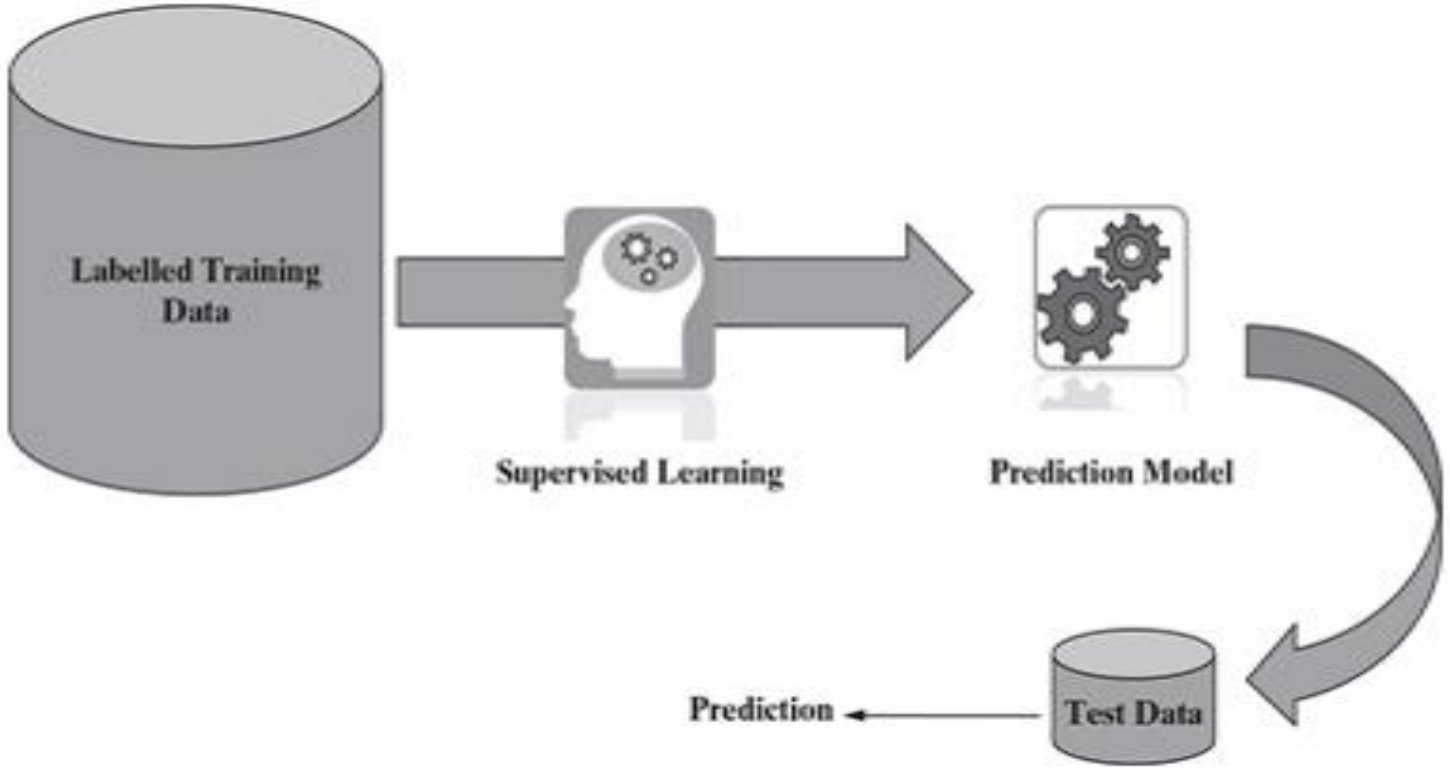
AIBO

Sony created a series of robotic pets called Aibo. It was built in 1998. Although most models sold were dog-like, other inspirations included lion-cubs. It could express emotions and could also recognize its owner. In 2006, Aibo was added to the Carnegie Mellon University's 'Robot Hall of Fame'. A new generation of Aibo was launched in Japan in January 2018.
https://us.aibo.com

Types of Machine Learning



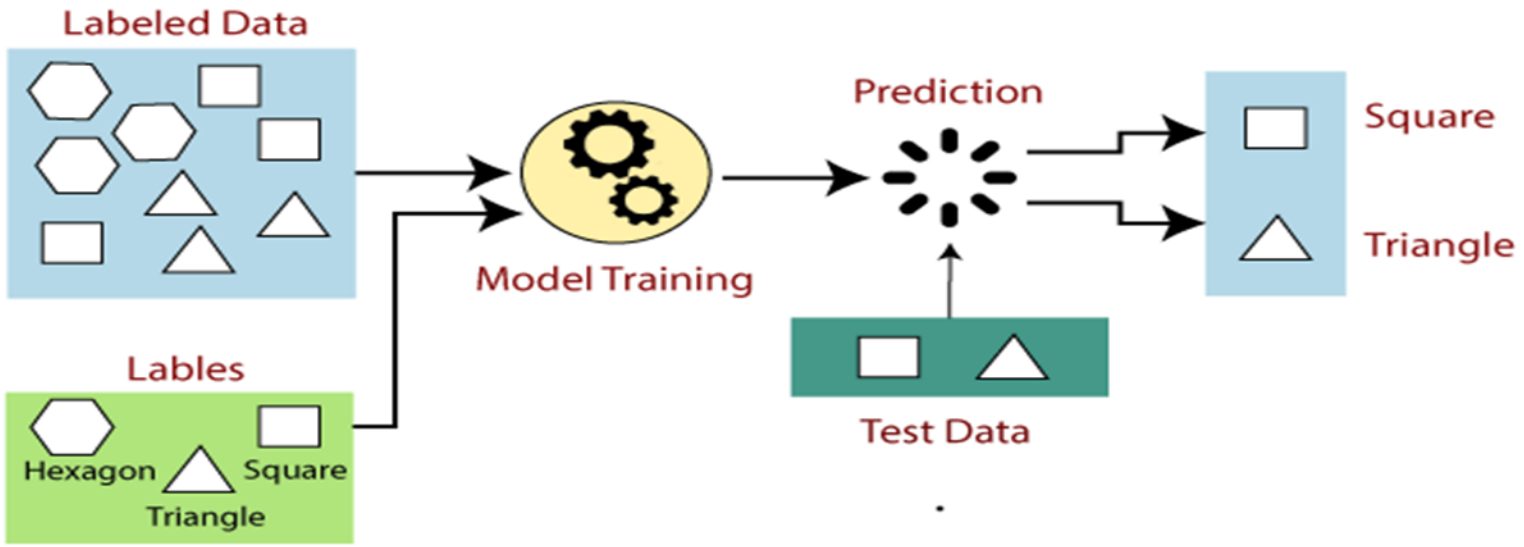
Supervised Learning



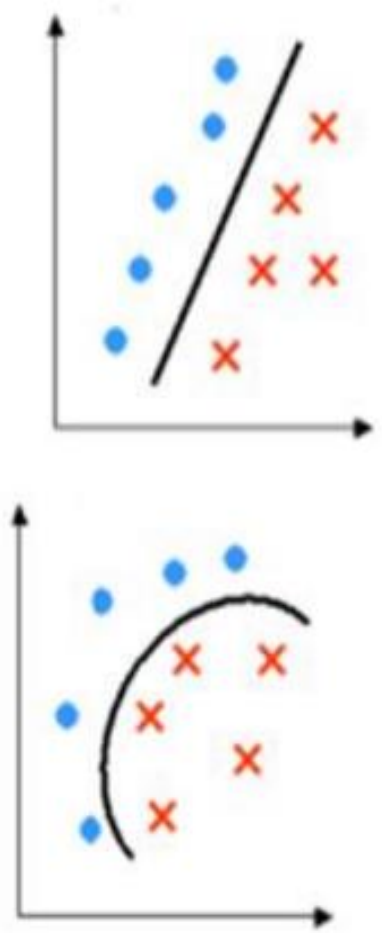
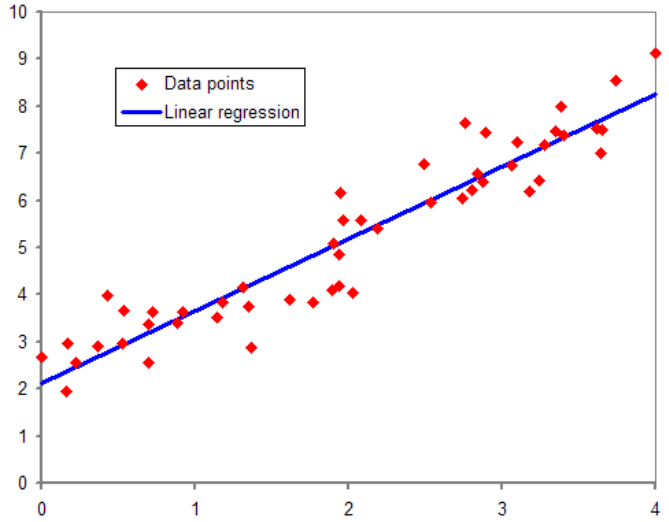
Supervised Learning

Supervised Learning

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

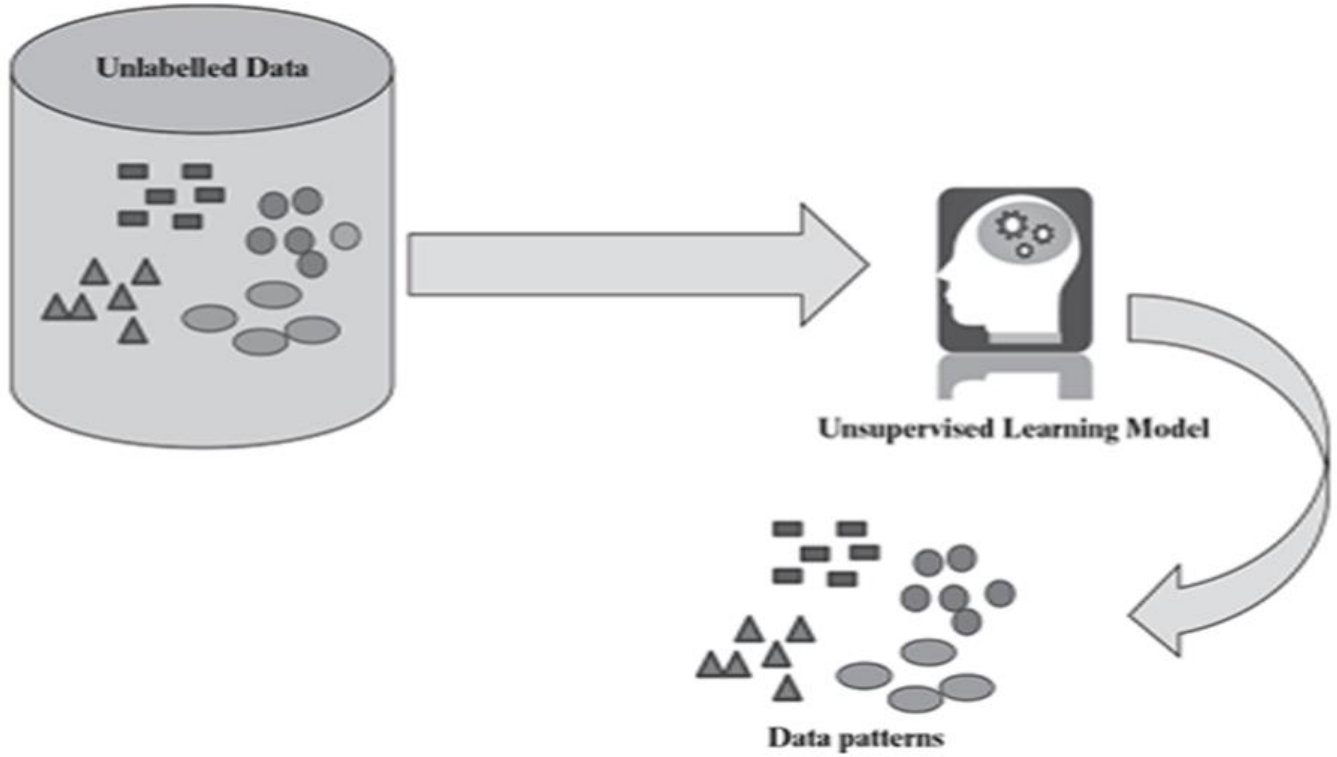


Types of Supervised Learning

Classification (Discrete value output)	Regression (Predict real value output)
	

