Machine learning

way for computers to learn from data without being explicitly programmed to complete a task.

Data : ML relies on data - examples or info the model learns from. Data can be numbers, text, images etc.

Features : These are individual measurable properties or characteristics of data.
eg. for a model that predicts house prices, features could include no. of rooms, location, square footage etc.

Labels : These are the actual answers or outcomes that the model tries to predict
In supervised learning
In supervised learning, the model learns from labelled data.

Training & testing :

training → model learns pattern from a dataset called the training set.

testing → model's accuracy is evaluated using a seperate testing set to see how well it perfor

Model → this is the str. or algorithm that makes prediction or decisions based on the data it's trained on. There are many types of models like linear regression, decision trees, neural networks.

Supervised learning

When model learns from labelled data. Tasks include classification (like identifying spam emails) & regression (predicting prices).

Unsupervised learning

model finds pattern in unlabelled data. Common tasks are clustering (grouping similar item) & dimensionality reduction (simplifying data)

Overfitting : when model learns a data too well, it struggles with new data because its too specific.

Underfitting : when model is too simple, it does not capture enough details & perform poorly on both

training & new data.

Cross validation :

a method for assessing how the model generalises to new data by splitting data into parts & training on some & validating on others.

Neural networks : Inspired by human brain, these algorithms designed to recognize patterns. They're used in deep learning & have layers that process data to capture complex patterns especially in tasks like image & speech recognition.

Hyperparameters

These are settings for a model that are set before training & influence how well the model learns. eg. how quickly the model adjusts is a hyperparameter.

I Data is the foundation of ML

Types of data

1.) Structured data :
This is organized & formatted in a way that is easily readable by machines, often in tabular form.
eg. Numerical data (continuous values like age, height etc.)
Categorical data (Discrete values representing categories like colors or types of animals)

2.) Unstructured data :
type of data does not have predefined format, making it more complex to analyse.
eg. Text, Images, Audio/video

3.) Semi Structured data :
This data does not fit neatly into a table but has some organizational properties like tags or markers.
eg. JSON files/ XML files/ HTML docs

data should be accurate, complete, consistent & up to date.

Data Prep

Before using data in ML, it often needs to be cleaned & prepared.

1.) **Data Cleaning** :

a) **Handling missing values** : You can remove data points with missing values or use other imputation techniques.

b) **Removing duplicates** : Check for & remove any duplicate entries

c.) **Correcting errors** : Look for & fix inaccuracies in the data like typos or incorrect entries.

2.) **Data Transformation**

a) **Normalization / Standardization** : Scaling numerical data to ensure all features contribute equally to the analysis.

b.) **Encoding categorical variables** :

Converting categorical data into Numerical form using techniques like one-hot encoding or label encoding.

3) Data Splitting :

dataset usually divided into 2 sets.

a) Training set : used to train data.

b) Testing set : used to evaluate model's performance on unseen data.

IV) Features are the individual measurable properties / characteristics of the data that a model uses to make prediction.

Each feature represents a piece of info that could help the model recognize patterns in the data.

Good features improve the model's ability to make accurate predictions, & vice versa.

Types of features :

a.) Numerical : can be measured on a scale

eg. age, height, weight

Slide 3 — Constraint Programming

b) Categorical - Discrete values that represent categories like gender, marital status, color etc.
These are often encoded into numbers for the model to process.

c) Binary - Categorical features often represented using 0/1.

d) Ordinal - categorical features with meaningful order or ranking like educational level.

~~Feat~~
Feature Engineering is process of selecting, modifying or creating features to improve performance of a ML model.

a) Feature selection :
selecting most relevant features & removing irrelevant or redundant ones.
To decide which features to keep, several techniques can be used :
(i) Filter methods : use statistical features, such as correlation with the target variable.

(ii) Wrapper methods : Test combination of features by running the model multiple times, can be time consuming but effective.

