1. Introduction to ML

Machine Learning (ML) is the design and study of algorithms, that can improve automatically through experience and by the use of data.

Examples:

- Algorithms to separate spam emails from legitimate emails.
- Predict the price of House, based on aspects such as neighborhood, income, No. of rooms, area, etc.,
- Finance:
  - Banks analyze past data to build models to process credit card applications, loan processing, fraud detection.
  - Determine prices of financial instruments such as options, futures
- Stock market:
  - To develop algorithms to predict stock prices.
- Manufacturing:
  - Learning models can used for optimization, control and troubleshooting.
- Medicine:
  - ML can be used for medical diagnosis using available medical tests, past history and symptoms.
- Telecommunications:
  - call data can procesed for Network optimization, resource allocation and BS installation, etc.,

Data mining

Advances in computer networks and storage have led to the ability to store and process large amounts of data.

Example :- Point of sale (POS) terminals of all outlets of a franchise.

Data, Customer ID, Items bought, Total spent

Application of ML techniques to such large databases is termed as Data mining.

Pattern recognition

Learning specific patterns from a large amount of data

Example : Face Recognition

- Analyzing large number of sample face images to capture patterns.
- Later, the algorithm can recognize the face by checking for the pattern in a new image.

Artificial Intelligence (AI)

- Designing smart machines capable of performing tasks requiring human intelligence
- An AI system should have the ability to learn in a changing environment.

Examples :

- Smart assistants : Siri, Alexa
- Self-driving cars
- Chat bots

Machine Learning approaches and algorithms:

① Supervised Learning.

⊙ Learn the mapping from the input (x) to the output using example / training input-output data

⊙ Linear Regression and Classification are Supervised Learning.

(i) Linear Regression

- System that can predict a variable using several other attributes / inputs / features.

- Example: Determine price (y) of a used car based on inputs (x) such as
  - Brand (x_1)
  - Year of manufacture (x_2)
  - Engine capacity (x_3)
  - Mileage (x_4)
  - Type of part owner (x_5)

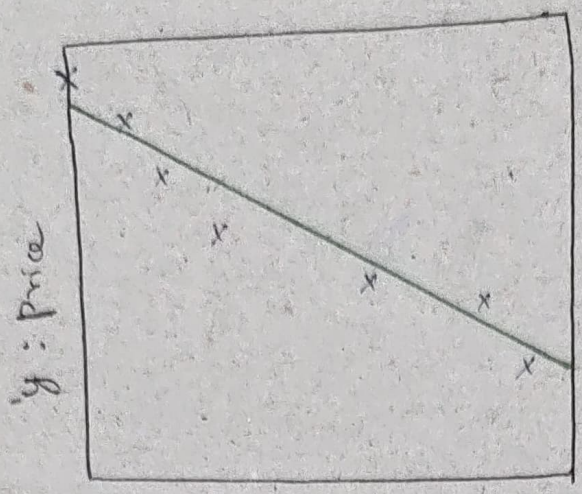
- Let x_i denote the regressors (Features) and y be the response (Price)