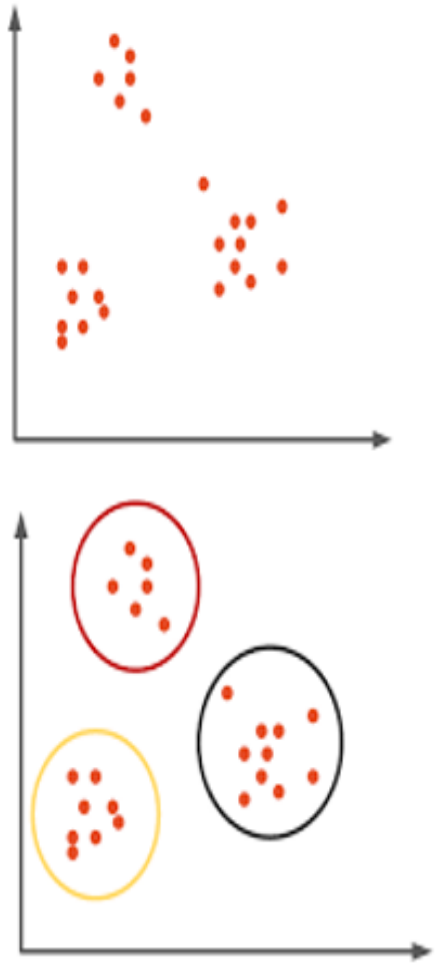
Unsupervised Learning

- **Unsupervised learning is a machine learning technique in which models are not supervised using training dataset.**
- Instead, models itself find the hidden patterns and insights from the given data. It can be compared to learning which takes place in the human brain while learning new things.



Types of Unsupervised Learning

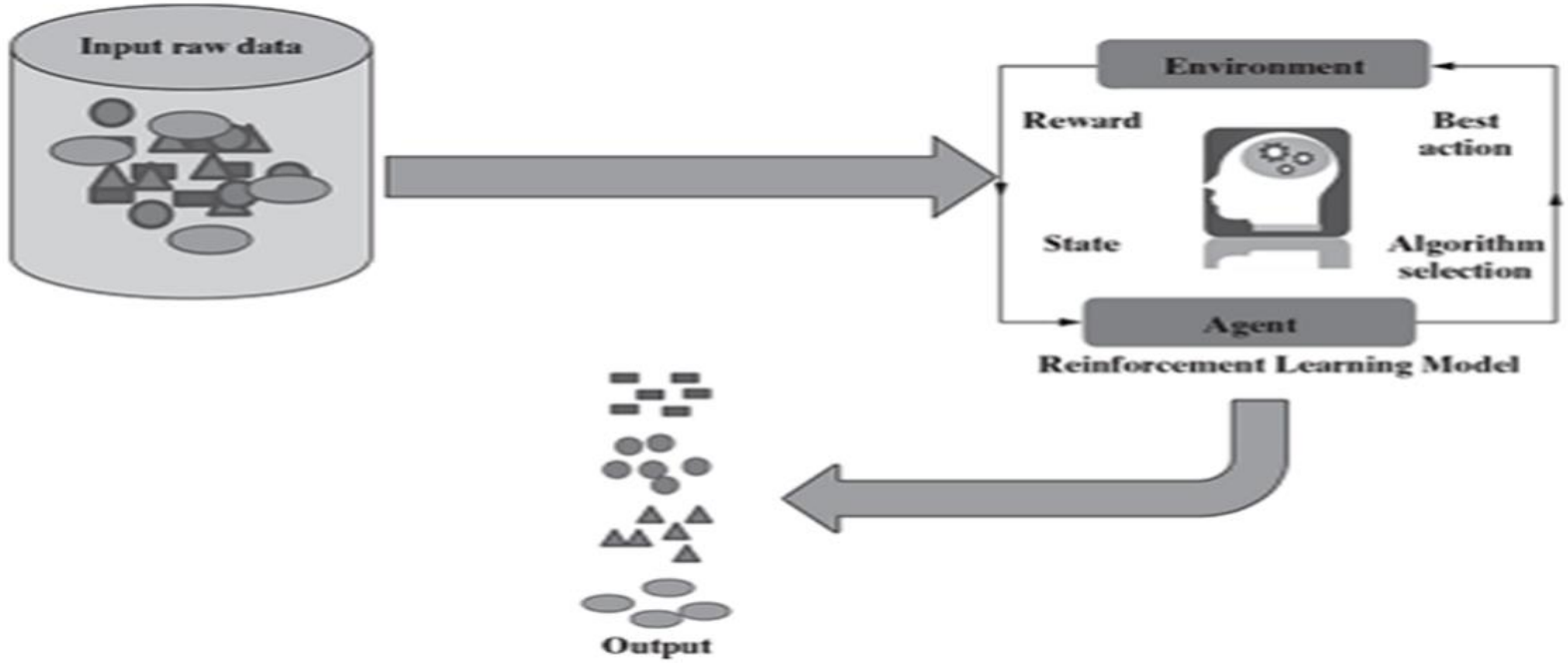
Clustering



Association

Reinforcement Learning

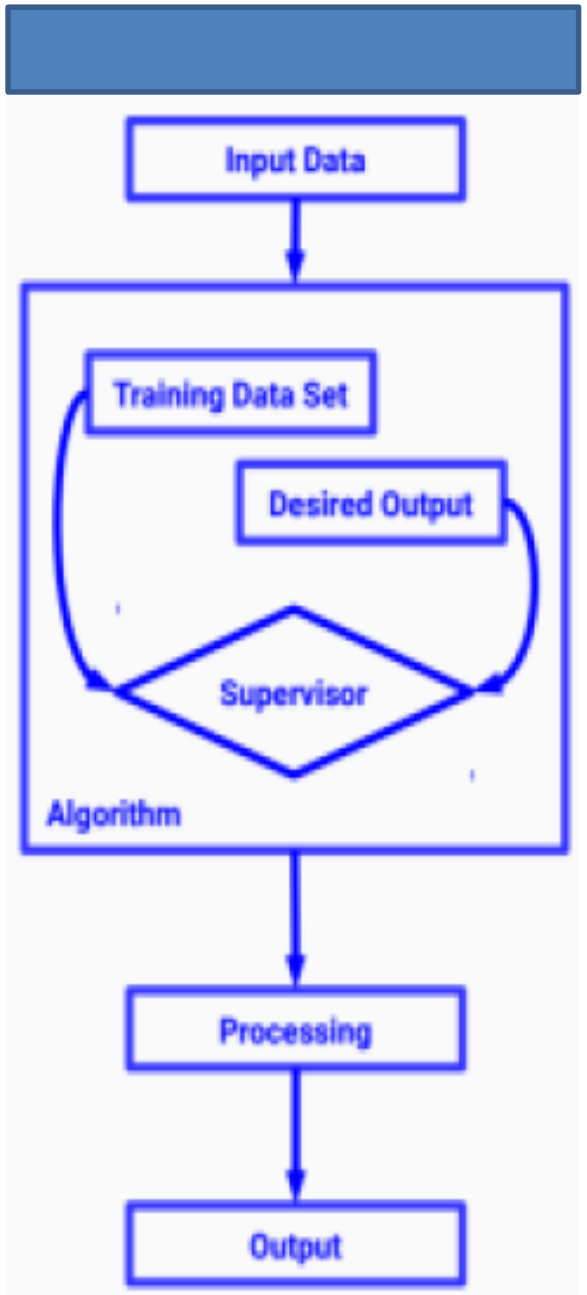
- Reinforcement Learning is a feedback-based (reward) Machine learning technique in which an agent learns to behave in an environment by performing the actions and seeing the results of actions.
- For each good action, the agent gets positive feedback, and for each bad action, the agent gets negative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

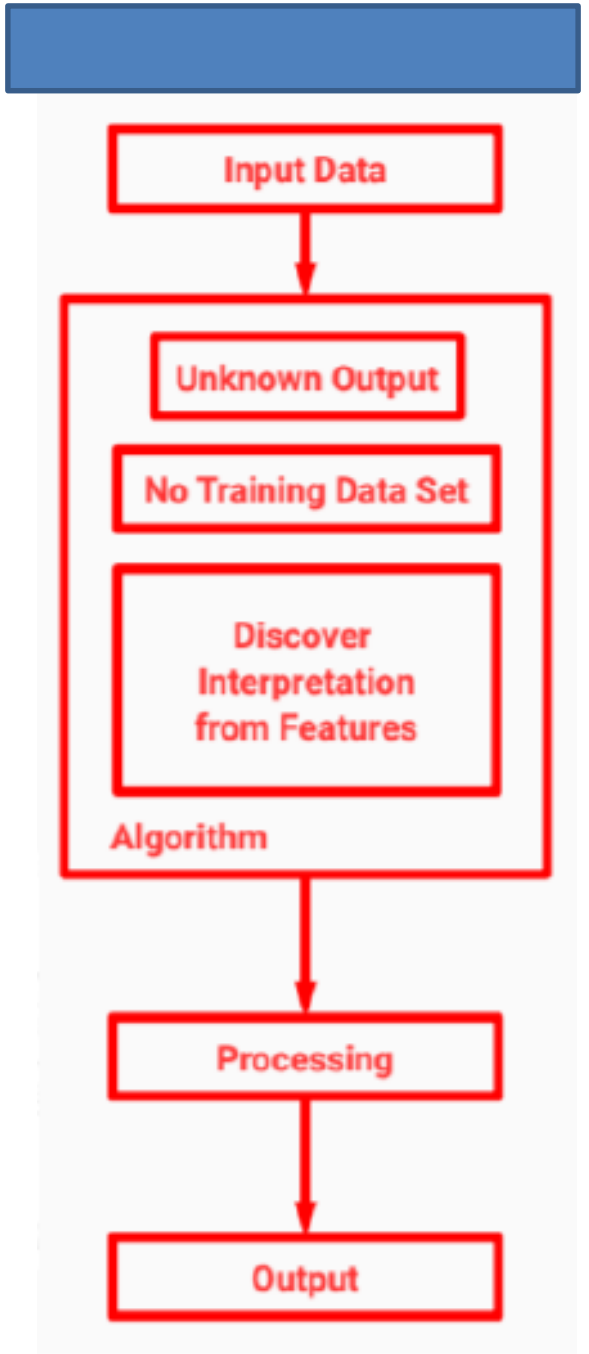


Supervised vs Unsupervised vs Reinforcement Learning

Criteria	Supervised ML	Unsupervised ML	Reinforcement ML
Definition	Learns by using labelled data	Trained using unlabelled data without any guidance.	Works on interacting with the environment (reward based)
Type of data	Labelled data	Unlabelled data	No – predefined data
Type of problems	Regression and classification	Association and Clustering	Exploitation or Exploration
Supervision	Extra supervision	No supervision	No supervision
Algorithms	Linear Regression, Logistic Regression, SVM, KNN, NB, DT.	K – Means, PCA, DBSCAN, Apriori	Q – Learning, SARSA
Aim	Calculate outcomes	Discover underlying patterns	Learn a series of action
Application	Risk Evaluation, Forecast Sales	Recommendation System, Anomaly Detection	Self Driving Cars, Gaming, Healthcare

Supervised Learning





Did you know?

