Unit. 4 Machine Learning

Syllabus:
- Intro to ML
- Supervised Learning
- Optimization Technique

ML

Supervised Learning | Unsupervised Learning | Reinforcement Learning

Open Book: Article Review
Implement following supervised learning algos from scratch: (Donot use any libraries). Implement every thing by developing your own fns.

Intro to ML

- Arthur Samuel (1959)
- ML – gives computers ability to learn w/o being explicitly programmed.
- wrote a checkers playing program
  - ↳ played 10,000 games against itself

E = 10,000 games
T = playing checkers
P = if you win or not

ML → "well posed learning problem":
↳ A computer program's ability is measured against performing a particular task. we can say that a computer program is said to learn from experience 'E' w.r.t some class of tasks 'T' & performance measure 'P'.

M/c has learnt if it's performance at task set 'T' improves w/ experience 'E'.

Data Warehouse (largest qty)
↑
Data Repository
↑
Dataset
↑
Data (smallest qty)

SR. NO	Cars	SUVs	Trucks	Owned by MRF
1				
2				Yes
3				No
:				Yes
10,000				

} Attributes / Features.

} classes' yes | no. Labels.

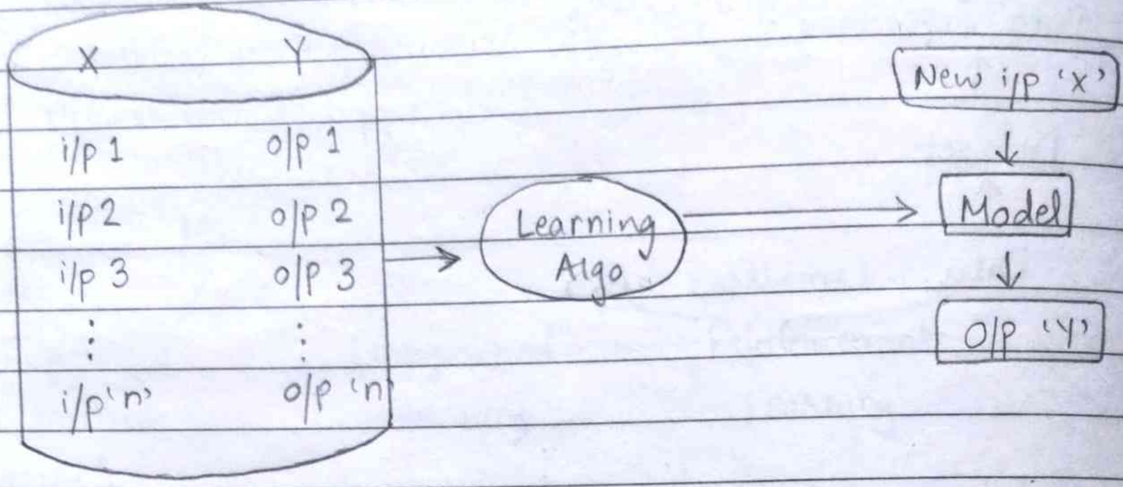
→ Supervised Learning

- It is a type of ML algos where labelled data (used as training data), i.e, Input (X) – Output (Y) : i/p - o/p paired data is used for training the ML algo. A fn that maps ~~the~~ an i/p to o/p based on i/p - o/p pairs can be written as $y = f(x) \in$ $Y = f(X)$

- In supervised learning algos, unlabelled data (testing data) w/ only i/p 'X' values is used to predict o/p 'Y' based on training.

- For any ML algo, the data is split into 80 : 20 ratio for training : testing purpose; where 80% data is used for training the ML algo & 20% data is used for testing.

Supervised Learning



In supervised learning, if the algo predicts numerical value, i.e, in terms of continuous o/p variable, then such algos are called regression algos.
For eg: Linear regression or decision tree w/ entropy value

Supervised learning algos that generate discrete o/p, i.e, categorical o/p such as Yes/No, True/False, etc are called as classification algos.
For eg: Support vector m/c, Random forest, etc.

07/04/25. Supervised Learning
- → Labelled data
- → Algos
- → Train/test → validate

X → features - $\{x_1, x_2, x_3, \cdots x_n\}$ → Independent variable
Y → target variable / output variable → dependent —"—

Supervised learning Algo

Regression · classification

→ ① Simple Linear Regression
↓
one independent variable
— " — dependent — " —

→ ② Multiple Linear Regression
↓
two/more independent variable
one dependent — " —.

eg:

Advertising Budget (lakh INR) (X)	Sales (Y) (%)
1	1.5
2	1.9
3	3.2
4	4.0
5.	4.8

$Y = \beta_0 + \beta_1 X.$ where β_0 : intercept
β_1 : slope

similar to straight line eqn: $y = mx + c$.

$$\boxed{\beta_1 \text{ (slope)} = \frac{\sum (x_i - \bar{X})(y_i - \bar{Y})}{\sum (x_i - x)^2}}$$

where $\bar{X}, \bar{Y}$: mean

In above eg, $\bar{X} = (1+2+3+4+5)/5 = 3$.
$\bar{Y} = (1.5 + 1.9 + 3.2 + 4.0 + 4.8)/5 = 3.08$

→

P.T.O.