$$y(k) = h_0 + h_1 x_1(k) + \dots + h_n x_n(k) + \epsilon(k)$$

where, $k \rightarrow$ time / index

- $\bar{h}$ is the n-dimensional regression coefficient vector

$$\bar{h} = \begin{bmatrix} h_0 \\ h_1 \\ \vdots \\ h_n \end{bmatrix}$$



Price ↑ Mileage ↓

Price ↓ Mileage ↑

Plot : 1D Linear Regression

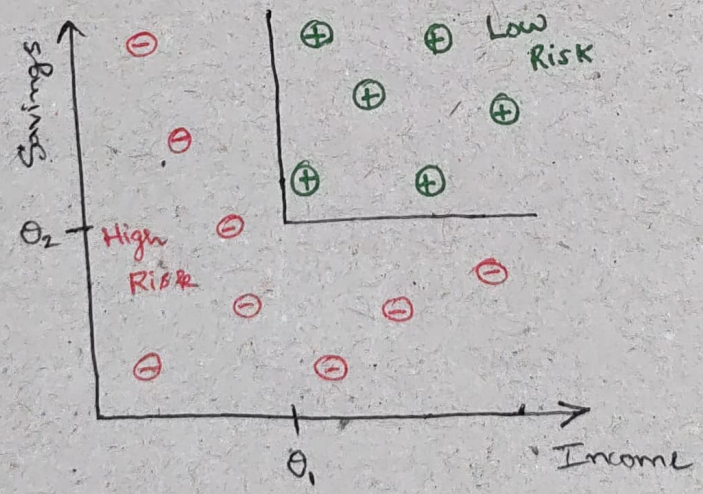
(iii) Classification

- Classify objects based on available data
- Typically 2 objects (sometimes more than 2)
  Eg. SPAM / NOT SPAM
- Example : To determine low-risk and high-risk customers for loan approval

IF (income > θ_1 , savings > θ_2)

THEN {low-risk}

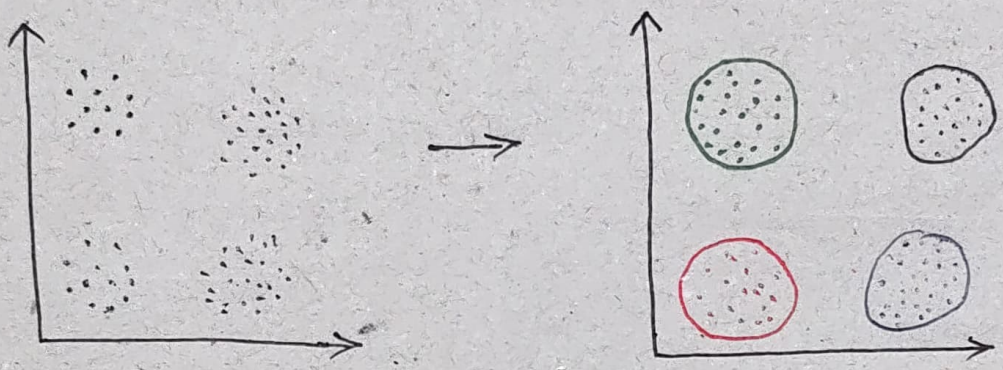
ELSE {High-risk}



② Unsupervised Learning

⊙ Identifying patterns in data sets that are NOT labeled.

⊙ Clustering is one of the most popular applications of Unsupervised Learning.
(i) discover groups of similar examples within the given data.



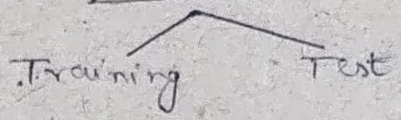
⊙ Clustering Applications : Customer Segmentation

Clustering customers in various groups based on specific attributes, provides a company with natural grouping of its customers.

Validation and Cross validation

Test and validation

For building an ML application, the available dataset is typically partitioned into two subsets.



(i) Training dataset

The subset used to fit the model or optimize the parameters of a model.

Example:

- Determine the weights of a neural net
- Compute the regression coefficients

(ii) Test dataset

The subset of data used to finally evaluate the performance of the ML model.

Example:

- Compute Mean-Squared Error (MSE) of regression $(\text{Actual price} - \text{Expected price})^2$
- False positive or False Negative of classification
  Eg. Covid Test result ☺