- Many video games are based on artificial intelligence technique called **Expert System**.
- This technique can imitate areas of human behaviour, with a goal to mimic the human ability of senses, perception, and reasoning.

Linear Regression with one Variable: Model and Cost Function, Model representation, Cost Function

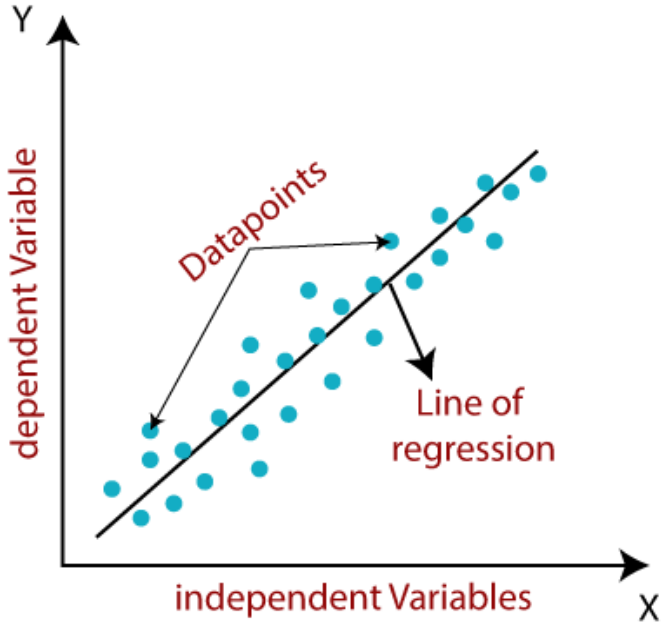
Parameter Learning: Gradient Descent for single variable

Did you know?

- It is a statistical method that is used for predictive analysis. **Linear regression makes predictions for continuous/real or numeric variables such as sales, salary, age, product price, etc.**
- Linear Regression is an approach to show the relationship between the independent variable x and dependent variable y.

y = mx + c

- **m: slope of line**
- **c: constant**
- **x: Independent variable**
- **y: Dependent variable**



- Our goal is to find the fit of the line. The best fit means where the error is minimum. It can make our prediction more accurate.

Linear Regression with One Variable

Types of Linear Regression

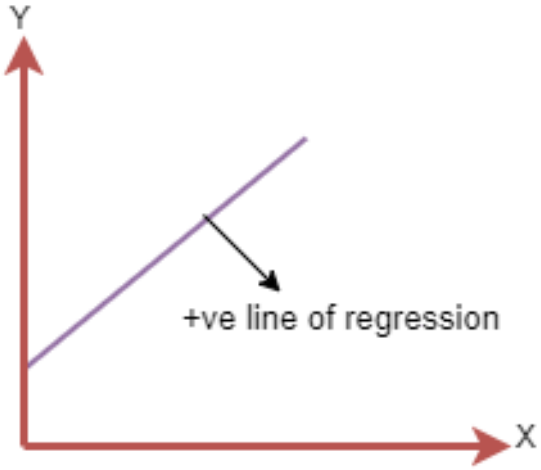
- Simple Linear Regression
- If a single independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Simple Linear Regression.

- Multiple Linear regression
- If more than one independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Multiple Linear Regression.

Linear Regression with One Variable

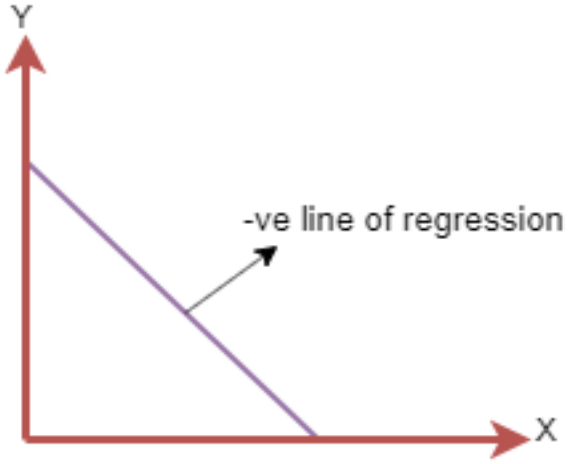
Linear Regression Line

Positive Linear Relationship



The line equation will be: $Y= a_0+a_1x$

Negative Linear Relationship



The line of equation will be: $Y= -a_0+a_1x$

Linear Regression with One Variable

How to find "Best Fit" line?

- Our main goal is to find the best fit line that means the error between predicted values and actual values should be minimized. The best fit line will have the least error.
- The different values for weights or the coefficient of lines (ao, a1) gives a different line of regression, so we need to calculate the best values for ao and a1 to find the best fit line, so to calculate this we use cost function.
- **Cost function** optimizes the regression **coefficients** or **weights**. It measures how a linear regression model is performing.
- We can use the cost function to find the accuracy of the mapping function, which maps the input variable to the output variable. This mapping function is also known as Hypothesis function.
- For Linear Regression, we use the Mean Squared Error (MSE) cost function, which is the average of squared error occurred between the predicted values and actual values.

Linear Regression with One Variable

$$MSE = 1\frac{1}{N}\sum_{i=1}^{n}(y_i - (a_1 x_i + a_0))^2$$

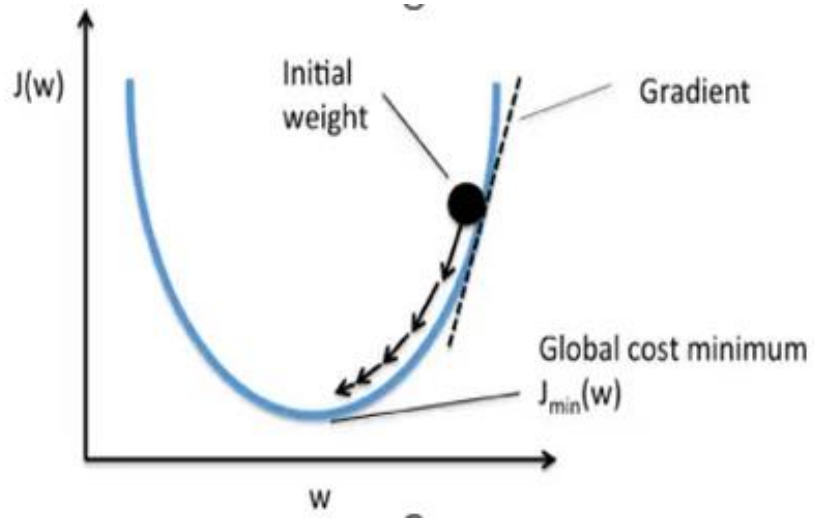
N=Total number of observation
Yi = Actual value
(a1xi+ao)= Predicted value.

Gradient Descent

- Gradient descent is used to minimize (optimize) the MSE by calculating the gradient of the cost function.
- A regression model uses gradient descent to update the coefficients of the line by reducing the cost function.
- It is done by a random selection of values of coefficient and then iteratively update the values to reach the minimum cost function.

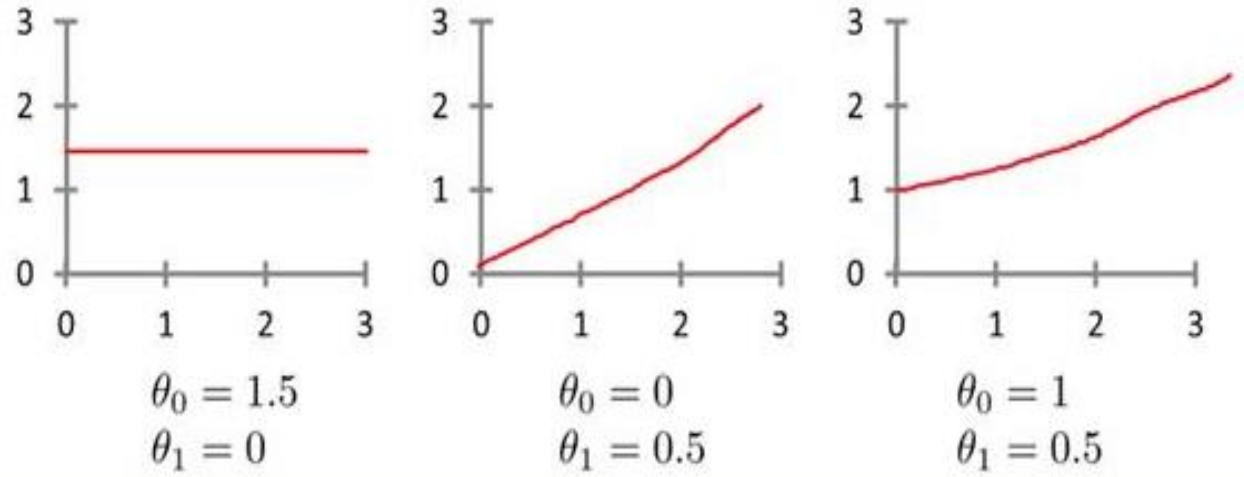
Linear Regression with One Variable

Gradient Descent



Standard Learning Rate
LR = 0.01 (Trial & Error method)

Hypothesis Example: hθ(x) = θ0+ θ1 x the equation of the line **θ0,θ1 are parameters** and how the effect,



DEMO

Linear Regression with One Variable

https://colab.research.google.com/drive/13x6JVgXhdbejVywU6bQWS8mV8HLoy7Fo?usp=sharing

Linear Regression with Multiple Variables: Multiple Features
Gradient Descent for Multiple Variables

Linear Regression with Multiple Variables

- Multiple Linear Regression is one of the important regression algorithms which models the linear relationship between a single dependent continuous variable and more than one independent variable.

