Page NO.

1. 1 to 18 – Mid Term
2. 19 to 45- Final Term

MD IFTAKHAR KABIR SAKUR

25th BATCH

COMPUTER AND COMMUNICATION ENGINEERING

International Islamic University Chittagong

COURSE CODE: CCE-4829

COURSE TITLE: Machine Learning

COURSE TEACHER:

Saiful Islam

Adjunct Faculty
Computer and Communication Engineering

Machine Learning

Midterm

24.02.24

CCE - 4829

Differences between ML Algorithm & Traditional Rule Based Algorithms:-

Rule Based :-	Machine Learning :-
⇒ ① Human crafted	→ Data Driven Approach.
② Interpretability	→ Automated Feature
③ Limited Adaptability.	→ Adaptability
④ Limited Scalability.	→ Scalability
⑤ Domain Specificity	→ Complexity
	→ Black Box Nature

Main problems solved by Machine Learning:-

AUT-22 2(b)

(i) Classification:-

In ML it is about training a model to predict the category or class of an input based on its features.

Ex:- Email is: spam or not spam

(ii) Regression:-

→ Drawing a line or curve through scatterplot of points to understand relationship between two or more variables.

Regression models predict continuous numerical values based on input features.

Ex:- Ice cream sales based on temperature.

(iii) clustering:-

Finding natural groupings within a set of objects based on their characteristics & similarities.

Ex:- A collection of ~~boxes~~ shapes of different colors. Where similar shapes will be grouped then based on their similarities inside the group

Machine Learning classification

Put-2 2(c)

① Supervised Learning :- (classification questions)

Means, having a teacher who teaches about which one is what.

The algorithm learns from labeled data. Where each example is associated with outcome. It allows the algorithm to make predictions on new, unseen data.

② ~~Unseen~~ Unsupervised Learning :- (clustering)

The algorithm learns patterns & structure ~~for~~ from unlabeled datas. The goal is to find similarities.

Ex:- 6 fruits are given. Now label them based on similarities.

③ Semi-Supervised :-

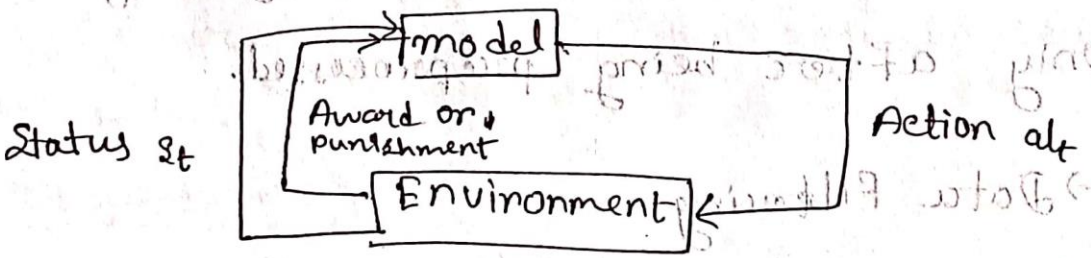
The algorithm learns from a combination of labeled & unlabeled data.

Use both the labeled & unlabeled images of animals. While you have some examples

of dogs & cats labeled, many others are unlabeled. You use both labeled & unlabeled data.

(iv) Reinforcement learning:-

Learns to make decisions by interacting with an environment. Receives feedback in the form of rewards or penalties based on actions & its goal is to learn a policy.



Basic Machine learning Concept

(i) Data Set:- A collection of data used in machine learning tasks. Each data record is called a sample.

(ii) Training Set:- Dataset used in training process. Where each sample is referred to as a

training sample. The process of creating model from data.

(iii) Test set:- process of using the model obtained after learning for prediction.

Data Cleansing

Most machine learning models process features which are numeric.

Collected data can be used by algorithms only after being preprocessed.

