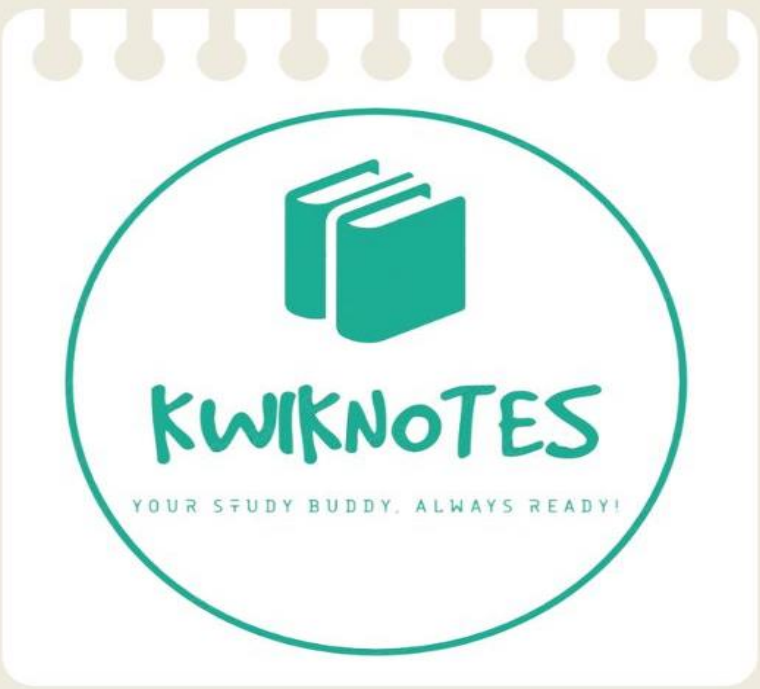


HANDWRITTEN NOTES

Download FREE Notes for Computer Science and related resources only at

Kwiknotes.in

Don't forget to check out our social media handles, do share with your friends.

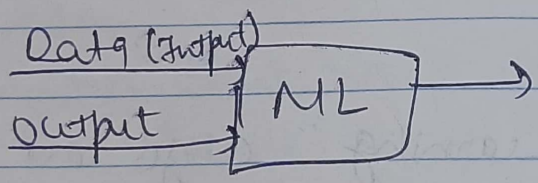
improve

ML

It is a subfield of AI that focus on developing algo & models that enables machines/ computers to learn themselves from data and make predictions without being explicitely programmed.

Arthur Samuel used term ML in 1956

ML algo create a mathematical model that helps to make decision with assitance of previous/ historical data



Working

ML works by training algorithm to learn patterns from data, make decision on those pattern & improve performace. It begin with data collection, preprocessing followed by selecting an appropriate ml algo The model is then trained on given dataset, adjusting its internal parameters to make accurate prediction. Once validated, it can be deployed for practical ~~implementation~~ apps from image recognition to reccommendation system, enabling automated decision making

Features

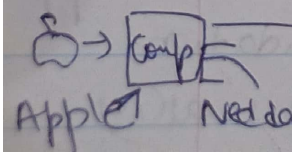
- Detect various patterns from data
- learn from past exp & improve
- data driven tech
- Similar to data mining
- Personalization exp on platform (Netflix)
- Inp
- Solving Complex problems
- find pattern in dataset, make predictio
- Rapid increment in prod of data
- finding hidden patterns.
- Image & speech Recognition in Automobile

Types

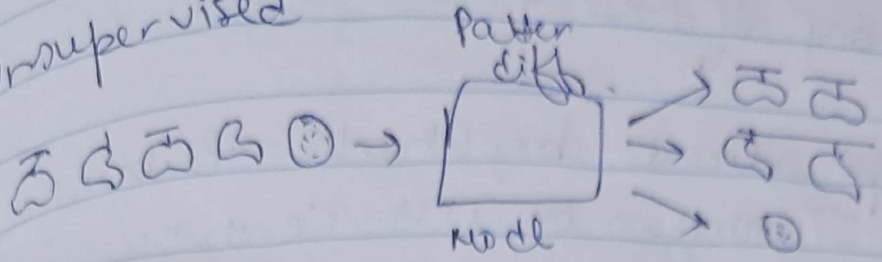
1) Supervised Learning: algo are trained on labelled dataset where each data point has input feature & corresponding target labels.
Sample data are provided to machine for training & system then predict outpu

After training, we test model with sample data to see if it accurately predict or not

Classification & Regression] two algo in supervised.



2) Unsupervised



Type of ml where machine learns without supervision. Training is provided to machine with the set of data that has not be labelled, classified, categorized, & algo need to act on unsupervised data

Goal is to restructure input data into new feature or group of object with similar pattern.

In this we dont have predetermined result. Machine tries to find useful insights from large amount of data.

Two algo:
Clustering, Association.

3) Reinforcement learning
is a feedback based learning method in which a learning agent gets a reward for each right actions & penalty for wrong action. This improves with performance. These system interact ~~&~~ with env & explors it. Aim is to get rewards & hence inprove performace. Ex- Robot.