Linear Regression

① Simple Linear Regression.

eg: Fit a linear regression model to predict the sales of a pdt based on it's advertising budget.

Advertising Budget (X) (lakh INR)	Sales (Y)
1	1.5
2	1.9
3	3.2
4	4.0
5	4.8

Step-1 is input data ↵

Step-2 Simple Linear Regression has one indepent variable & one dependent variable.

So, mathematically it is represented as eqn:

$$\boxed{Y = \beta_0 + \beta_1 \cdot X}$$

where

β_0 : intercept

β_1 : slope

X: independent variable

Y: dependent variable.

Step-3: First we will calculate mean of X & Y as:

$$\overline{X} = \frac{(1+2+3+4+5)}{5} = 3$$

$$\overline{Y} = \frac{(1.5+1.9+3.2+4.0+4.8)}{5} = 3.08$$

Step-4 : $\beta_1 \text{ (slope)} = \frac{\Sigma (x_i - \overline{X})(y_i - \overline{Y})}{\Sigma (x_i - \overline{x})^2}$

→

x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{Y}$	$(x_i-\bar{x})(y_i-\bar{Y})$	$(x_i-\bar{x})^2$
1	1.5	-2	-1.58	3.16	4
2	1.9	-1	-1.18	1.18	1
3	3.2	0	0.12	0	0
4	4.0	1	0.92	0.92	1
5	4.8	2	1.72	3.44	4
				8.7	10

$\therefore$

$$\beta_1 \text{ (slope)} = \frac{8.7}{10} = 0.87$$

Step-6. Calculate intercept:

$$\boxed{\beta_0 \text{ (intercept)} = \bar{Y} - \beta_1 \bar{x}}$$

$$= 3.08 - (0.87 \times 3)$$

$$\beta_0 = 0.47$$

Now, Predict sales:

Predict sales when company's advertising budget is 6 lakhs:

Using eqn:

$$Y = \beta_0 + \beta_1 x$$

$$= 0.47 + (0.87 \times 6)$$

$$Y = 5.69$$ — Ans.

eg. Suppose we want to predict the study hours a student needs to invest/spend based on the no. of practice problems they solve.

$\longrightarrow$

P.T.O.

Practice Problems (X)	Study Hours (Y)
10	2
15	3
20	3.5
23	4
25	5

Our goal is to find no. of hours required if the student solves 48 practice problems.

Step-1. Means:

$$\bar{X} = \frac{(10+15+20+23+25)}{5} = 18.6$$

$$\bar{Y} = \frac{(2+3+3.5+4+5)}{5} = 3.5$$

Step-2.

X	Y	$(x_i-\bar{X})$	$(y_i-\bar{Y})$	$(x_i-\bar{X})(y_i-\bar{Y})$	$(x_i-\bar{X})^2$
10	2	-8.6	-1.5	12.9	73.96
15	3	-3.6	-0.5	1.8	12.96
20	3.5	1.4	0	0	1.96
23	4	4.4	0.5	2.2	19.36
25	5.	6.4	1.5	9.6	40.96
				26.5.	149.2

Step-3 β_1 (slope) $= \frac{26.5}{149.2} = 0.177613941$.

Step-4. β_0 (intercept) $= \bar{Y} - \beta_1\bar{X}$

$= 3.5 - (18.6 \times 0.177613941)$

$= 0.196380697$

Step-5 For 48 problems;

$Y = \beta_0 + \beta_1 X$

$= 0.196380697 + (0.177613941 \times 48)$

$Y = \underline{8.72184965}$ hours — Ans.

o/p : output
i/p : input

Predicting categorical data — binary data → (T/F or yes/no)
⇓
Logistic Regression Algo. → (1/0)

M.L.E → Maxi. Likelihood Estimation

→ **Logistic Regression (L.R)**

In L.R, the o/p is always a binary one, i.e, categorical o/p. For eg: True/False, Yes/No, Pass/fail, etc. In Logistic Regression, our goal is to build a ML model that predicts categorical o/p based on i/p features. Generally, L.R uses sigmoid fn to calculate probability of given categorical data. The fn is:-

$$Y = \frac{1}{1+e^{-(+\beta_0+\beta_1 x)}}$$

where
$e \approx 2.71828$
x = independent variable / i/p feature
β_0, β_1 : coefficients
β_0 → intercept
β_1 → coeff. for i/p feature 'x'