→ Data Filtering

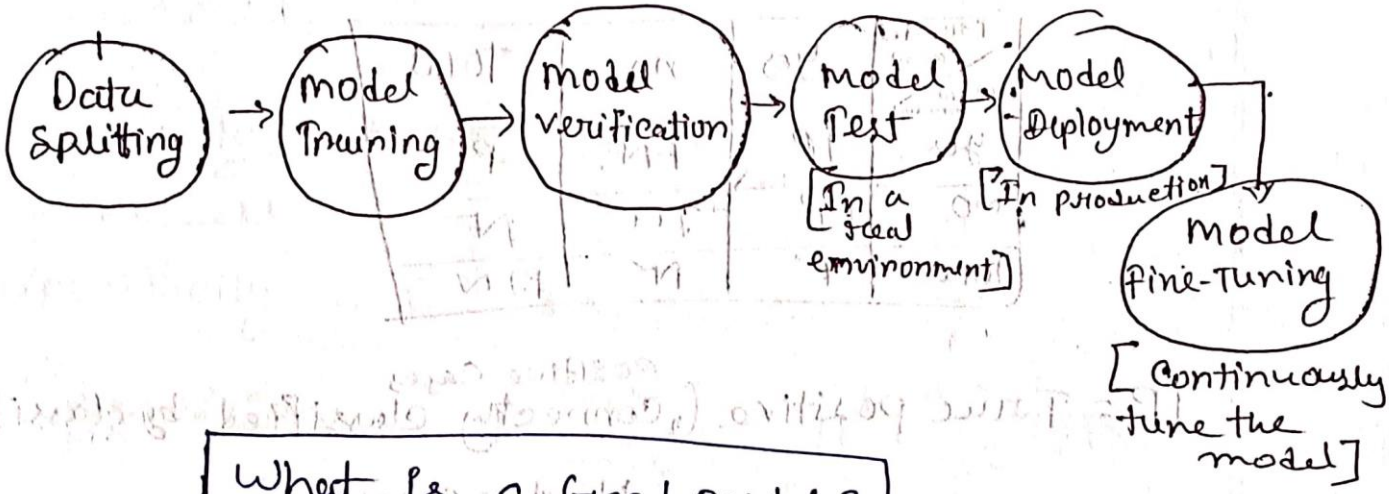
→ processing of lost data

→ processing of possible exceptions, errors

→ Combination of data from multiple data sources.

→ Data consolidation.

Overall procedure of Building a model



What is a Good model?

→ Generalization Capabilities:-
  ↳ Accurately predict the actual service data.

→ Interpretability:-
The prediction result easy to interpret.

→ prediction Speed:-
  How long does it take

→ Practicability:-
  Is the prediction rate acceptable or not.

Machine Learning performance Evaluation

Estimated Amount / Actual Amount	Yes	no	Total
Yes	TP	FN	P
NO	FP	TN	N
Total	P'	N'	P+N

TP = True positive (Positive cases correctly classified by classifier)

FN = False Negative (Negative cases incorrectly " ")

P = positive (The number of real positive cases)

FP = False positive (positive cases incorrectly classified by classifier)

TN = True Negative (Neg. cases correctly " by)

N = Negative (Neg. cases in data)

(a) Accuracy & recognition rate :- $\frac{TP+TN}{P+N}$

(b) Error rate & misclassification rate : $\frac{FP+FN}{P+N}$

(c) Sensitivity, true positive rate & recall : $\frac{TP}{P}$

(d) Specificity & True Neg. rate :- $\frac{TN}{N}$

(e) precision = $\frac{TP}{TP+FP}$ | recall → $\frac{TP}{TP+FN}$

(f) Harmonic mean of the recall rate = $\frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$

(g) F_β (β = Non Negative real Number) = $\frac{(1+\beta^2) \times \text{precision} \times \text{recall}}{\beta^2 \times \text{precision} + \text{recall}}$

Hyperparameters of a Model

→ Learning rate, Number of layers, Neurons in each layer, epoch these are hyper parameters.

→ Manually specified

→ Code :- model.coef_
model.intercept_

→ Often used in model parameter estimation process.

→ Often specified by the practitioner

→ Can often be set using heuristics.

Hyperparameter Search procedure & Method

① Dividing dataset into a training set, validation set & test set.

② Optimizing the model parameters using training set.

③ Searching for the model's hyper-params. using the validation.

④ perform step-2 & step-3 alternately.

Search Algorithm (Step-3):-

① **Grid search**:- Exploring a grid of possible combinations of hyperparameters.

② **Random Search**:- Randomly exploring different regions of the hyperparames space

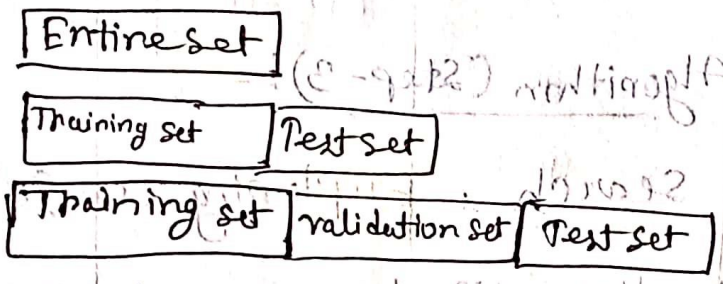
③ **Heuristic Intelligent Search**:- To find approximate solutions to complex problems efficiently.

④ **Bayesian Search**:- It is like using past observations to guide the search to get better hyperparameters.

Cross validation

Cross validation is a powerful technique for estimating a machine learning model's performance & assessing its ability to generalize to new data.

The basic idea is to divide the original dataset into two parts. Training set & validation set.



K-Fold Cross validation is a hyperparameters

K-fold Cross validation :-

used to evaluate the performance of a ML model more reliably than a simple train-test split.

~~K-Fold Cross~~

- Divide the raw data into K groups
- Each subset as validation test.
  - K-1 subsets as the training set.
  - K models as the performance indicator of the K-ev Classifier.

Reasons for underfitting:-

→ High bias & low variance
→ Training dataset not enough
→ Model is too simple
→ Training data is not cleaned & having noise in it.

Techniques to reduce underfitting:-

→ Increase model complexity
→ Increase number of features, performing feature Engineering
→ Remove noise from the data
→ Increase the number of epochs or increase the duration of training to get better results.