Test-Train split

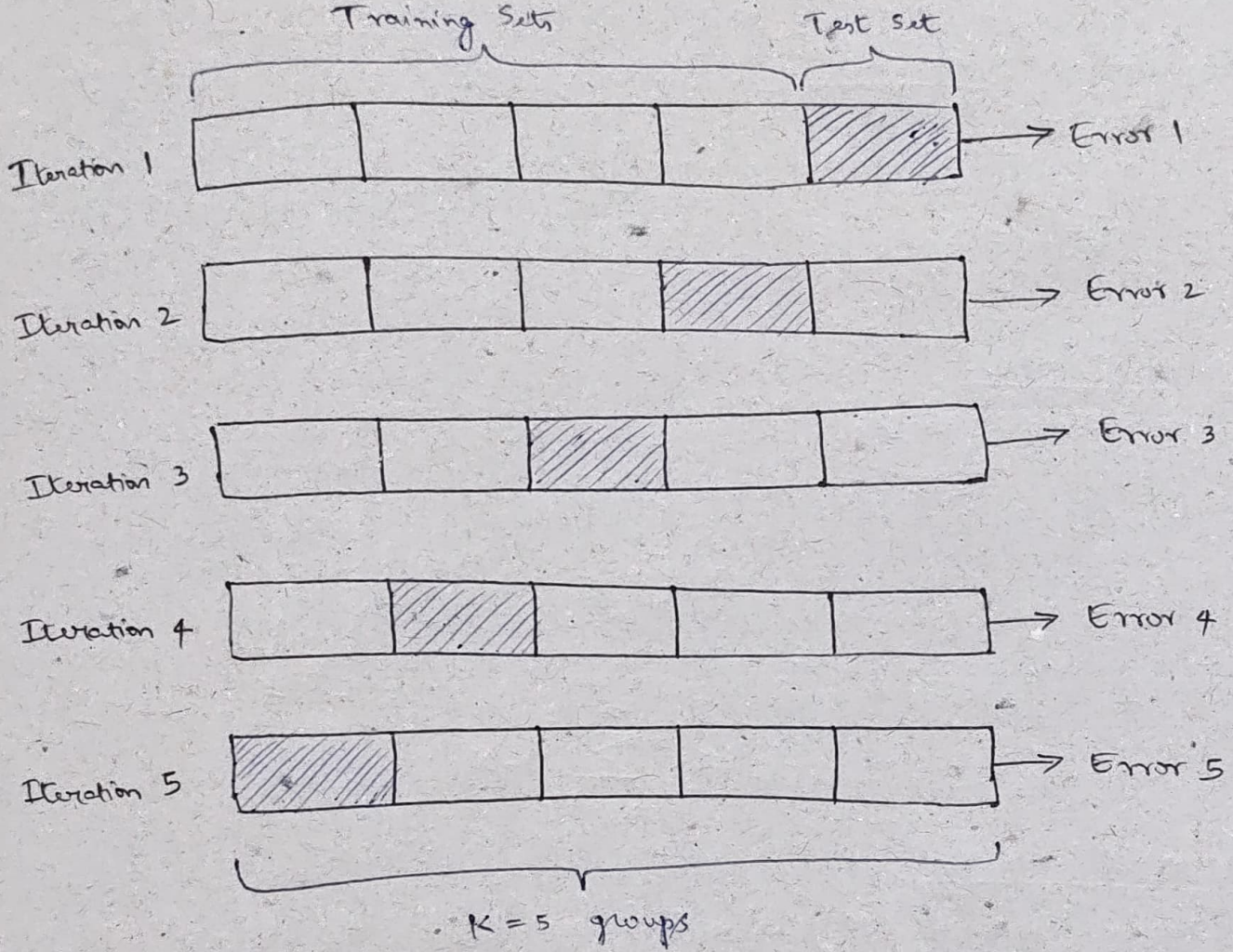
Frequently, a "test-train" split is employed.

- Hold out 10-20% of data for testing
- Rest is used for training

K-fold cross-validation

- Divide data into K groups
- For each group as Test data, rest groups are used as Training data, to fit model and evaluate score on test data.
- Final score is average of scores over all groups.

5-fold cross-validation



$$\text{Error} = \frac{1}{5} \sum_{i=1}^{5} \text{Error}_i$$