Eqn is represented as:-

$$\boxed{P(Y_{class=c_i}|x) = \frac{1}{1+e^{-(\beta_0+\beta_1 x)}}}$$

where
$P(Y_{class=c_i}|x)$ is probability of $Y=c_i$ where c_i is the set of 2 binary categories for eg
c_i = {Yes, No} or c_i = {True, False}, etc.
c_i = {Pass, Fail}, or c_i = {0, 1}.

eg: Input data:-

Study Hours (X)	Result (Y)
1	0 (Fail)
2	0
3	1 (Pass)
4	1
5	1

Assume $\beta_0 = -3$
$\beta_1 = 1$

$$P(\text{Pass} \mid \text{Study hours}) = \frac{1}{1+e^{-(\beta_0+\beta_1 X)}}$$

$$P(\text{Pass} \mid 1\text{hr}) = \frac{1}{1+e^{-(-3+1(1))}}$$

$$= \frac{1}{1+e^2} = 0.1192 = 11.92\% \text{ passing}$$

$$P(\text{Pass} \mid 2\text{hrs}) = \frac{1}{1+e^{-(-3+1(2))}}$$

$$= \frac{1}{1+e} = 0.2689 = 26.89\% \text{ passing}$$

$$P(\text{Pass} \mid 3\text{hrs}) = \frac{1}{1+e^{-(-3+1(3))}}$$

$$= \frac{1}{1+1} = 0.5 = 50\% \text{ passing}$$

→
P.T.O.

$$P(Pass \mid 4hrs) = \frac{1}{1+e^{-(-3+4)}}$$

$$= \frac{1}{1+e^{-1}} = 0.73105 = 73.10\% \text{ passing}$$

$$P(Pass \mid 5hrs) = \frac{1}{1+e^{-(-3+5)}}$$

$$= \frac{1}{1+e^{-2}} = 0.88079 = 88.079\% \text{ passing}$$

Step-3 Making predictions using threshold

$0.119 < 0.5 \Rightarrow$ class = Fail $\Rightarrow$ predict (Fail)
$0.268 < 0.5 \Rightarrow$ predict (Fail)
$0.5 \geq 0.5 \Rightarrow$ predict (Pass)
$0.73 \geq 0.5 \Rightarrow$ predict (Pass)
$0.88 \geq 0.5 \Rightarrow$ predict (Pass)

* Imp Note:-
Finding β_0 & β_1
Unlike linear regression; in logistic regression there is no direct close form sol^n to find β_0 & β_1. Here, iterative optimization algo. named as Maximum Likelihood Estimation is used. This algo internally tries different values of β_0 & β_1 until the algo finds the one that maximizes the likelihood of observing given data.

eg: (continued previous example)
If a student's study hours are 8 & 10; predict the student's result.

$$P(\text{Pass} | 8\text{hrs}) = \frac{1}{1+e^{-(-3+8)}}$$

$$= \frac{1}{1+e^{-5}} = 0.9933 = 99.33\% \text{ passing}$$

$$P(\text{Pass} | 10\text{hrs}) = \frac{1}{1+e^{-7}} = 0.99908 = 99.908\% \text{ passing}$$

* Imp. points in Logistic Regression

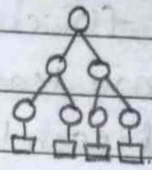
1) Finding optimal coefficient:
In a real-world scenario, you need to use statistical softwares/methods to find optimal values of β_0 & β_1 that best fit data using Max. Likelihood Estimation. Note that this involves iterative calculations.

2) Model Evaluation:
After building a logistic regression model, it is crucial to evaluate model's performance using matrix-like accuracy, precision, recall, etc.

$\longrightarrow$

(continued on next page)

Supervised Learning
- Linear Regression
- Logistic Regression
- Decision Tree → Binary Tree



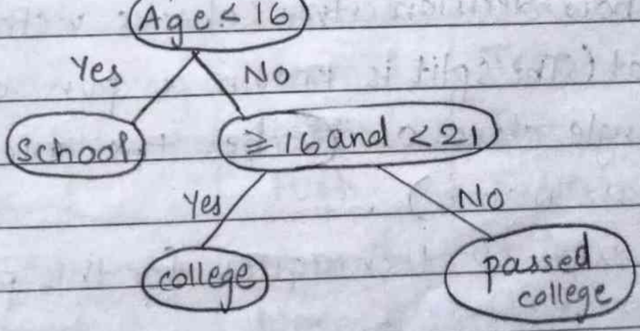
Decision Tree
↓
Split
Pure   Impure

f_1	f_2	f_3	O/P
			Yes
			No

needed for construct decision tree:
1. Entropy
2. Ginni Index / Ginni Impurity (G.I)
3. Information Gain.