Reasons for overfitting:-

→ High variance low bias
→ The model is too complex
→ Size of the training data

Technique to reduce overfitting:-

→ Increase training data
→ Reduce complexity.

Avoid overfitting:-

→ Cross validation
→ Training with more data
→ Removing features
→ Early stopping the training

K-means clustering

→ Similar গুলোকে একসাথে রাখার জন্য clustering করা হয়।

Make 2 clustering based on given values:-

Serial NO	Age	Amount
C_1	20	500
C_2	40	1000
C_3	30	800
C_4	18	300
C_5	28	1200
C_6	35	1400
C_7	45	1800

⇒ Centralized value:-

C_1 (20, 500)

C_2 (40, 1000)

For, C_3

Formula:- $d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$

(x_2, y_2 → centroid; x_1, y_1 → centroid)

So, For $C_3 = \sqrt{(30-20)^2 + (800-500)^2}$ [This calculation based on C_1]

$= \sqrt{(10)^2 + (300)^2}$

$= 300.\ldots$

C_3, value based on C_2

$C_3 = \sqrt{(30-40)^2 + (800-1000)^2}$

$= 200$

So, $200 < 300$

So, C_3 will be in C_2

[20, 500] [40, 1000] ⇒ Updated C_2 value

Centroid value updatation:-

$\frac{40+30}{2} = 35$

$\frac{800+1000}{2} = 900$

So

[35, 900] value.

For, $C_4 = \sqrt{(18-20)^2 + (300-500)^2} =$

$C_4 = \sqrt{(18-35)^2 + (300-900)^2} =$

→ Updation:-

$\frac{18+20}{2} = 19.5$

$\frac{500+300}{2} = 400$

[19.5, 400]

Example-2 [6, 7, 8, 9, 10, 11, 12, 13, 14].

C_1	C_2	C_3	
9	10	11	[Randomly chosen]

Now select each value & find distance between Centroid

9-6 = 3 | 10 - 6 = 4 | 11-6 = 5

So, (9 6) (10) (11)

[To be continued]

... Until the clusters don't change anymore.

Model Validity

Generalization Capabilities:-

The goal of ML is the model should be working well on new samples not only on the training sample. And this capability of applying a model to new samples is called generalization or robustness.

Error:-

The difference between the sample that are predicted by the model obtained after learning & actual sample ~~data~~ result.

→ **Training error:-** Error during training

→ **Generalization error:-** Error when the model is run on new sample.

Underfitting :-

Occurs when the model or the algorithm does not fit the data well enough.

Overfitting:- When learning is small but the generalization error is large.

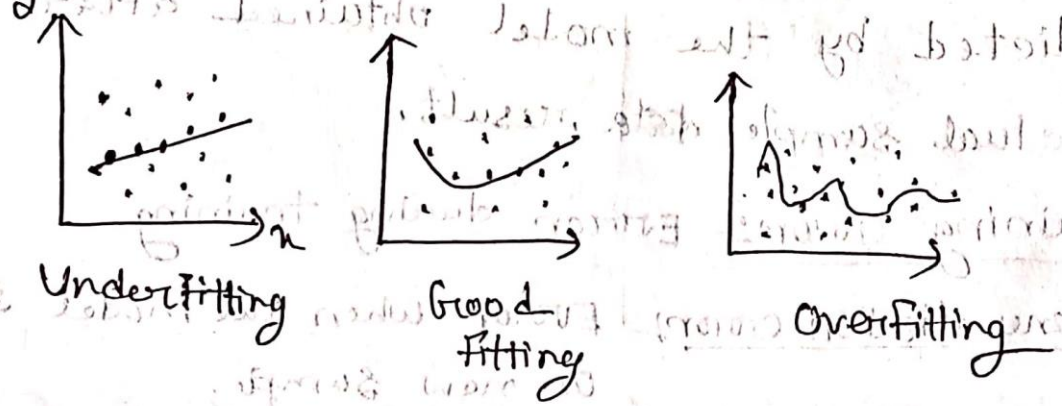
Model Capacity :-

Model's capacity of fitting functions which is called as model complexity.

→ If model suits the task then algorithm is optimal

→ Model with insufficient capacity can't solve complex tasks.

→ High capacity can solve but solve complex tasks.