History of ML

Conceptualized by Charles Babbage & Alan Turing in 21st Century
ML experianced challenges during AI winter but emerged stronger with neural n/w & reinforcement learning
Notable milestone include IBM Deep Blue defeating a chess champion

Supervised

Ad 1) Predict things accurately when they have clear ex to learn from
2) Easy to understand
3) Capable to solve complex prob

Is Need lot of labelled data
Only predict things seen before

Unsupervised learning
Find hidden patterns
Dont need labelled data
Simplify new / complex data

Dis No specific goal
No exact predictions

Reinforcement

Ad Good for tasks with decision making
Get Better with experiance

Dis Takes more compute power & time to train
May be slow in learning new things

Clustering → used in unsupervised learning, to group & data points on basis of hidden pattern & similarity without predefined labels.

Classification (Supervised learning)
sorting things into categories / groups
used when we predict from which category something belong

Regression (Supervised)
Predicting quantity or number like price, temperature

(Unsupervised)
Association - finding patterns when Clustering things.

Page No.
DATE: / /

Semi Supervised Learning
Type of ML with both qualities of supervised & unsupervised.
In this you have to access data with both labelled & unlabelled dataset.

This aim to leverage benefit of labelled data along with making use of larger pool of unlabelled data.

	Supervised	Unsupervised
→	Includes input & output labels	Include only input feature, no labels
→	Predict on base of input feature	Discover pattern by grouping
	Uses classification, Regression	Uses clustering, Association
	Feedback is direct	No Feedback
	Predict output	find pattern

Scikit

Also known as SKlearn is most useful & robust library for machine learning in python

Involves selection of efficient tools & techniques for ml and statistical modelling including classification, regression, clusturing, etc

→ This library is written in python & built upon Numpy, Scipy & Matplotlib
→ Opensource
→ Used for data mining & analysis.

Features

→ Provide consistent & straightforward API
→ Rich set of Algos
→ Data preprocessing → handelling missing value, encoding, etc
→ Predefined dataset

Dataset is a collection of data organised in a structured way & stored in Memory for processing, analysis, etc.
In Scikit learn some dataset are

already provided to easily access data

Iris dataset → measures of sepal & petal length
digit dataset → from 0 to 9
Wine, Breast Cancer, Diabities dataset

These can be used in python script & Jupyter Notebook using datasets.load_* where * is name of dataset

data.load_iris()

```
from sklearn import datasets
iris = data.load_iris()
x = iris.data
y = iris.target
```

Applications of Scikit

Classification → prediction on input
Clustering
NLP
Anomaly Detection.

Problems ML can Solve

1) Image Classification → used to classify diff category of images ex - dog or cat

2) Natural language Processing : ML can understand human lang & take decision

3) Recommendation system → ML can suggest movies, songs, etc.

4) Regression Analysis : ML can be used to analyse numbers ex - prices

5) Healthcare - imp in personalized treatment

Classification of ML Techniques

- Supervised
  Input data paired with correct output, prediction of base of input data
  Algo - Linear Regression, Decision Tree

- Unsupervised
  And paterns / structures
  Algo - hierarchical clustering, K-means clustering

- Semi Supervised
  Mixture of supervised & unsupervised

- Use small labelled data & large amt of unlabelled data.
- Expensive, time consuming

Reinforcement Learning
- Agents make decision acc to environment
- Trial & error used
- Reward for correct & penalties for wrong
Algo - DQN, Q-learning

Deep Learning
Subset of ML that focus on deep neural n/w with multiple layers.
Used for speech & image Recognition
Algo - RNN, CNN

ML Lifecycle

1) Data Collection
   Gather Relevant Data
2) Data Preprocessing
   Clean and prepare data.
3) Feature Engineering
   Choose Relevant feature to Represent data.
4) Model Selection
   App ML algo Selection

5) Model Training
Train algo on dataset

6) Model Evaluation
Access performance

7) Model Deployment
For real world use

8) Monitor & Maintenance
Update when needed

Types Of Dataset

Numerical, Text, Image, Time, Geospatial (location), Categorical (gender)

DATA PREPROCESSING

It is a critical step in ML lifecycle that involve cleaning, transforming & preparing data for training & testing machine models.

Tasks included are data cleaning, (outlier analysis) handelling missing value, feature Scaling, encoding categorical data, Spliting & testing

It is for successful machine learning modelling.

1) Outlier Analysis

Outlier are datapoints that are significantly diff from majority of data. They are like exceptions

They can affect accuracy of analysis or prediction

Ex- A list of score with all marks between 60-90 ✓ then some score comes with 20-30, these are outlier

- we can Remove them
- Analyze seperately
- Transform them.

