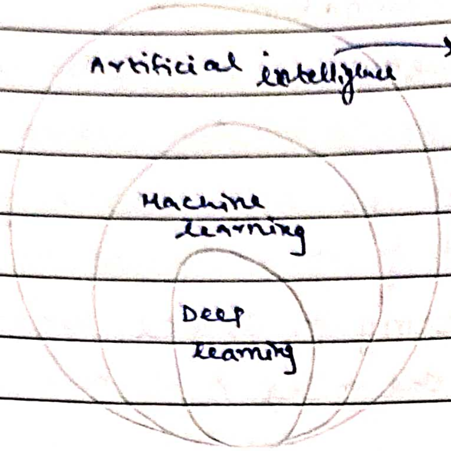
Intro to ML



→ mimics human behavior.

Arthur Samuel: Gives computers the ability to learn without being explicitly programmed.

Input, Data
↓
Intelligent System
↓
Decisions, output, Actions.

- Descriptive : explains what happened
- Predictive : predicts what could happen
- Prescriptive : use data, suggest further actions.

Sometimes solution is not trivial. It may be dependent on other features (ie, it is not linear).

General Strategy: Given (x, y)
Predict y

$\Rightarrow$ Given a new x, $y = F(x)$

x → Features   y → Prediction.

- x may be an N-dim vector
- can be entities other than numbers
- May be a collection of pictures. (image classification x : images, where image is represented by features)
- Sound bytes: Distinguish sound tracks. How can we tackle the feature set for sounds and classified.

- when we get datasets, there are too many features in the dataset / raw data.
- we need to find relevant information that may be hidden.
  Ex: Relevance of two images (are they similar).
- leads to feature extraction: Extracting useful info (x) from raw data.

Representation : From Raw data to Features

Convert all data into a vector of real numbers: X

★ Points in a feature space

No - 0   Yes - 1

Convert all predictions into an int / real number : Y

Ex: Suppose housing data has 'locations'.

how do we deal with categorical data?

Colours

how to represent in integers.

Categorical data

↙ ↘

Ordinal Data

- Meaningful order/ rank

- Ex: Satisfaction Rating

Poor	0
Fair	1
Good	2

- Integer encoding

low - 0
medium - 1
high - 2

Nominal Data.

- categories are names/ labels with no inherent data.

- Ex:

Colours	Pets
Red	Dogs
Blue	Cats
Gree	Fish

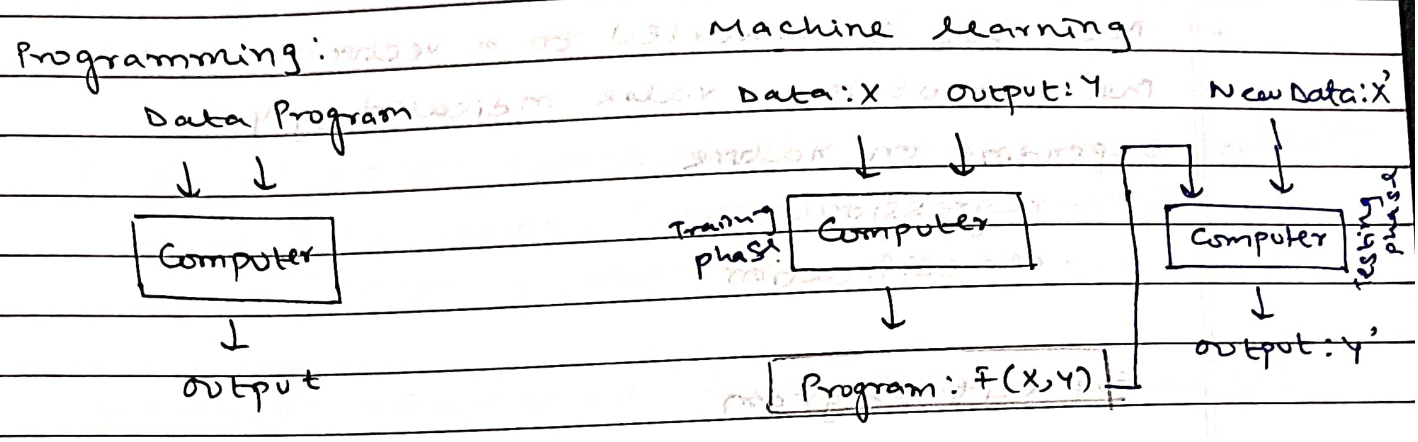
- One-hot encoding
  - for no order data or to prevent ~~data~~ algos to think data is ordi