AUT-2022

(1) (b) Working principle OF ML:-

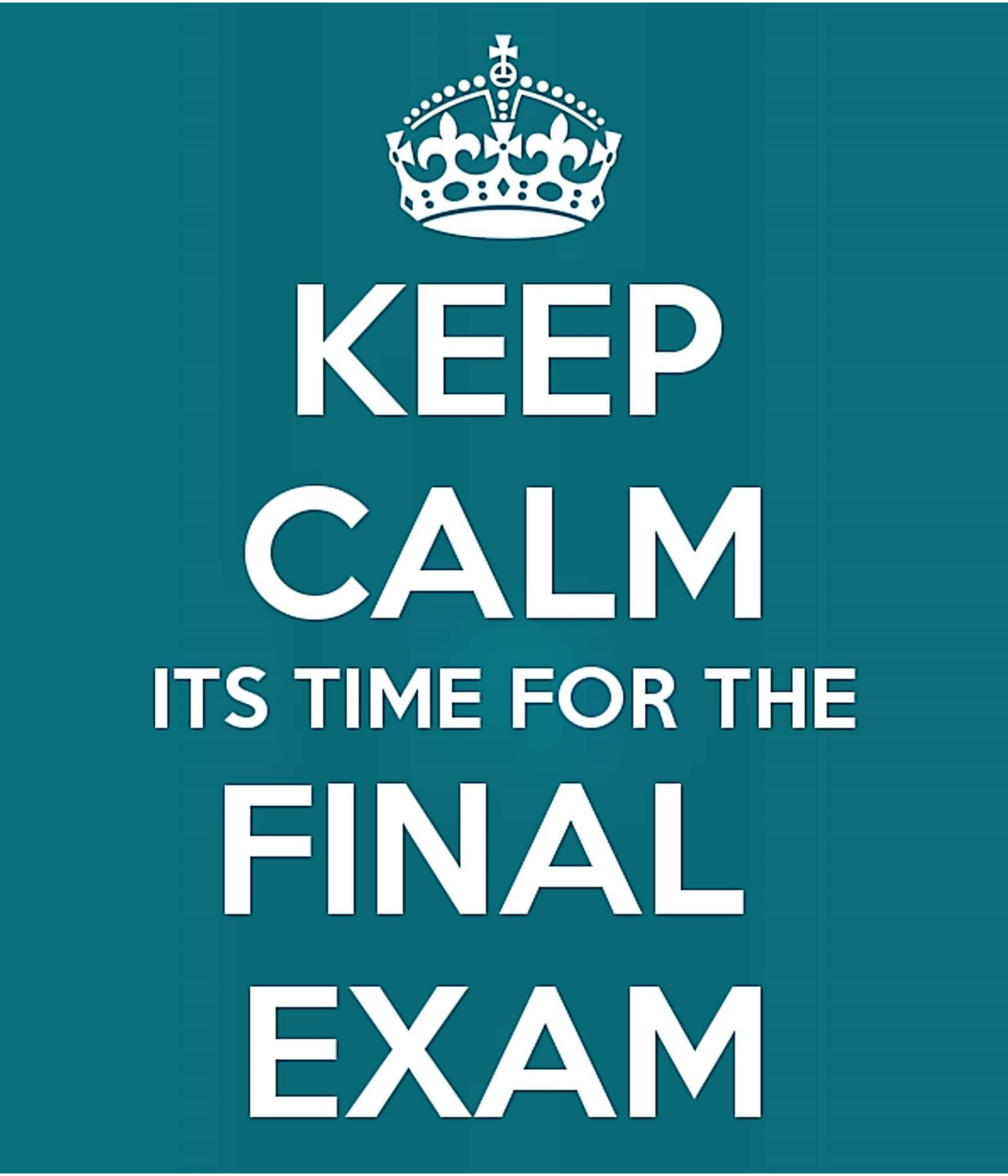
→ Data Collection
→ Data Preprocessing
→ Model selection
→ Model Training
→ Model Evaluation
→ Model deployment

(c) Image recognition most Common Application:-

3 (a)

Clustering play in unsupervised learning:-

→ Grouping Similar data
→ NO prior labels
→ Clustering Algorithms
→ Evaluation
→ parameter tuning.


KEEP
CALM
ITS TIME FOR THE
FINAL
EXAM

[CCE-4829]

[Machine Learning]

[Final]

What is Hidden Markov Model?

=> A statistical model that represents a system containing hidden states where the system evolves over time.

It is hidden cause the state of the system is not directly visible to the observer.

But the observer can see some outputs that depend on the state

→ Markov models are characterized by the Markov property which states that the future state of a process only depends on its current state, not on the sequence of it.

It is used in : Speech, Handwriting, gesture recognition, part-of-speech tagging, musical score following

Components of Hidden Markov model :-

(1) Hidden States :- Not directly observable. ex:- Speech recognition system.

(2) Observations :- Data point or output that can be observed. ex:- In speech recognition, the observations could be the audio signal feature.

(3) Transition probabilities :-

Transition from one hidden state to another.

These are usually represented in a matrix form.

(4) Emission probabilities :-

These describe the probability of an observed output given a hidden state.

(5) Initial State probabilities :-

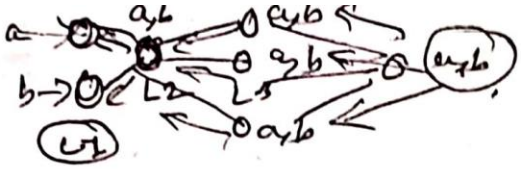
These represent the probabilities of the system being in a particular state at the beginning of the process.

How Hidden Markov model works?

① **Evaluation problem:** Given the model parameter & an observation sequence, determine the probability of the observation sequence. This is solved using forward-backward algorithm.

② **Decoding problem:** The decoding problem is an observed sequence of data; this decoding problem is to determine the most likely sequence of hidden state.

