"Assignment No 4."

Q1) What is PAC Model?

Ans:- PAC stands for probably Approximately Correct (PAC) Learning.

- A good learner will learn with high probability and close approximation to the target concept.
- with high probability the selected hypothesis will have lower the error ("Approximatly Correct") with the parameter ε & δ
- PAC learning requires,
  - small parameter- ε & δ
  - with probability at least $(1-\delta)$, a system learn the concept with error at most ε.

- ε is upper bound on the error in accuracy i.e. the hypothesis with error less than ε.
  Accuracy $1-\varepsilon$

- ε give the probability of failure in achieving the accuracy. δ $(0<\delta<1)$, the hypothesis generated is approx. correct at least $1-\delta$ of the time
  confidence $1-\delta$

Q2) Explain Sample Complexity of finite hypothesis Space

→ - We seek to generalize Occam's razor to

infinite hypothesis space, To do so, we look at the set of behavior $\Pi_H(S)$ of hypotheses from H on a sample S.

$$\Pi_H(S) = \{\{h(x_1), \ldots h(x_m)\} : h \in H\}$$
for $S = \{x_1, \ldots, x_m\}$

$\Pi_H(m) = \max_{S: |S|=m} |\Pi_H(S)|$ defines the growth function H.

- our goal is to modify Occam's razor to get a bound of the form

$$err(h) \leq O\left(\frac{In\,\Pi_H(2m) + In\,\frac{1}{\delta}}{m}\right)$$

- 1st, some definations, for our proof of this bound, we fix ϵ and let D denote our target distribution.

Step 1 : $Pr[B'[B] \geq 1/2$
step 2: $Pr[B] \leq 2Pr[B']$
step 3 = $Pr[B''] = Pr[B']$
step 4 = $Pr[b(h) \cdot S, S'] < 2^{-m\epsilon/2}$

Define the event $B' = [\exists h \in H ; h$ consist with S and $M(h, S') > \frac{m\epsilon}{2}]$.

Q3) Define & explain "shattering a set of finite hypothesis spaces

→ — shattering is a key notion in ML that refers to a classifier's capacity to accurately distinguish any arbitrary labeling of group of point.

— Strictly speaking, a classifier breaks a collection of points if it can divide them into all viable binary categories.

— the greatest no of points that a classifier is capable of shattering is specified by the the VC dimension

— When classifier is able to properly distinguish any potential labeled of the points, it is said to be "shattering" a collection of points.

Q4) Explain esemble learning model in detail.

→ — Esemble learning refers to a ML approach in which the predictions from multiple models are merged to enhance the accuracy of the ultimate forcast

— Simple Esemble Techniques

1) Max Voting
2) Averaging
3) Weighted Averaging

1) Max Voting :- This method generally

used for classification problems. In this technique, multiple models are used to make predictions for each data point.

2) Averaging :-
- This method is similar to max voting technique multiple prediction are made for each data point in averaging.
- In this method, we take an average of prediction from all models and use it to make the final prediction.

3) Weighted Average :-
- This is an extension of averaging method, All methods are assigned different weight defining the important of each model for prediction.

- Advanced Ensemble techniques
1) Stacking :-
- It uses predictions from multiple method to build a new model. It will make predictio test set

2) Blending :-
- It follows same approach that stacking following but uses only a holdout set from the train set to make predictions.

3) Bagging :-
— It combining the result of multiple model. to get a generlized result.

4) ~~Bootin~~ Boosting :-
— Boosting is a sequential process, where each subsequent model attempts to correct the error of the previous model.

Q5) Explain VC dimension.

→ — The Vapnik - Chervonenkis dimension, more commonly known as VC dimension, is model capacity measurement used in statistics and ML.
— It is termed informally as a measure of model's capacity, It is used frequently to guide the model selection process while developing ML applications.

- How to find VC dimension.
— let us consider binary classification model, which state that for all points (a, b), such that $a < x < b$, label them as 1, otherwise lable them as 0.