One hot encoding.

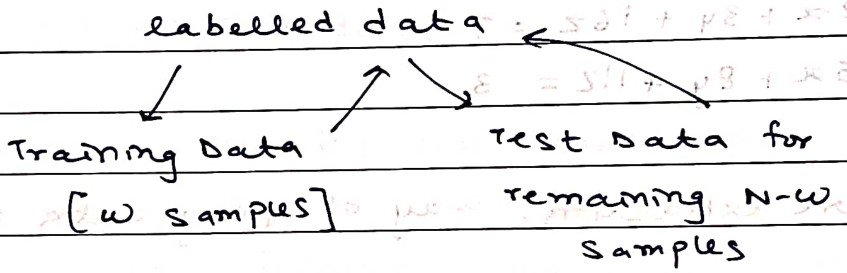
	x_6	x_7	x_8
Pets	Cat	Dog	Fish
Cat	$\boxed{1}$	0	0
Cat			
Dog			
Fish.			

→ Can lead to a significant increase in feature column.

Flow
- Identify features
- Conv features to integral.
- Chose ML algo to find relations/ patterns
- Now, it can predict new values



ML Based Train-Test Data



*Learning is concerned with accurate prediction of future data
Better Prediction - ML model strong
learnt well.

SUMMARY

$Y = F(x)$

o/p — Prediction — feature/representation

Note: Traning and Testing set comes from same distribution

(concept of iid)

- The input is converted to a vector x.
- The output is a value indicated by y.
- Depending on nature
  - regression
  - classification

Explicit program

x_1	x_2	x_3	y
2	3	16	7
5	8	11	3

$$2x + 3y + 16z = 7$$
$$5x + 8y + 11z = 3$$

feature extraction: way of giving data to ML data.

* Representations:

- In ML, it refers to the way data is transformed or encoded into a format that is suitable for a learning algo to process.

- Images: Raw Pixel Representation
  Deep learning based features.

how can we diff two images?

- they can be pixelized.
- each pixel has a different ~~color~~ number.
- Each number has all the info carried by a number

→ Sum of all Pixels
→ Number of boundary Pixels } Pixel Analysis
→ Edge detection

Instead of us to find best features to distinguish b/w 2 images, let ML model find out. => DEEP LEARNING

→ SOUND : waveform representation, spectrogram represent^n,
Mel-frequency Cepstral Coeff (MFCC)
↳ Technique by which given wavelength, wave frequency.

spectogram

→ Text data: Differentiate Regular & Spam messages

Approaches

- N grams
- Bag of words
- Term Frequency - Inverse Doc Frequency
- word Embeddings.

Ex: N grams : one file/doc about happy
one file/doc about sad

check frequency at which words are repeated
or even synonyms.
or combination of words
helps distinguish b/w documents

Ex: Bag of words

S1: weather is sunny today
S2: weather was rainy today.

Represent these sentences in form of a vector
and then distinguish.
Conv text data into numbers.

why AI ?

Now : very good, ~~hitg~~ highly quality datasets

improved architectures

Proficient ML models, improved architectures

- Massive parallel computing

- software platforms, cloud compute, API's libs,

- New Regularization techniques, Robust optimizers.

- Identify features, without good features may lose out on imp relations.
- Data can be in form of image, Sound, data.
- complex data - Numeric data