③ **Learning:** The learning problem is to estimate the model parameters. The Baum-Welch algorithm is often used for this.



Hidden Markov Model Algorithm :-

Step-1 :- Define state space & observation space
(All possible hidden state & observation space)

Step-2 :- Define the initial state (The probability distribution)

Step-3 :- Define the state transition
(Transition from one state to another)

Step-4 :- Define the observation likelihoods.
(probabilities of generating each observation from each state)

Step-5 :- Train The model.
(Baun welch Algorithm, Forward-backward algorithm)

Step-6 :- Evaluate model (Accuracy, precision, F1 score)

Limitation of Hidden Markov Model

(1) Limited Modeling Capabilities:
→ designed to model sequences of data.
→ Ex:- In speech recognition, the complex relationship between the speech sounds & the corresponding acoustic signals may not fully captured by the simple structure of an HMM.

(2) Overfitting:- When the model fits too well for the training data & can't generalize to new data. As a result high error rates can result. So, needs to be careful to choose the number of hidden state.

(3) Lack of Robustness:- limited in their robustness to noise & variability in the data.

Ex: In speech recognition the acoustic signal there the data can have noise in it. which can result in poor performance.

Computational Complexity :

→ limited in its
→ Specially when dealing with large amounts of data.
→ And to address this limitation, it is often necessary to use parallel computing techniques or to use approximate that reduce the computational complexity of the model.

Topic:

Topic: Decision Tree (Theory)

=> A supervised learning algorithm which looks like an inverted tree, each node represent a predictor variable, the link between the nodes represents decision tree & each leaf node represent an outcome.

How does the decision Tree algorithm works?

=> Step-1: Select the Feature that best classifies the data set into desired classes.

Step-2: Traverse down from root node.

Step-3: Route back to step-1.

Decision Tree Structure:

→ Root Node: The starting point of a tree.

→ Internal node: Represents a decision point that eventually leads to the prediction of the outcome.

→ Leaf Node: Represents the final class of the outcome.

Branches:- Are connections between nodes.

Expressiveness of decision trees:-

Can represent any boolean function of the input attributes.

A	B	A & B
F	F	F
F	T	F
T	F	F
T	T	T

And

A	B	A OR B
F	F	F
F	T	T
T	F	T
T	T	T

OR

A	B	A ⊕ B
F	F	F
F	T	T
T	F	T
T	T	F

X-OR

$A \cdot \bar{B} + \bar{A} \cdot B$

$F + F$

$F \cdot F + T$

How a decision tree is created:-

→ Using ID3 Algorithm.

→ ID3 = Iterative Dichotomiser 3 Algorithm

→ Uses the concept of Entropy & information gain to generate a decision tree for a given set of data.

ID3 Algorithm:

① Best Attribute selection. (Separate the dataset into different classes. Can get:-
↳ Information Gain
↳ Entropy)

② Assign A as a decision variable for the root node.

③ For each A, build a descendant of the node.

④ Assign classification levels to the leaf node.

⑤ If data is correctly classified :- Stop

⑥ else: iterate over the tree.

Entropy:- Measures the impurity or uncertainty present in the data.

$$\text{Entropy} = -\sum P(x) \cdot \log P(x)$$

Info. gain:- How much information a particular feature/ variable gives us about the final outcome.

Info gain(Cx) = entropy(parent) - [weighted avg]* entropy(children)

= entropy(S) - Entropy(S,T)

Entropy(children) with weighted average:-

[weighted avg] Entropy(children) = (no. of outcomes in left child node)/(total no. of outcome in parent node)* (entropy of the left node)

+ (no. of outcome in right child node)/(total no. of outcomes in parent node) * (entropy of right node)

Entropy for two variable:-

$$\text{Entropy}(S,T) = \sum P(c) \cdot E(c)$$

Topic:- Classification Analysis

=) Classification algorithms are used to predict/ classify the discrete values.

Types of ML classification:

1. Logistic algorithm
2. K-Nearest Neighbours
3. Support vectors Machines
4. Kernel SVM
5. Naive Bayes
6. Decision Tree Classification
7. Random Forest classification

Regression:-

→ Relation between dependent & independent variables.

→ It is all about predicting a quantity.

→ Used to predict the continuous values such as (price, salary, age etc).

Types of Regression Algorithm :-

→ Simple Linear Regression
→ Multiple Linear Regression
→ Polynomial regression
→ Support vector Regression
→ Decision Tree regression
→ Random Forest Regression

Difference (Regression & Classification)

Regression	classification
(1) Must be continuous	(1) Must be discrete value
(2) Map the input value (x) with the continuous output variable (y).	(2) Map the input (x) with discrete output (y)
(3) used with cont. data	(3) Used with discrete data
(4) Find the best fit line	(4) Divide dataset into different classes
(5) Algorithm used to solve regression problem.	(5) classification problem solve.

⑥ Can be divided into Linear & Non-Linear regression.	⑥ can be divided into binary classifier & multiclass classifier
⑦ Weather prediction, House price etc.	⑦ Speech recognition, Cancer cells etc.