$h(x) = 1$ if $a < x < b$
$h(x) = 0$, otherwise
$(a, b) \in R^2$

we take two points m & n for these two points, there can be 2^2 distinct labels in binary classification. list are follows

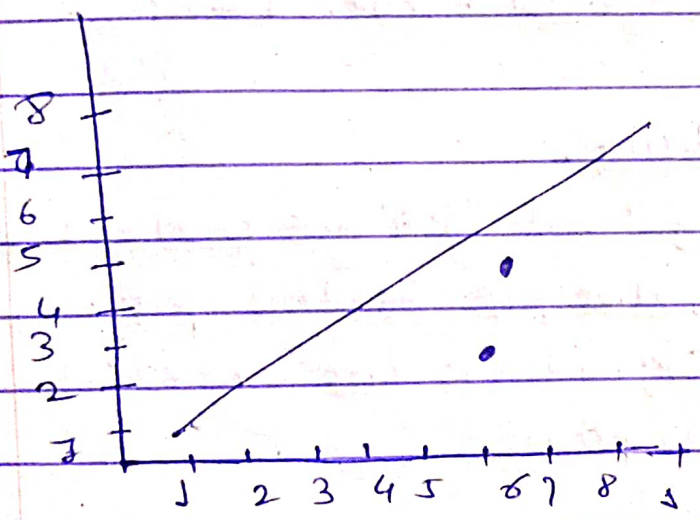
$h(m) = 0; h(n) = 0$

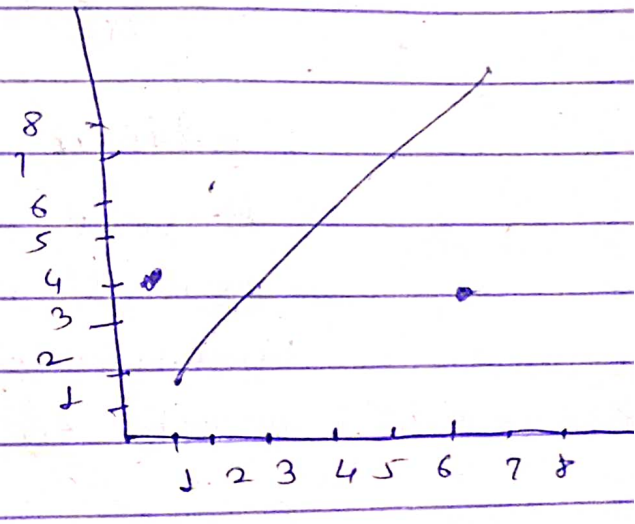
$h(m) = 0; h(n) = 1$

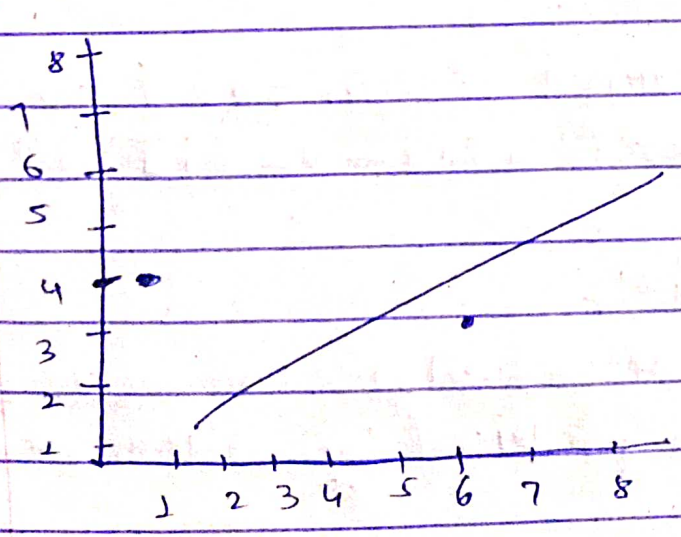
$h(m) = 1; h(n) = 0$

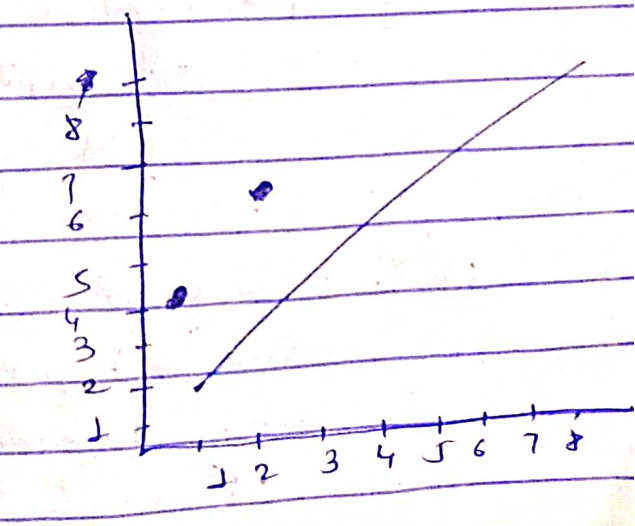
$h(m) = 1; h(n) = 1$.

- we can observe that for all the possible labelling variations of mm and nn.
- The model can divide the points into two segments.