(iii) Embedded methods : Select features while building the model, using algorithms that automatically determine feature importance. eg. decision trees.

b) Feature Creation :
Sometimes new features can be created by combining or transforming existing ones.
eg.
- Polynomials : For a model predicting sales, creating a feature that is square of 'ad spend' might help capture a non linear relationship

- Interactions : combining 2 features like "age" & "income" to create "age : income" ratio.

c.) Feature Transformation :

- Normalization / Standardization :
scaling each v feature to make sure numerical each has a similar range/ distⁿ. This is imp. for neural networks or k- nearest neighbours.

- Encoding categorical data :
Converting categories to numerical values using methods like one hot encoding (each category becomes its own feature) or label encoding (assigning each category a unique no.)

d.) Dealing with missing or inconsistent data :

Missing values can be filled using techniques like mean imputation or features can be flagged to indicate missing data.

III Labels in ML

A Label is the actual outcome or answer for each piece of data in a supervised learning problem.
Labels are what the model is trying to predict or classify.

When training, model learns from this ~~data~~ labels to recognize patterns & make predictions on new unseen data.

Types of labels

a) **Binary Labels :**

Used in problem with only 2 possible outcomes.
like "Yes/ No"
eg. In a spam detection model, the label would be 'spam' or 'no spam'

b) **Multi Class Label :**

Used when multiple categories to choose from.
eg. image classification to identify animal
Could be Cat, dog etc.

c) Continuous Labels:

Used in regression task when label is continuous value
eg. for model predicting house prices, label would be actual price like $250,000.

* In supervised learning → labels are available & model learns by matching features to labels

In unsupervised learning → No labels are provided, instead, the model tries to find patterns or clusters within the data without specific guidance.

V Model in ML:

① Linear models:

(i) Linear regression:

predicts a continuous outcome based on a linear relationship b/w features & the target.
eg. predicting house p- based on sq. footage.

(ii) Logistic Regression :

Used for binary classification, predicting the probability of a binary outcome.

eg. predicting whether a customer will buy a product.

(2) Tree based models :

(i) Decision Trees : These models split data based on features to make decisions

eg. Classifying animals based on characteristics (eg. does it fly? Is it furry?)

(ii) Random Forests :

An ensemble of decision trees where multiple trees work together to improve accuracy & reduce overfitting.

(iii) Gradient boosting mechanisms :

A method that builds multiple decision trees in sequence, focusing on improving areas where the previous trees struggled.

(3) Nearest Neighbour Models :

a) k - Nearest Neighbours :

A simple algo that classifies new data based on the closest points in the training data.
eg. Classifying an unknown fruit based on similar fruits in dataset

(4) Support Vector Machines (SVM) :

A model that finds optimal boundary that best seperates diff. classes. It works well in high dimensional spaces & is used for both classification & regression.

(5) Neural Networks : inspired by human brain, neural networks consist of layers of interconnected nodes. These models are powerful for ~~both classification &~~ complex tasks like image & speech recognition.

(6) Clustering Models :

a) k - Means clustering : An unsupervised model that groups similar data points together into clusters.

b) Hierarchical clustering : Builds a hierarchy of clusters that can be visualised as a

(4) Probabilistic Models:

a) Naive Bayes : probabilistic model often used in text classification, like spam detection.
It assumes features are independent making it fast & effective for certain tasks.

* Loss fn : Calculates the diff b/w model's prediction & the actual labels. The model adjusts to minimize the loss.
eg Mean squared Error, Cross entropy loss.

* Optimization Algorithm :

This is the method, the model uses to minimize the loss fn. The most common optimization algorithm is Gradient Descent where the model iteratively adjusts the parameters to reduce errors.