☐ What is a support vector?

⇒ Support vectors are simply the co-ordinates of individual observation.

→ ~~It~~ is

☐ How to find the SVM for case in hand?

⇒

☐ SVM Kernel Technique:-

It is a function that takes low dimensional input space & transforms it to a higher dimensional ~~page~~ space.

Mostly useful in non-linear seperation problem.

Types of SVM:-

2 types:-

① **Simple SVM:-** used for linear regression & classification problems

② **Kernel SVM:-** Can add more features to fit a hyperplane instead of a two-dimensional space.

Kernel function:-

↳ Linear
↳ polynomial
↳ Gaussian Radial Basis function
↳ Sigmoid.

Linear :-

→ Linear kernel works really well when there are a lot of features.

→ Are faster than others.

→ Have fewer parameters to optimize.

→ $f(x) = w^T * x + b$

w = weight
x = data
b = linear co-efficient

Polynomial :-

→ Not used in practice.

→ Not computationally efficient.

→ predictions are not accurate.

→ $f(x_1, x_2) = (a + x_1^T \times x_2)^b$

Gaussian Radial Bias Function (RBF) :-

→ Usually choice for non-linear data.

→ $f(x_1, x_2) = \exp(-\text{gamma} * ||x_1 - x_2||^2)$

gamma = How much a single training point has on the other data point.

$||x_1 - x_2||$ = dot product between feature

Sigmoid:

→ Used in Neural Network.

→ $f(x,y) = \tanh(alpha * x^T * y + c)$

alpha = weight vector

c = An offset value

V.Vim (For both ML & AI)

Math

Step-1: (Determine the decision column.)

Play golf (14)
Yes	9
No	5

Step-2: Calculating entropy for the classes (playgolf)

we know that,

$$\text{Entropy} = -\sum P(w) \cdot \log_2(P(w))$$

So, $\text{Entropy}(5,9) = -\left\{ \frac{5}{14} \cdot \log_2\left(\frac{5}{14}\right) + \frac{9}{14} \cdot \log_2\left(\frac{9}{14}\right) \right\}$

SAKUR-E201037

Step-3:- (Calculating entropy for other attribute after split)

① E (playgolf, outlook)

② E (playgolf, temperature)

③ E (playgolf, Humidity)

④ E (playgolf, windy)

Entropy for two variables we know,

$$\text{Entropy}(S,T) = \sum P(c) \cdot E(c)$$

So,

Entropy

① E (playgolf, outlook) = P(Sunny) · E(Sunny) + P(Rainy) · E(Rainy) + P(Overcast) · E(Overcast)

② E(playgolf, temperature) = P(Hot) · E(Hot) + P(mild) · E(mild) + P(Cool) · E(Cool).

③ E(playgolf, Humidity) = P(High) · E(High) + P(Normal) · E(Normal)

④ E (playgolf, windy) = P(True) · E(True) + P(False) · E(False)

E(Playgolf, outlook) calculation:

		Playgolf		Total
		yes	NO	
Outlook	Sunny	3	2	5
	Rainy	2	3	5
	Overcast	4	0	4

$$E(\text{Sunny}) = E(3,2)$$

$$= -\left\{ \tfrac{3}{5} \cdot \log_2\left(\tfrac{3}{5}\right) + \tfrac{2}{5} \cdot \log_2\left(\tfrac{2}{5}\right) \right\}$$

$$= -\{ 0.442 + 0.53 \}$$

$$= 0.972$$

$$E(\text{Rainy}) = E(2,3)$$

$$= -\left\{ \tfrac{2}{5} \cdot \log_2\left(\tfrac{2}{5}\right) + \tfrac{3}{5} \cdot \log_2\left(\tfrac{3}{5}\right) \right\}$$

$$= 0.972$$

$$E(\text{Overcast}) = E(4,0)$$

$$= -\left(\tfrac{4}{4} \log_2\left(\tfrac{4}{4}\right) + \tfrac{0}{4} \cdot \log_2\left(\tfrac{0}{4}\right) \right)$$

$$= 0$$

$\therefore E(\text{Play golf, Outlook}) = \frac{5}{14} \cdot E(3,2) + \frac{4}{14} \cdot E(4,0) + \frac{5}{14} \cdot E(3,2)$

$= 0.35 +$

$= \frac{5}{14} \times 0.972 + \frac{4}{14} \times 0 + \frac{5}{14} \times 0.972$

$= 0.35 + 0 + 0.35$

$= 0.7$

Now, I (Play golf, Temperature) Calculation :-

	Playgolf		Total
~~Hot~~	Yes	No	
Mild	4	2	6
Cool	3	1	4
Hot	2	2	4

$E(\text{Mild}) = E(4,2)$

$= -\frac{4}{6} \log_2\left(\frac{4}{6}\right) - \frac{2}{6} \log_2\left(\frac{2}{6}\right)$

$= -(0.39 + 0.53)$

$= 0.92$

$E(\text{Cold}) = E(3,1)$

$= \frac{3}{4} \cdot \log_2\left(\frac{3}{4}\right) + \frac{1}{4} \cdot \log_2\left(\frac{1}{4}\right)$