Size $(feet)^2$ X1	No.of Bedrooms x2	No.of Floors x3	Age of Home x4	Price ($) y
2104	5	1	45	460
1416	3	2	40	232
1534	3	2	30	315

$$h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 + \theta_4 x_4$$

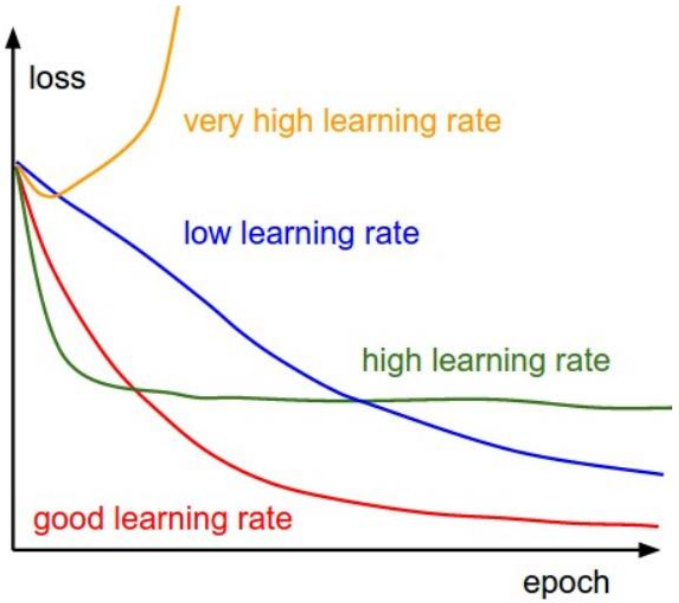
Linear Regression with Multiple Variables

Gradient Descent for Multiple Variables

$$h_\theta(x) = \theta^T x = \theta_0 x_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_n x_n$$

- Cost Function

$$\text{MSE} = \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2$$



DEMO

Linear Regression with Multiple Variables

https://colab.research.google.com/drive/1iqtvvlGkyBk5C5eWKlpGAXkRVbVk-Joi?usp=sharing

Polynomial Regression

- Polynomial regression is a form of regression analysis in which the relationship between the independent variable x and dependent variable y is modeled as an nth degree polynomial of x. That is, if your dataset holds the characteristic of being curved when plotted in the graph, then you should go with a polynomial regression model instead of Simple or Multiple Linear regression models.

- The equation for Polynomial Regression looks very similar to that of Multiple Linear Regression.
- $y=b_0+b_1*x_1+b_2*(x_1)_2+....+b_n*(x_1)_n$

- The main difference is that in Multiple Linear Regression, there are several variables of the same degree but here the single variable has different powers.

DEMO

Polynomial Regression

Did you know?

Did you know?

- Machine learning saves life – ML can spot 52% of breast cancer cells, a year before patients are diagnosed.
- US Postal Service uses machine learning for handwriting recognition.
- Facebook's news feed uses machine learning to personalize each member's feed.

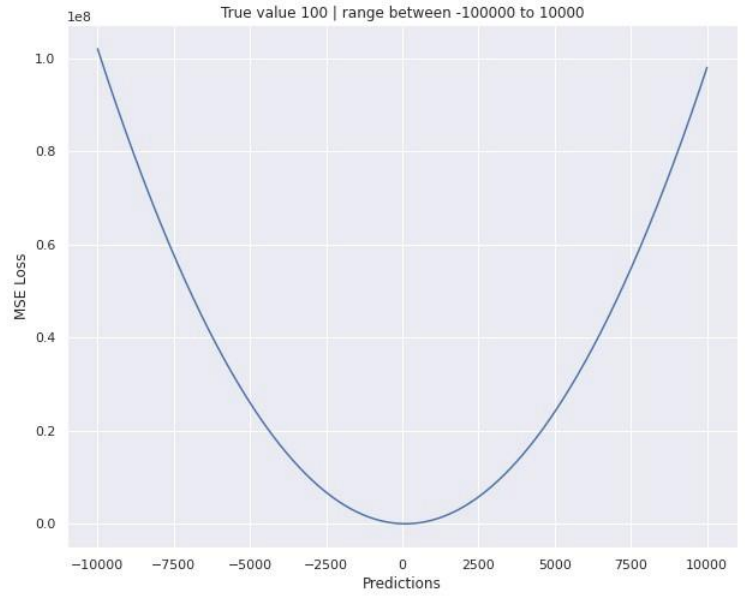
Loss Functions and Generalization

Loss Functions

- It's a method of evaluating how well specific algorithm models the given data.
- If predictions deviates too much from actual results, loss function would cough up a very large number.
- Gradually, with the help of some optimization function, loss function learns to reduce the error in prediction.
- **Regression Losses**
  - **Mean Square Error/Quadratic Loss/L2 Loss**
  - **Mean Absolute Error/L1 Loss**
  - **Mean Bias Error**
- **Classification Losses**
  - **Hinge Loss/Multi class SVM Loss**
  - **Cross Entropy Loss/Negative Log Likelihood**

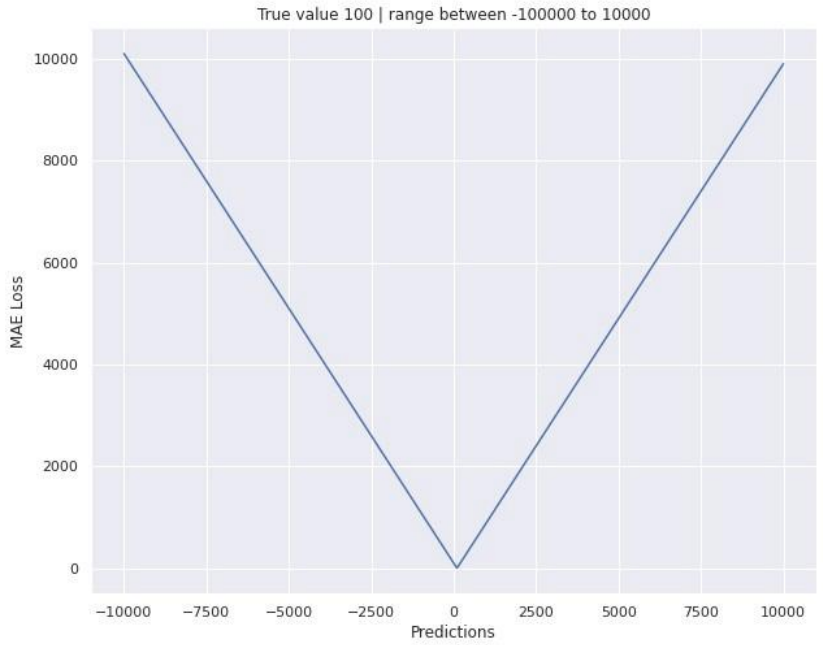
Loss Functions

- **Mean Square Error Loss**
- It’s the square difference between the current output y_pred and the expected output y_true divided by the number of output.
- The MSE function is very sensitive to outliers because the difference is a square that gives more importance to outliers.
- If we had to predict one value for all targets, the prediction should be the **mean**.



Loss Functions

- **Mean Absolute Error Loss**
- At the difference of the previous loss function, the square is replaced by an absolute value. This difference has a big impact on the behavior of the loss function which has a "V" form.
- It's like a median, outliers can't really impact the behavior.

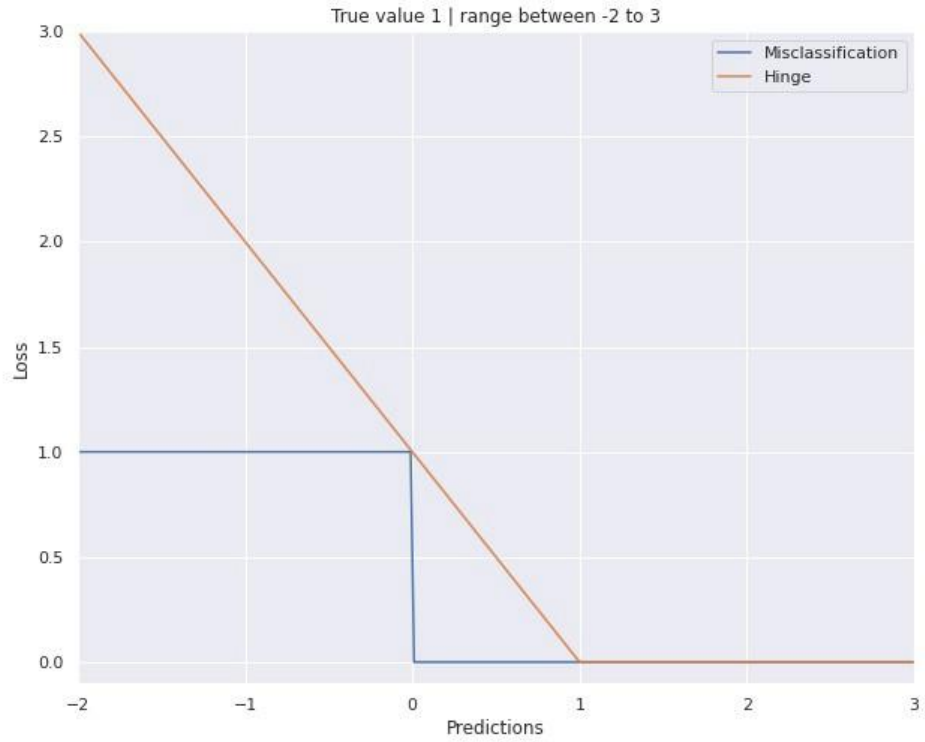


Loss Functions

- **Mean Bias Error**
- This is same as MSE with the only difference that we don't take absolute values.
- **It could determine if the model has positive bias or negative bias.**

Loss Functions

- **Hinge Loss**
- The Hinge loss function was developed to correct the hyperplane of SVM algorithm in the task of classification.
- The goal is to make different penalties at the point that are not correctly predicted or too closed of the hyperplane.