Techniques
- Visual Inspection (bar, plot, etc)
- Statistical Method (Z-score)
- Domain Knowledge (to determine)

2) Treating Missing Value

Missing values can lead to biased & inaccurate results while traing a model
Ex of missing data - in a survey people might forget to ans some ques

Treating
1) Deletion - Remove RONS / cols with missing value (not recommended)
2) Imputation → fill missing value with mean, median, mode, etc
3) Advanced Technologies - Use algo to predict missing value ex- decision tree, K-nearest neighbour

3) Encoding Categorical Data.
ML work with data sets such as numbers, categorical data etc which need to be encoded to numerical form to be understood by machine

In ml we have blue or red for color, dog or cat for animal. Machine dont understant these words instead convert them to Numbers to process.

Techniques.
Label Encoding - Assign unique no. to each category

One-hot encoding - binary cols for each category

Ordinal encoding - numerical value to each category acc to order/importance

4) Splitting Dataset into Training/Test Dataset

To evaluate performance of model, it is imp to split dataset into two parts

70%, 80% or 1) Training Dataset → train model
30%, 20% 2) Test Dataset → evaluate model

Example to teach robot to identify cat.

x_train, x_test

5) Feature Scaling
Process of standardizing/normalizing feature in dataset to ensure they have consistent scale

It is like converting all dataset to same unit/scale.

Ex- height is in centimer & weight in kg may confuse machines.

Technique
Standerdization
Min-Max Scaling b/w (0 & 1)

Feature Selection Technique

Feature Selection is process of choosing a subset of relevant feature (var/column) from a larger ~~dataset~~ of features in dataset.
Goal is to improve model performance, reduce complexity.

1) Filter Method
These are like detectives that look at each fature (column) one by one in datastt.

* They decide if a feature is important without asking ML model.
Done by simple maths & statistics.

Ex- if in a subset A is related to task, it is choosen.

2) Wrapper Method
These are like detectives who use special ML model to decide if feature is important or not.

They create list of feature, & suspect & find best for solving problem.

3) Embedded Method
are like detectives who are so smart that they learn while working.

They use ML models to figure out which feature is important while solving the problem.

Effient as they perform feature selection while building model.

Supervised Learning Techniques

make prediction based on labelled data set without explicit programming

Techniques

1→ Linear Regression
→ Assumes a linear relationship between input & output

Objective - to find best fitting line that minimize sum of squared difference between predicted & actual value of variable

Used in economics, social science, busines to find cause-effect Relationship & practical applications.

Ex- prediction on base of score & time of study.

2→ Logistic Regression
Type of Regression to predict whether something will happen or not.
Ex- Is this spam email.
It uses a special curve instead of straight line called sigmoid curve

to make prediction b/w 0 & 1

Goal is to use binary classification tasks to find 2 possible outcome yes or no / 0 or 1.

3→ KNN K-Nearest Neighbour
It is supervised ML algo used for classification & Regression.

It is non parametric, instance based algo that make prediction on based similarity of input points to other points in training dataset

K represent no. of nearest neighbour to be considered while making prediction

Make prediction by asking closest neighbour what is happening & get the idea

Ex - If we want to find color of a ball & all nearest ones are blue then most likely it is also a blue ball.

4) SVM Support vector Machine

Powerful classification algo
Identifies support vectors, which are data points closest to decision boundary (hyperplane) that seperate diff classes.

Can handle both linear & non linear data

finds best hyperplane that seperate classes
Aim to
Maximize margins b/w classes

5) Naive Bayes Algo
Based on Bayes Theorm & is a probabilistic approach.

Particularly useful for text classification & spam detection.

Assumes that features are conditionally independent

It is like making quick guess on simple rules
ex- pdf se dekhle.
If 'money' & 'buy' apper in 90% then
it may be spam mails.

Called naive as they are independent.

6) Decision Tree

ML algo that recursively divides an dataset into subset. Create a tree like structure where each internal node represent a decision based on feature values.

Tree nodes leads to answers.

Model Evaluation

When we make models like decision tree, evalution is needed to see how good they worked

Calculate performance & effectiveness of trained Model.

Measure how well model is making prediction

Important Measures

1) Accuracy - percentage of correct ans on test. 90% accurate means 90% of time it is right

2) Recall: No. Of subset we got from all data. High recall means we found most matched.

Precision : it is like how accurate you are at finding subsets from all data.

F1 score - It is balance b/w Recall precision. It tells how good we were are picking subset.

CONFUSION MATRICES

Table used in ML & stats to evaluate the performance of classification Model particularly in binary classification.

Help to understand how well model make prediction by comparing prediction to actual outcome.

A typical Confusion Matrix has 4 values.

1) True Positive (TP) : that were correctly predicted as true by model
2) True Negative : correctly predicted as negative (Type 1 error)
3) false positive : in correctly predicted as positive by model . Orginally were neg
4) false Neg : incorrectly predicted as neg but were actually positive . (Type 2 error)

Simply Confusion matrix is like a chart that help to understand how good/bad a comp model is at making prediction/ decision.

Used when comp say yes/no to something

Implementation of Learning Technique & determination of performance

1) Collecting data
2) Getting data Ready
3) Choose Technique
4) Train Computer
5) Evaluate performance
6) Make it Better
7) Use in Real life
8) Keep Eye
9) Document & Sharing