eg:

```
age = 12
if age < 3:
    print ("baby")
elif (age ≥ 3 and age < 16):
    print (" school")
elif (age ≥ 16 and age ≤ 21):
    print ("college")
else:
    print ("passed")
```

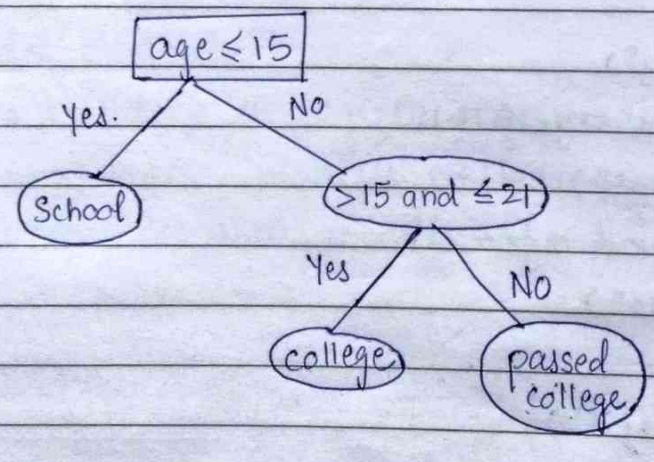


Page No.
Date 09 | 04 | 25

Decision Tree

- It is a supervised learning classification technique which constructs a flowchart-like structure where each internal node (i.e, non-leaf node) denotes a test attribute/feature and each leaf node denotes a class prediction.
- At each node, the algo chooses the best attribute to split the data into individual classes.
- For eg:

```
age = 12
if age ≤ 15:
    print ("school")
elif (age > 15 and age ≤ 21):
    print ("college")
else:
    print ("passed college")
```



- Mathematically, how decision tree checks wether the split is pure split or not (The split is known as pure split when it denotes a single class or reaches the decision of a class label / class name).
- Decision tree uses 2 techniques for this purpose:-

① Entropy.
② Ginni Index / Ginni Impurity

* Entropy

It is a measure of impurity or randomness in the dataset.

$$\text{Entropy}(S) = -\sum_{i=1}^{n} P_i \cdot \log_2 P_i$$

$$H(S) = -P_+ \log_2 P_+ - P_- \log_2 P_-$$

where P_+ : positive probability
P_- : negative —"—
P_i : probability of 'i' class.
$H(S) = \text{Entropy}(S)$

P_i: proportion of examples in class 'i'

In every class or decision tree constructed, entropy is zero when all examples are in one class, i.e, pure split.

NOTE: Entropy is max. when classes are evenly split.

eg:

Outlook	Temperature	Humidity	Wind	Play Tennis.
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	Weak	Yes.
Rain	Mild	High	Weak	Yes.
Rain	Cool	Normal	False	Yes.
Rain	Cool	Normal	True	No.
Overcast	Cool	Normal	True	~~No~~ Yes
Sunny	Mild	High	False	~~Yes~~ No.
Sunny	Cool	Normal	False	Yes.
Rain	Mild	Normal	False	Yes
Overcast	Mild	High	True	Yes.

Total Instances = 11.

Play Tennis → Yes = 7 ∴ $P(Yes) = 7/11$
→ No = 4 ∴ $P(No) = 4/11$

∴

$$Entropy(s) = -\left[\frac{7}{11}\cdot \log_2\left(\frac{7}{11}\right) + \frac{4}{11}\cdot \log_2\left(\frac{4}{11}\right)\right]$$

$$= 0.9456.$$

For outlook = Sunny.

Total instances = 4.

Play Tennis → Yes = 1
→ No = 3.

∴

$$Entropy(Sunny) = -\left[\frac{1}{4}\log_2\left(\frac{1}{4}\right) + \frac{3}{4}\log_2\left(\frac{3}{4}\right)\right]$$

$$= 0.8112.$$

For outlook = Overcast

Total instances = 3

Play Tennis → Yes = 3 } Pure split
→ No = 0 }
↓
∴ Entropy = 0.

|04|25

** Imp.

Supervised Learning	Unsupervised Learning
- Labelled Data	- Unlabelled Data
- X-Y Pairs	- Only X
- Training of algo. based on X-Y pairs.	- Training of algo based on similarity b/w features
- classes	- Groups / clusters
- eg: Linear Regression, Logistic —"—, Decision Trees.	- eg: K-means clustering, K-medoids —"—, Hierarchical —"—, DBSCAN —"—, Spectrum —"—