Loss Functions

- **Cross Entropy Loss/Negative Log Likelihood**
- This is the most common setting for classification problems.
- **Cross-entropy loss increases as the predicted probability diverges from the actual label.**

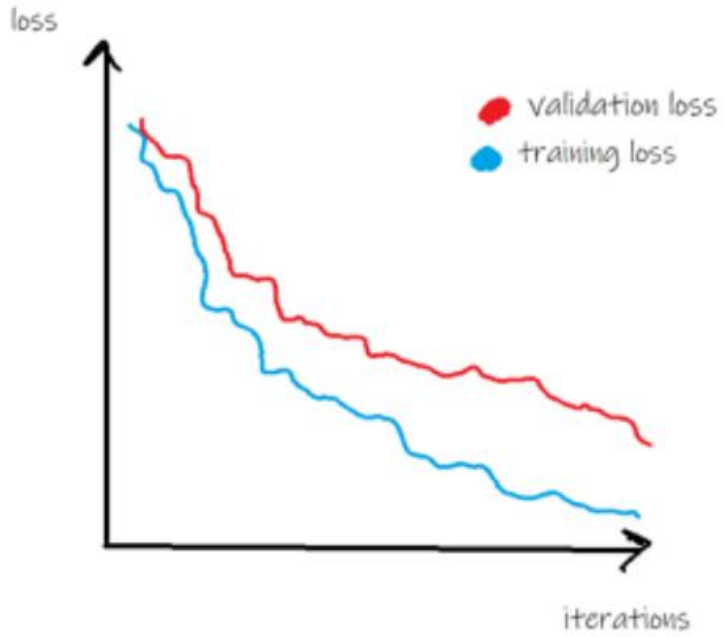
Generalization

- The term 'generalization' refers to a model's ability to adapt and react appropriately to previously unseen, fresh data chosen from the same distribution as the model's initial input.
- In other words, generalization assesses a model's ability to process new data and generate accurate predictions after being trained on a training set.
- Over-training on training data will prevent a model from generalizing. In such cases, when new data is supplied, it will make inaccurate predictions. Even if the model is capable of making accurate predictions based on the training data set, it will be rendered ineffective.
- This is referred to as overfitting.
- The contrary is also true (underfitting), which occurs when a model is trained with insufficient data.

Generalization

- The ultimate goal of machine learning is to find statistical patterns in a training set that generalize to data outside the training set.
- Take the following simple NLP problem: Say you want to predict a word in a sequence given its preceding words.
- For example, the sequence "the cat ___" may be followed by
- ***sleeps*, *enjoys*, or *wants***
- For example, the sequence "the plane ____"
- We would expect the model to predict **departs** with a higher probability than **depart**, and **leaves** with a higher probability than **leave**. In that case, the model has learned a pattern that is more generally true.

Generalization



If validation loss decreases as well, the learned patterns seem to generalize.

- If training loss in fact does decrease as expected, it doesn't automatically mean that whatever the model has learned is also useful. This is where the validation loss comes into play.
- Things look good if the validation loss decreases alongside the training loss. In that case, the learned patterns seem to generalize to the unseen validation data.
- The validation loss will typically be higher than the training loss, however, since not all patterns generalize, as you can see in the following graphic.

Bias

- Bias is defined as the average squared difference between predictions and true values. It's a measure of how good your model fits the data.

- Zero bias would mean that the model captures the true data generating process perfectly. **Both your training and validation loss would go to zero. That is unrealistic**.

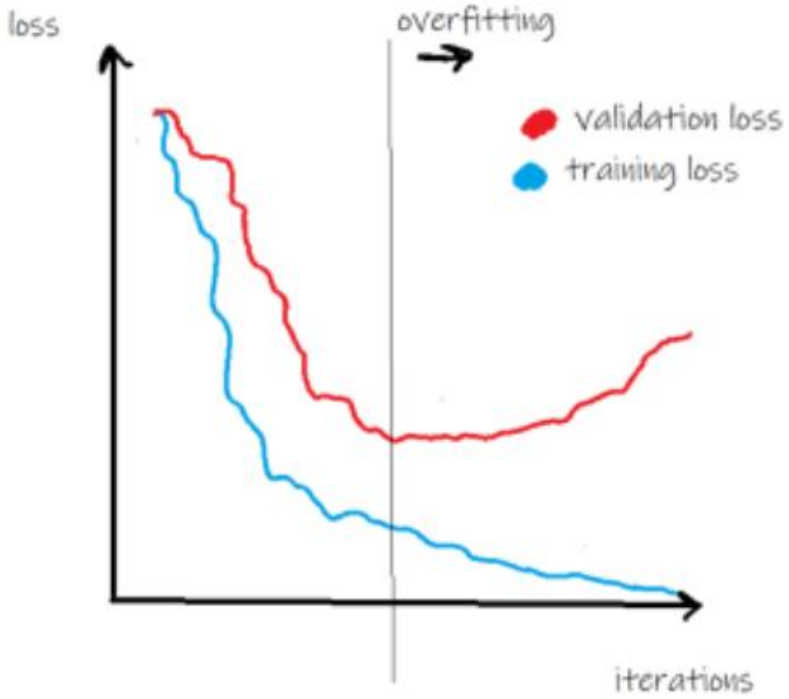
- However, as data is almost always noisy in reality, so some bias is inevitable — called the irreducible error.
- if losses do not decrease as expected, it probably signals that the model is not a good fit for the data.
- **For example, if you tried to fit an exponential relationship with a linear model — it can simply not adequately capture that relationship.**

Variance

- A model is said to have high variance if its predictions are sensitive to small changes in the input.
- High variance often means overfitting because the model seems to have captured random noise or outliers.

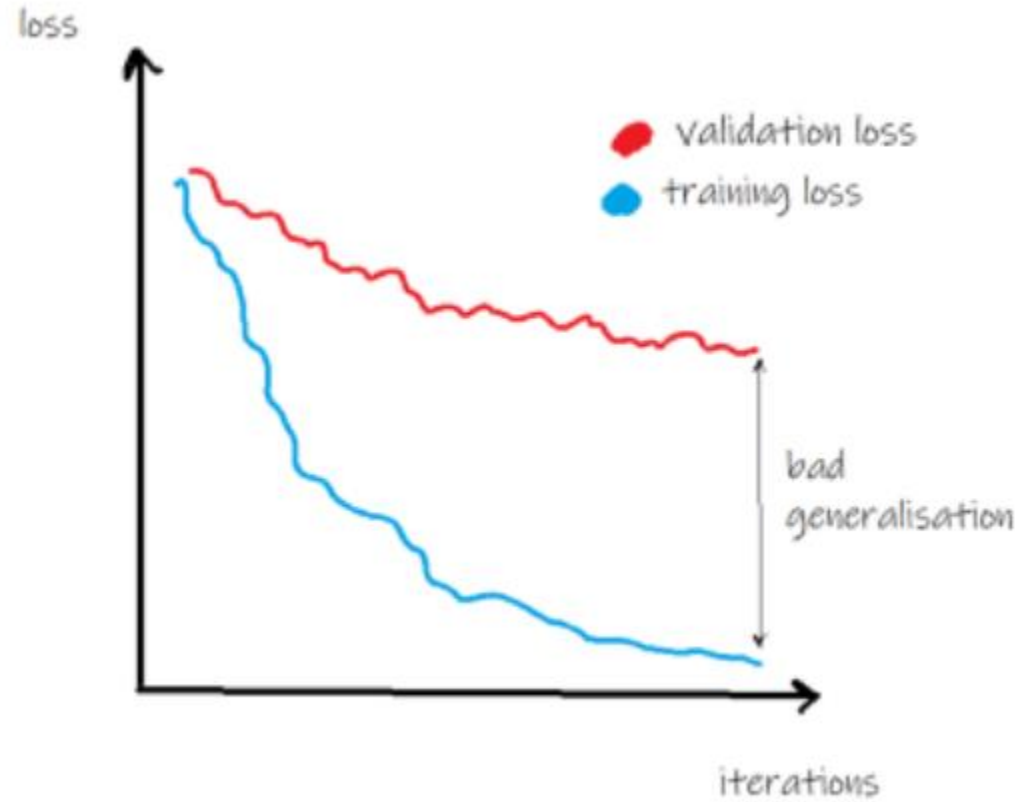
Overfitting

- At some point during the training of a model, the validation loss usually levels out (and sometimes even starts to increase again) while the training loss continues to decrease. That signals overfitting.
- In other words, the model is still learning patterns but they do not generalize beyond the training set (see graphic below).
- Overfitting is particularly typical for models that have a large number of parameters, like deep neural networks.



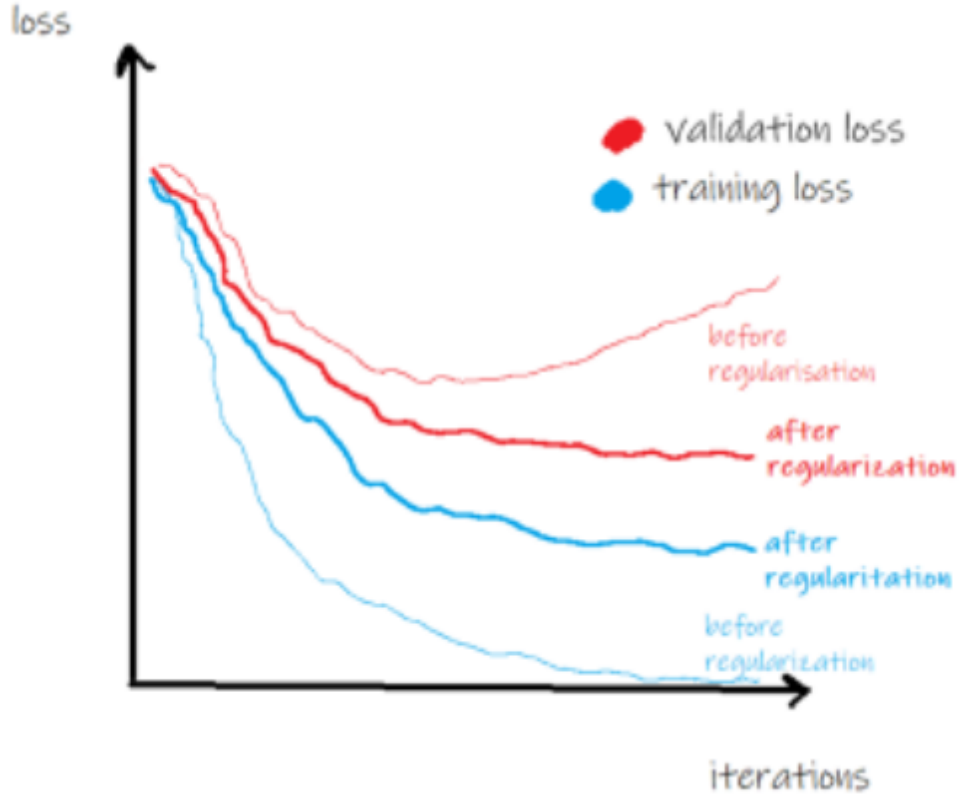
Overfitting

- A large gap between training and validation loss is a hint that the model does not generalize well and you may want to try to narrow that gap.
- That is to stop the train simplest solution to overfitting is early-stopping , that is to stop the training loop as soon as validation loss is beginning to level off.



Regularization

- Regularization is a method to avoid high variance and overfitting as well as to increase generalization.
- Without getting into details, regularization aims to keep coefficients close to zero (shrinking).
- L1 (Lasso) and L2 (Ridge regression) regularization are two widely used methods.



Probability Theory

What is Probability Theory?

- Probability theory describes probabilities in terms of a probability space, typically assigning a value between 0 and 1, known as the probability measure, and a set of outcomes known as the sample space.
- Outcomes are often referred to as the results of an event. Probability theory in general attempts to apply mathematical abstractions of uncertain, also known as non-deterministic, processes.
- The tools that are common in probability theory are discrete and continuous random variables, probability distributions, and stochastic processes.

What is Probability Theory?