$= -(0.31 + 0.5)$

$= 0.81$

$E(\text{HOT}) = E(2,2)$

$= \frac{2}{4} \cdot \log_2\left(\frac{2}{4}\right) + \frac{2}{4} \cdot \log_2\left(\frac{2}{4}\right)$

$= -(0.5 + 0.5)$

$= 1$

$\therefore E(\text{playgolf, Temperature}) = E(\text{Mild}) \cdot P(\text{Mild}) + E(\text{Cold}) \cdot P(\text{Cold}) + E(\text{HOT}) \cdot P(\text{HOT})$

$= \frac{6}{14} \cdot 0.92 + \frac{4}{14} \times 0.81 + \frac{4}{14} \times 1$

$= 0.39 + 0.231 + 0.286$

$= 0.907$

F(PlayGolf, Humidity) :-

	PlayGolf		Total
	Yes	No	
High	3	4	7
Normal	6	1	7

$\therefore E(High) = 3/7 \cdot \log_2(3/7) + 4/7 \cdot \log_2(4/7)$

$= -(0.524 + 0.461) = 0.985$

$\therefore E(Normal) = E(6,1)$

$= \left(\frac{6}{7}\right) \log_2(6/7) + \left(\frac{1}{7}\right) \log_2(1/7)$

$= -(0.191 + 0.401)$

$= 0.592$

$E(PlayGolf, Humidity) = 7/14 \cdot 0.985 + 7/14 \times 0.592$

$= 0.492 + 0.296$

$= 0.79$

R.E (playgolf, windy) :-

	Playgolf		Total
	Yes	No	
True	3	3	6
False	6	2	8

$E(\text{True}) = E(3,3)$

$= 3/6 \cdot \log_2(3/6) + 3/6 \cdot \log_2(3/6)$

$= 0.5 + 0.5 = 1$

$E(\text{False}) = E(6,2)$

$= 6/8 \cdot \log_2(6/8) + 2/8 \cdot \log_2(2/8)$

$= 0.311 + 0.5$

$= 0.811$

$\therefore E(\text{playgolf, windy}) = 6/14 \times 1 + 8/14 \times 0.811$

$= 0.43 + 0.463$

$= 0.893$

Step-4 (Calculating Info gain):

① Gain (playgolf, outlook) = Entropy (playgolf) * –

Entropy (playgolf, outlook)

$= 0.94 - 0.7$

$= 0.24$

② Gain (playgolf, Temp) = E (playgolf) – ~~Entropy~~

E (playgolf, Temp)

$= 0.94 - 0.907$

$= 0.033$

③ Gain (playgolf, Humidity) = E (playgolf) –

E (playgolf, Humidity)

$= 0.94 - 0.79$

$= 0.15$

④ Gain (playgolf, windy) =

$\therefore$ E (Play..) – E (playgolf, windy)