VI

Supervised Learning	Unsupervised Learning
model is given a set of labelled data (input features & output labels) During training, the model adjusts its parameters to minimize the diff. b/w its predictions & actual labels. once trained, model can predict the label for new data based on patterns it learned during training	model is given a dataset with only input features & no labels. The model explores the data to find patterns, clusters or associations within it. The output is often in the form of groupings, clusters, or lower dimensional representations rather than specific labels.
Input : Labeled data (features + labels)	Unlabeled data (features only)
Goal : Predict labels or outcomes	Find patterns
Common Tasks : Classification Regression	Clustering, association

eg.	Spam detection, price prediction	Customer segmentation, recommender systems
Algorithm eg:	Decision Trees, SVM, Linear Regression	K-means, PCA, Apriori.

① How to prevent overfitting?

1) Simplify the model
2.) Reduce features to keep only relevant features
3.) Regularization techniques like L1 & L2 regularization add penalties to complex models
4.) Early Stopping: Stop training when model's performance on the validation set starts to decrease.
5.) Cross validation: helps ensure the model generalizes better by validating its performance on diff. subsets of data.

Q How to prevent underfitting?

1) Increase model complexity
2) Add more features
3) Train for longer
4) Reduce regularization

VII Cross Validation

Cross validation is a technique, used to ~~validate~~ evaluate how well a ML model generalises to new data. It involves splitting a data into multiple subsets, training & testing the model on different combination of these subsets & averaging the results.

(1) Data Split into folds:

- The data is divided into k equal sized subsets or folds
- The choice of k depends on dataset size & complexity

(2) Training & Testing Process:

- The model is trained on k-1 of the folds & tested on the remaining fold.
- The process is repeated k times, with each fold getting a turn as the test set.

(3) Averaging results:

- The performance metrics (like accuracy or error rates) from each fold are averaged to give an overall estimate of model performance.

Types of cross validation:

1) k-fold cross validation (most common)

- Data split into k folds, the model is trained & evaluated k times, once for each fold.

eg. In a 5 fold cross validation, data is split into 5 parts & model is trained on 4 parts & tested on 5th part. rotating test set with each iteration.

2.) Stratified k-fold Cross validation:

- Ensures that each fold has a similar dist^n of classes, which is particularly useful for classification tasks with imbalanced classes.

- eg. If you have dataset with 90% class A & 10% class B labels, each fold will retain this ratio.

3.) Leave one out Cross Validation:
A special case of k-fold cross validation where K equals the no. of data points, meaning each fold contains only one data point for testing.

4.) Time Series Cross Validation:

- Used for time series data, where the order of data matter.
- training set grows with each fold & only future data points are used as test data.

eg. In forecasting, you might train the model on Jan data, test on Feb then train on Jan to Feb & test on March & so on.

VIII Neural Networks

series of algorithms designed to recognize patterns in data. Inspired by the structure of the human brain, neural network consist of layers of nodes (or neurons) that process data by adjusting connections based on learned patterns.

Str. of Neural Networks :

1) Input layer : first layer of the network. Each neuron in the i/p layer represents one feature of the data.
2) Hidden layer : intermediate layer where computation occurs
Neural networks can have one or multiple hidden layers, depending on the complexity of the task.
3) Output layer : final layer that produces the output

How neural networks work?

- Forward propagation :
- Loss calculation
- Backward propagation

Types of Neural Networks :

- Feed forward Neural Network
- Convolution Neural Network
- Reaurrent Neural Network
- Deep Neural Network
- Generative Adversarial Network

IX Hyperparameters.

are settings / configuration for a ML model that are fixed & tuned to help optimize model performance.

Common hyperparameters

1) Learning rate : Controls how much the model adjusts its parameters with each step of training.

2.) No. of epochs : Complete no. of passes through the entire dataset.

3.) Batch Size :

- The no. of training examples the model processed in one go before updating parameters.

4.) Tree depth & No. of trees :

i) Tree depth : Controls the max. depth of each tree. Shallow trees generalise better

ii) No. of trees : In random forests, more trees generally improve accuracy but increase computational cost.

5.) Bayesian Optimisation :

- A more sophisticated method that uses probability to model the relationship b/w hyperparameters & model performance