- For example: consider that you have two bags, named A and B, each containing 10 red balls and 10 black balls. If you randomly pick up the ball from any bag (without looking in the bag), you surely don't know which ball you're going to pick up. So here is the need of probability where we find how likely you're going to pick up either a black or a red ball.

- **Conditional Probability**
- **P(Red ball)= P(Bag A). P(Red ball | Bag A) + P(Bag B). P(Red ball | Bag B)**, this equation finds the probability of the red ball.

Conditional Probability

- P(Bag A) = 1/2 because we've 2 bags out which we've to select Bag A.
- P(Red ball | Bag A) should read as "probability of drawing a red ball given the bag A" here "given" word specifies the condition which is Bag A in this case, so it is 10 red balls out of 20 balls i.e. 10/20.

- P(Red Ball)= 1/2. 10/20 + 1/2. 10/20 = 1/2

Conditional Probability Formula

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

Probability of A given B

Probability of A and B

Probability of B

What is Probability Theory?

BAYES THEOREM

Likelihood

Class Prior Probability

$$P(c \mid x) = \frac{P(x \mid c)P(c)}{P(x)}$$

Posterior Probability

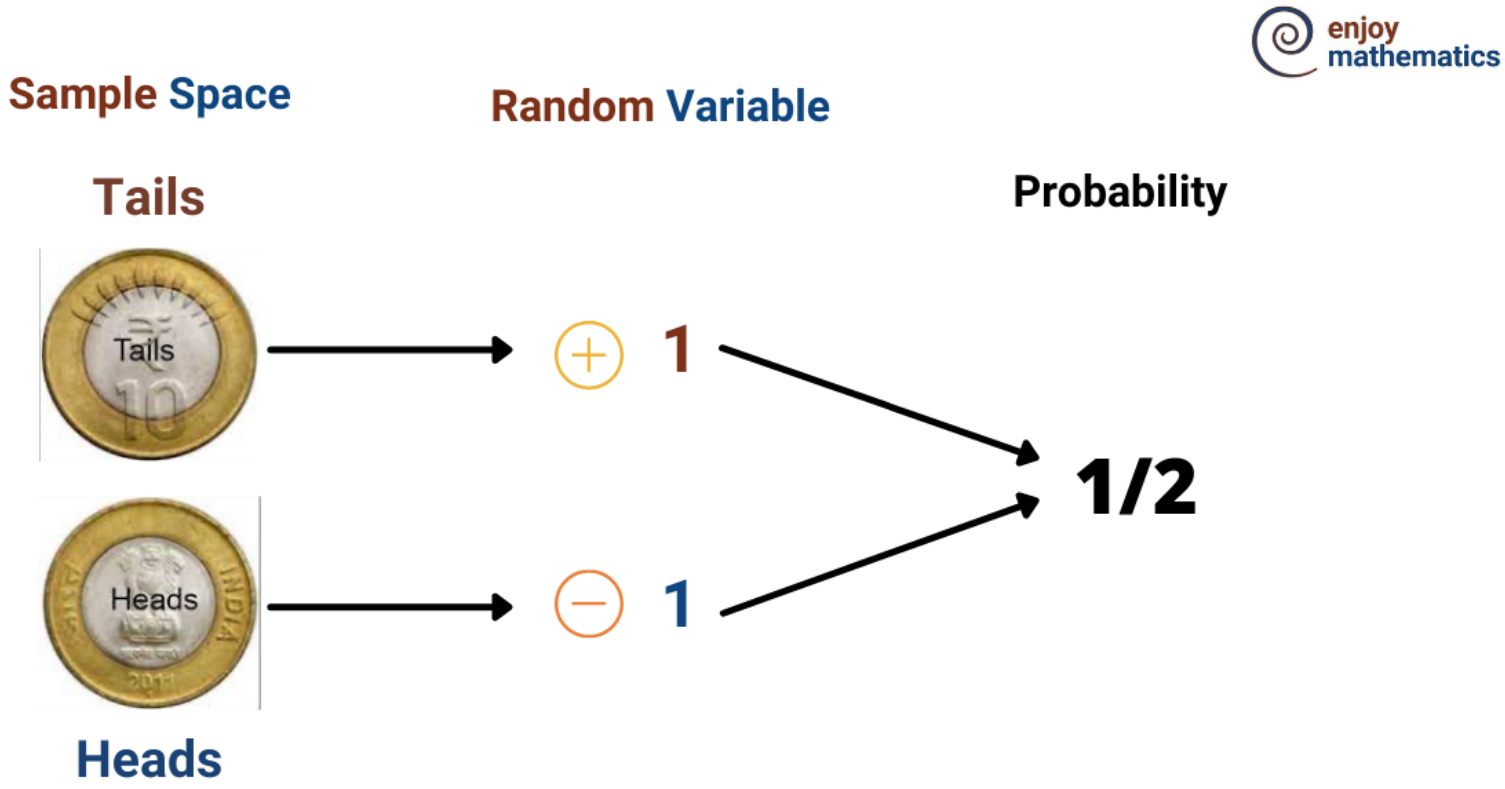
Predictor Prior Probability

$$P(c \mid X) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

BAE'S THEOREM

$$P(chill \mid Netflix) = \frac{P(Netflix \mid chill)P(chill)}{P(Netflix)}$$

What is Probability Theory?



Naïve Bayes Algorithm Demo

https://colab.research.google.com/drive/1qerX_b8boEeV2ZwpyRSyVaq-rolmbKm4?usp=sharing

Parametric vs Nonparametric Methods

Parametric Method

- The basic idea behind the parametric method is that there is a set of fixed parameters that uses to determine a probability model that is used in Machine Learning as well.
- Parametric methods are those methods for which we priory knows that the population is normal, or if not then we can easily approximate it using a normal distribution which is possible by invoking the Central Limit Theorem CLT. Parameters for using the normal distribution is as follows:

- Mean
- Standard Deviation

- Confidence interval – is used with mean, SD, population methods to determine the classification.
- Examples: Logistic Regression, Naïve Bayes Model, etc.

Nonparametric Method

- The basic idea behind the parametric method is no need to make any assumption of parameters for the given population or the population we are studying.
- In fact, the methods don't depend on the population. Here there is no fixed set of parameters are available, and also there is no distribution (normal distribution, etc.) of any kind is available for use.
- This is also the reason that nonparametric methods are also referred to as distribution-free methods.

- Low complexity so easy to apply.
- **Methods:**
- Spearman correlation test
- Sign test for population means
- U-test for two independent means

- Examples: KNN, Decision Tree Model, etc.

Elements of Computational Learning

Computational Learning

- Difficulty in ML algorithms.
- Capabilities of ML algorithms.
- Under what conditions a successful learning is possible/not possible?

- We focus here on the problem of inductively learning an unknown target function, given only training examples of the target T and a space of candidate hypothesis.

- How many training examples are sufficient to learn the target T. (Sample complexity)
- How much computational effort is needed for a learner (Computational complexity)
- How many mistakes it makes before it finds the correct hypothesis. (Mistake bound)

Computational Learning

- Historically, support vector machines have largely been motivated and analysed using a theoretical framework known as computational learning theory, also sometimes called **statistical learning theory.**

- This has its origins with Valiant (1984) who formulated the probably approximately correct, or PAC, learning framework.
- **The goal of the PAC framework is to understand how large a data set needs to be in order to give good generalization.**
- It also gives bounds for the **computational cost** of learning, although we do not consider these here.

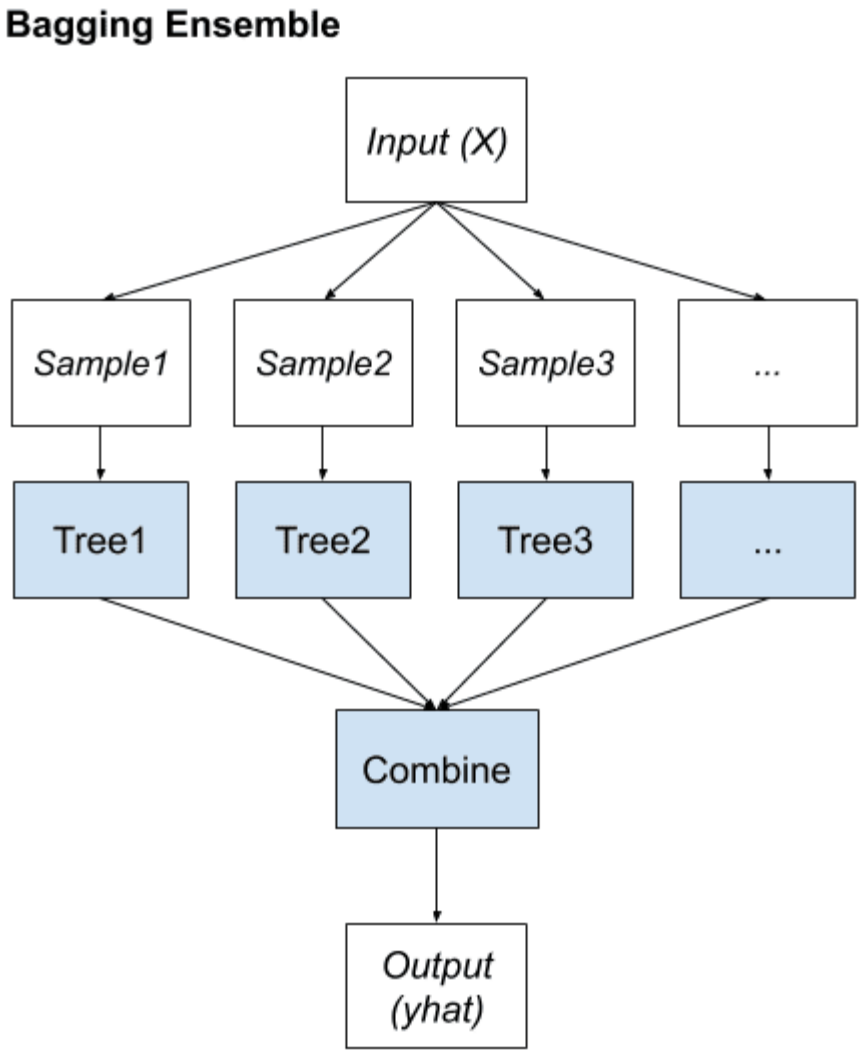
Theory of Ensemble Learning

Ensemble Learning

- Ensemble modeling is a powerful way to improve the performance of your model.
- **Ensemble learning** is a general meta approach to machine learning that seeks better predictive performance by combining the predictions from multiple models.
- The three main classes of ensemble learning methods are bagging, stacking, and boosting, and it is important to both have a detailed understanding of each method and to consider them on your predictive modeling project.