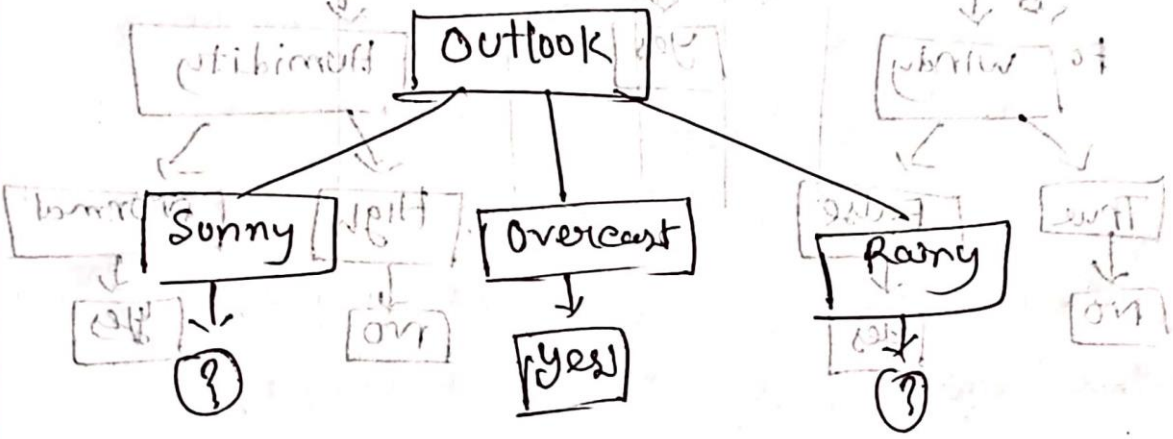
$= 0.94 - 0.893$

$= 0.047$

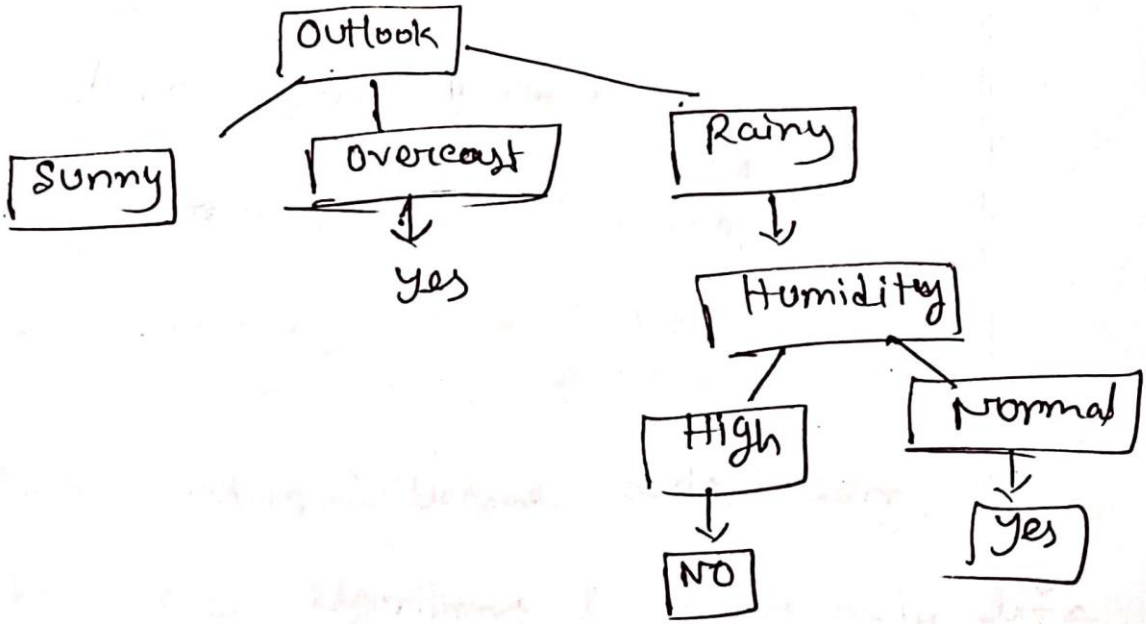
Step-5 :- (perform the first split)

Using highest information gain we will do that.

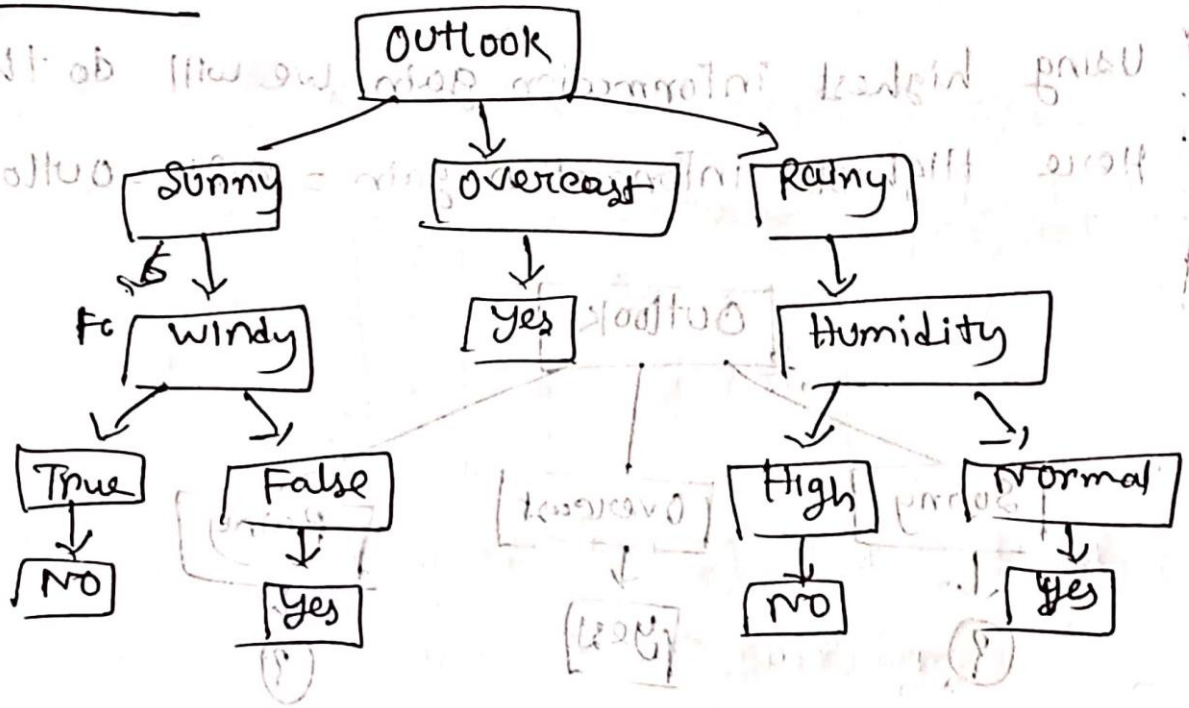
Here Highest information gain = 0.24 = Outlook



Step-6 :-



Step-7:-


Outlook
Sunny
Overcast
Rainy
Windy
Yes
Humidity
True
False
High
Normal
NO
Yes
NO
Yes