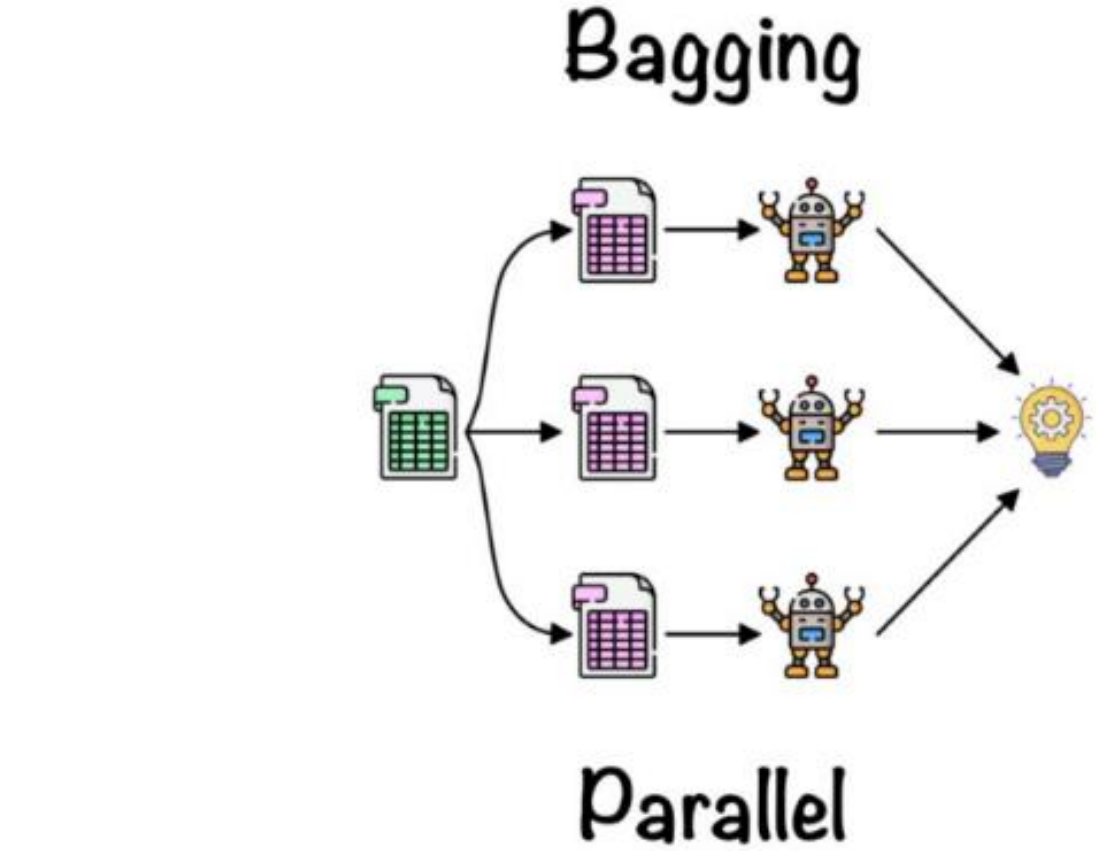
Bagging

- Bagging involves fitting many decision trees on different samples of the same dataset and averaging the predictions.



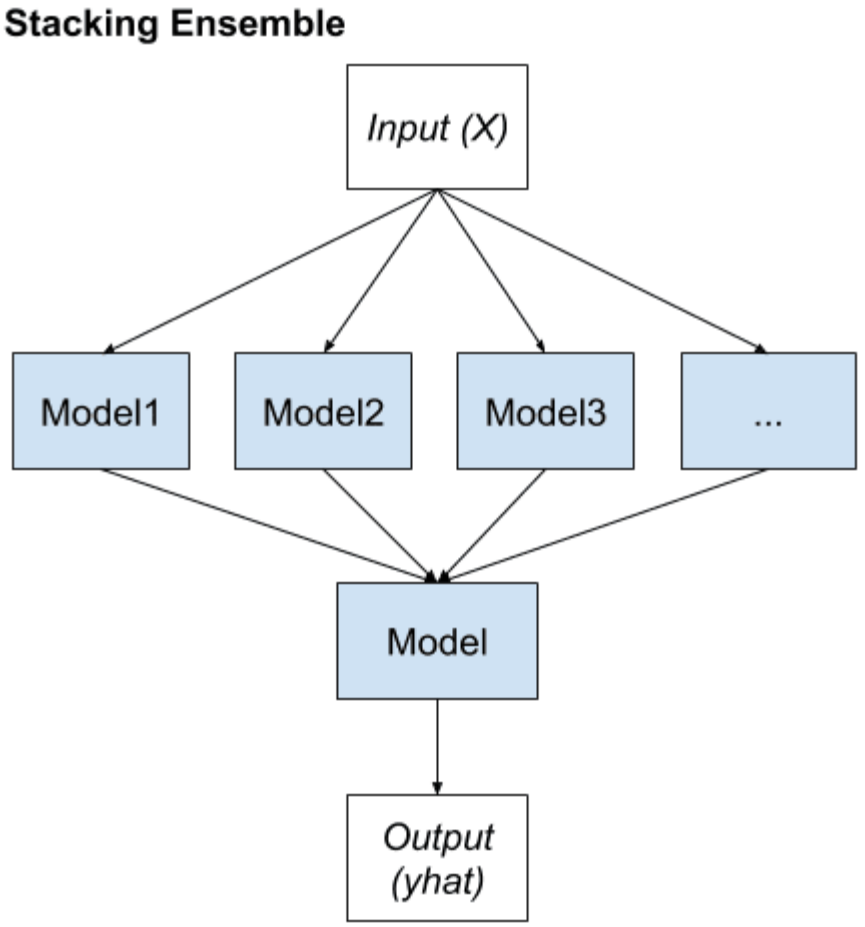
Bagging

- It creates a different training subset from sample training data with replacement & the final output is based on majority voting. For example, Random Forest.



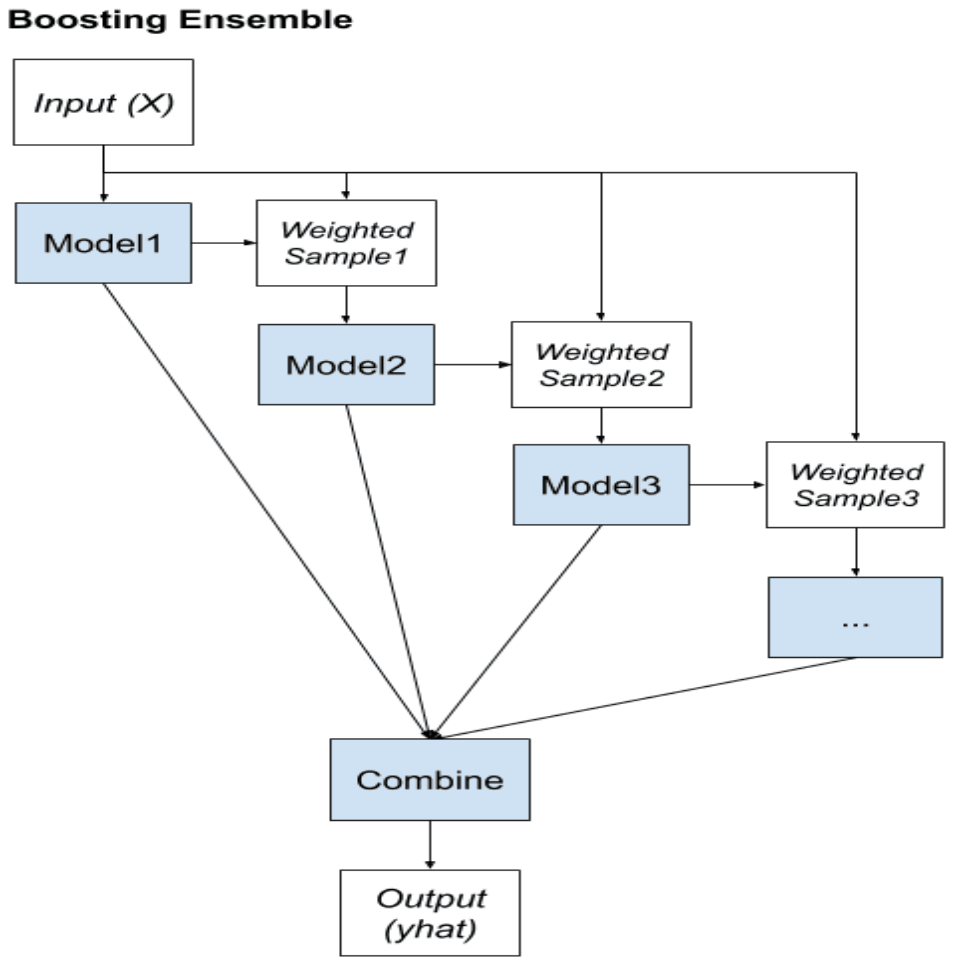
Stacking

- Stacking involves fitting many different models types on the same data and using another model to learn how to best combine the predictions.



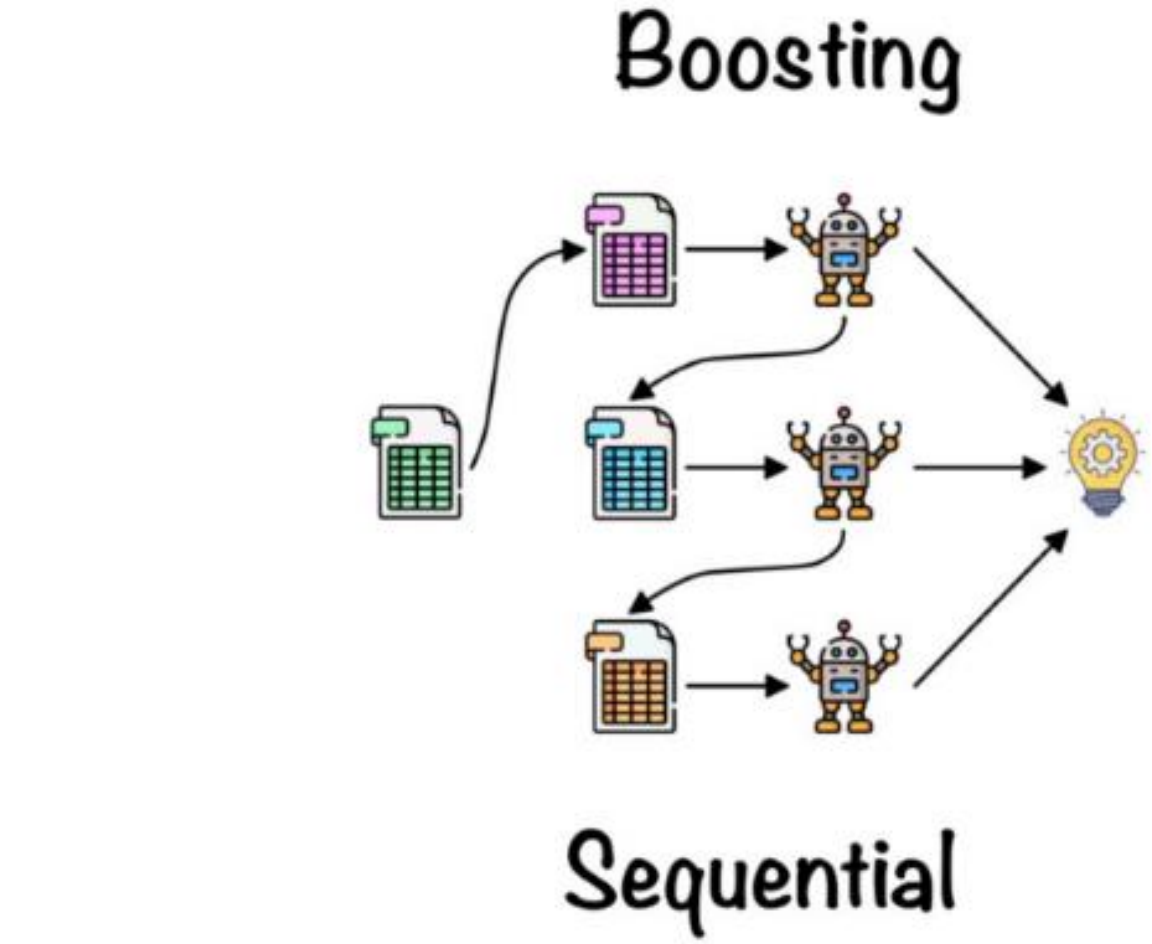
Boosting

- Boosting involves adding ensemble members sequentially that correct the predictions made by prior models and outputs a weighted average of the predictions.



Boosting

- It combines weak learners into strong learners by creating sequential models such that the final model has the highest accuracy. For example, ADA BOOST, XG BOOST



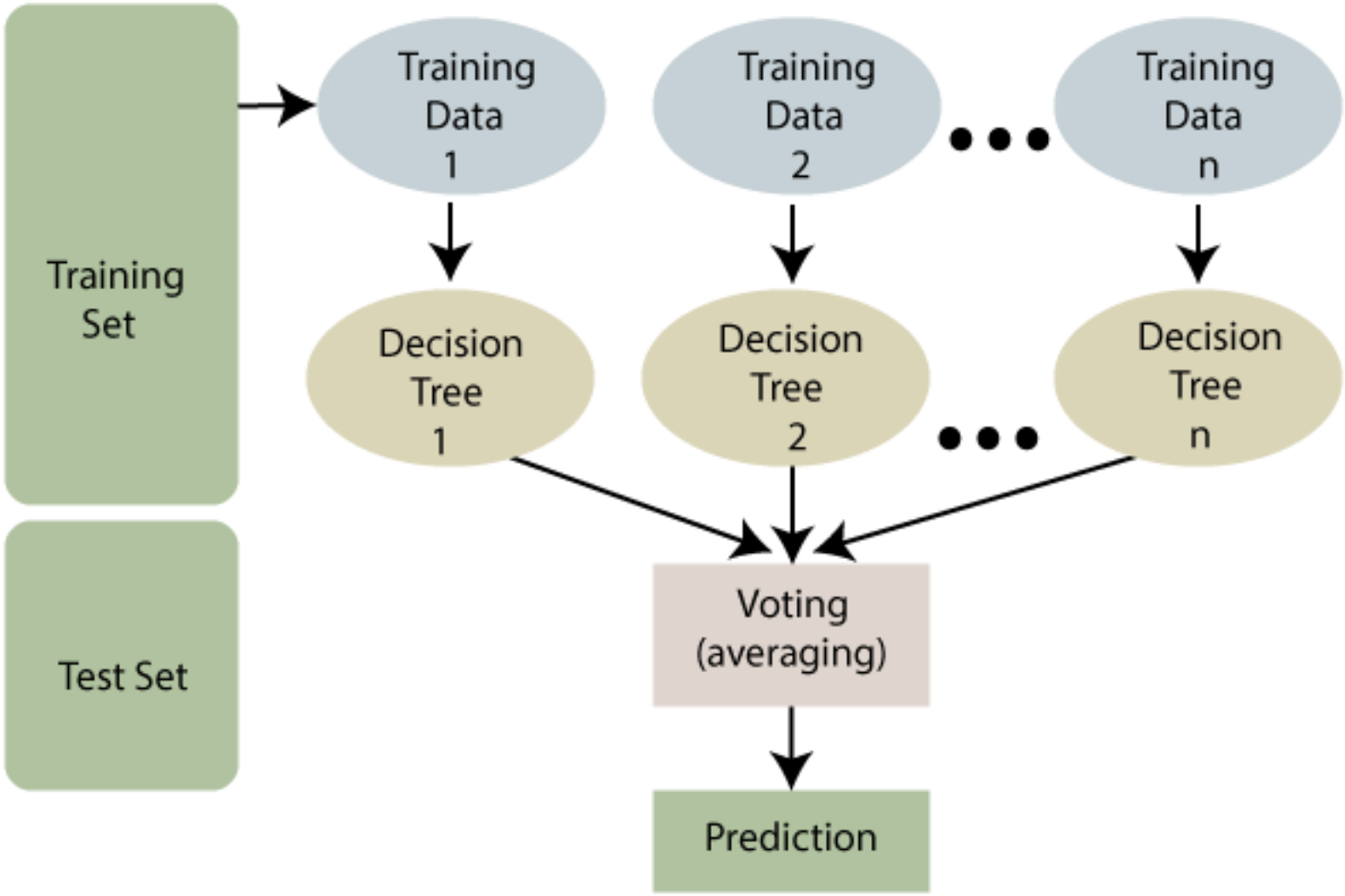
Random Forest

Random Forest

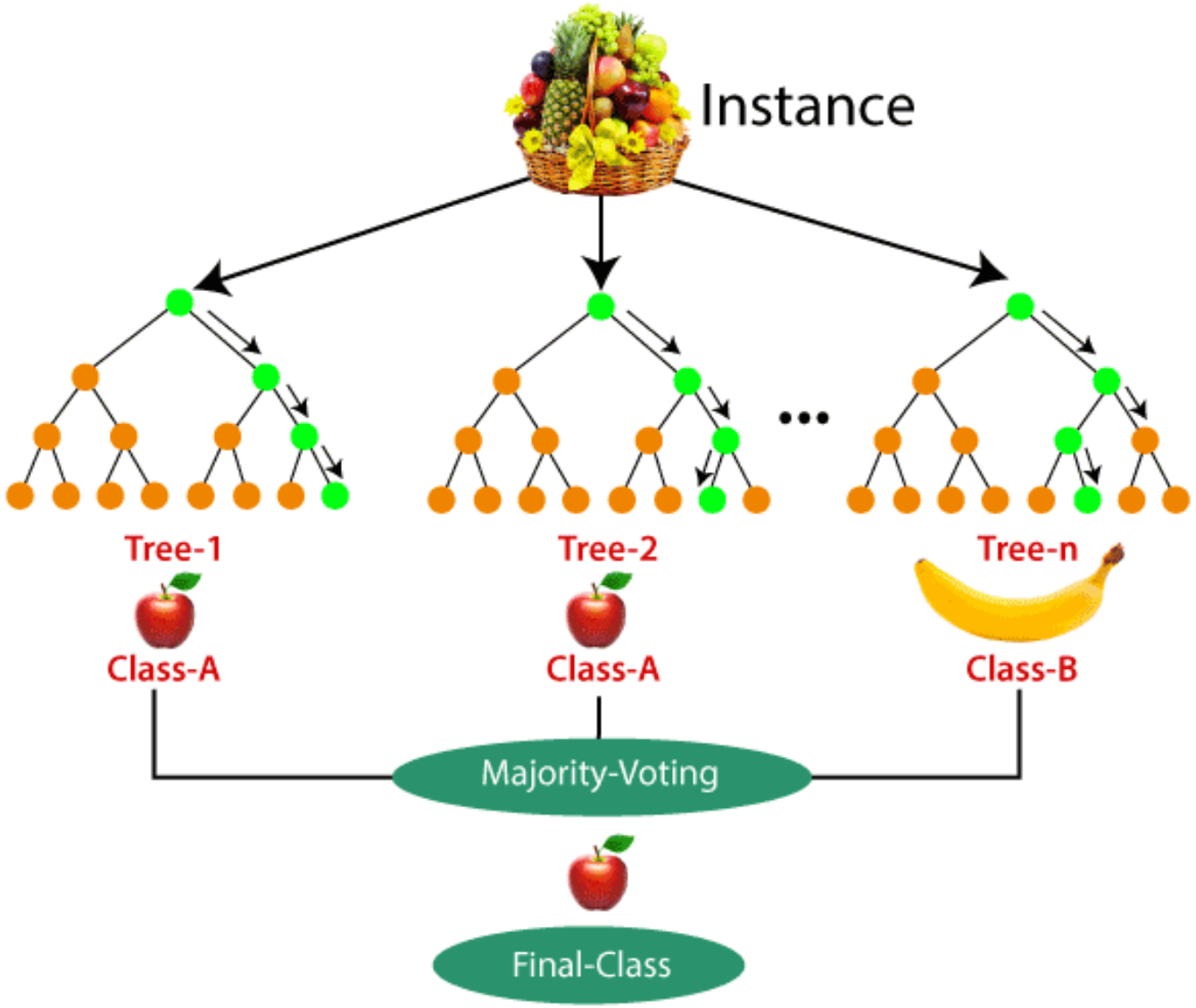
- Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique.
- It can be used for both Classification and Regression problems in ML.
- It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model.

- **Advantages:**
- It takes less training time as compared to other algorithms.
- It predicts output with high accuracy, even for the large dataset it runs efficiently.
- It can also maintain accuracy when a large proportion of data is missing.

Random Forest



Random Forest



Random Forest Demo

https://colab.research.google.com/drive/1SMaro_y73SC4WW7ZNUqIxDRw8lcUGmlO?usp=sharing

Research Papers References

1. **Supervised Learning**
   Data-analytics-for-the-identification-of-fake-reviews-using-supervised-learning
   https://drive.google.com/file/d/11F0360HJ4Zvzig4pegV6_wA7g4Tgjrr6/view?usp=sharing
   Deep Transfer Learning Based Classification Model for COVID-19 Disease
   https://drive.google.com/file/d/1UQyPkNJZt8Af22NUHhsUO7Ztag_61vWO/view?usp=sharing
   Predictive-modelling-and-analytics-for-diabetes-using-a-machine-learning-approachApplied-Computing-and-Informatics
   https://drive.google.com/file/d/1E9Zqvh0kM23OzILbrWQlMn3X3XWr_E0B/view?usp=sharing
2. **Unsupervised Learning**
   A Classification Algorithm-Based Hybrid Diabetes Prediction Model
   https://drive.google.com/file/d/159Q2CKlBF7wELcNlZOGJIiD-yPIqdsZt/view?usp=sharing
3. **Reinforcement Learning**
   Optimal-deep-reinforcement-learning-for-intrusion-detection-in-UAVs
   https://drive.google.com/file/d/1NOxQ636g94HkF2SDGi7VoGz14lV5Lixc/view?usp=sharing
   Multiagent-deep-reinforcement-learning-a-surveyArtificial-Intelligence-Review
   https://drive.google.com/file/d/1NrKbU-BbwiDKY5b6sqYlCRsxIAZk4a07/view?usp=sharing

Up Next

- Kernel Methods for Non-linear Data
- Support Vector Machines
- Kernel Ridge Regression
- Structure Kernels
- Kernel PCA
- Latent Semantic Analysis

End of Unit 1

Thank You.