Assigment No 5

Q1) Write note on :-

a) Manhattan distance

- Manhattan distance is distance measure that is calculated by taking the sum of distance between x and y co-ordinate.
- This distance also known as Manhattan length. In other word measured along axes at right angles.

formula :

Manhattan distance =

$$|x_1 - x_2| + |y_1 - y_2|.$$

b) Euclidean distance.

- It is a widly used distance metric. It works on the principle of the pythagoras theorm and signifies the shortest distance between two points.

formula :

Euclidean Distance = (sum for i to N (abs (V1[i] - V2[i]))^p)^(1/p)

c) Elbow method :-

- The elbow method plots the value of the cost function produced by different values of k.

Q2) - Explain Gaussian Mixture model in detail.

→ - Suppose there are K clusters so, μ & Σ and Σ are also estimated for each k. Had it been only one distribution, they would have been estimated by the maximum-likehood method.

- But since there are K such clusters and the probability ~~dett~~ density is defined as a linear function of densities of all these k distributions i.e

$$p(X) = \sum_{k=1}^{k} \Pi_k \cdot G_1(X|\mu_k, \Sigma_k).$$

where Π_k is mixing coefficient for k^{th} distribution, for estimating the parameters by the maximum log-liklihood method, compute $p(X|\mu, \Sigma, \Pi)$.

Q3) Explain Maximum likelihood & least square error hypothesis

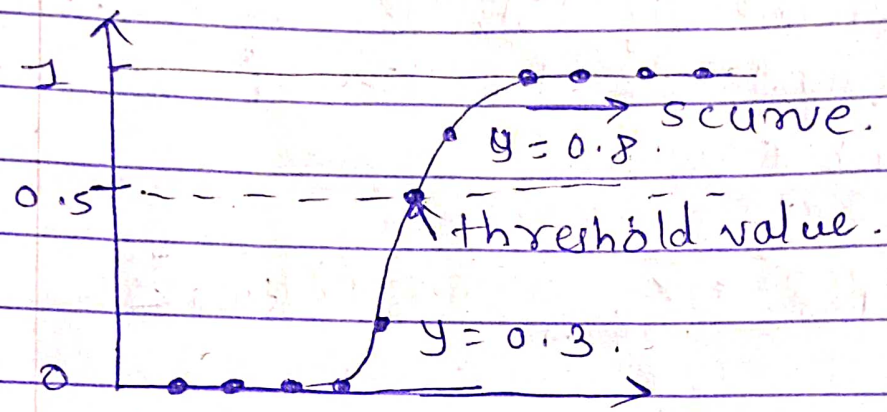
→ - Maximum Likelihood Estimation (MLE is a probabilistic based approach to determine values for the parameters of the model.

- Parameter could be defined as blueprint for the model because on the algorithm works. MLE is widly used technique in ML, ~~it is~~

-

-Least square is commonly used method in regression analysis for ~~estabtis~~ estimating the unknown parameters by creating a model will minimize the sum of squared error between the observed data & the predict data

e.g. of Maximum Likelihood Estimation.



Q4) Predict the class of new point X=1 & y=1 using KNN algorithm assume K=3 ?

→

X	y	class
-1	1	-
0	1	+
0	2	-
1	-1	-
1	0	+
1	2	+
2	2	-
2	3	+

$(x_1, y_1) = (1, 1)$ & $(x, y) = (-1, 1)$.

$$d_1 = \sqrt{(x_1 - x)^2 + (y_1 - y)^2}$$
$$= \sqrt{(1-(1))^2 + (1-1)^2}$$
$$= \sqrt{4}$$
$$= 2$$

$d_2 = (x_1, y_1) = (1, 1)$ & $(x, y) = (0, 1)$.

$$= \sqrt{(x_1 - x)^2 + (y_1 - y)^2}$$
$$= \sqrt{(1-(-1))^2 + (1-1)^2}$$
$$= 1$$

$d_3 = (x_1, y_1) = (1, 1)$ & $(x, y) = (0, -1)$.

$$= \sqrt{(x_1 - x)^2 + (y_1 - y)^2}$$
$$= \sqrt{1 + 1}$$
$$= \sqrt{2}$$

$d_4 = (x_1, y_1) = (1, 1)$ & $(x, y) = (1, -1)$.

$$= \sqrt{(x_1 - x)^2 + (y - y_1)^2}$$
$$= \sqrt{(1-1)^2 + (1-(-1))^2}$$
$$= \sqrt{4}$$
$$= 2$$

$d_5 = (x_1, y_1) = (1, 1)$ & $(x, y) = (1, 0)$

$$= \sqrt{(x_1 - x)^2 + (y_1 - y)^2}$$
$$= \sqrt{(1-1)^2 + (1-0)^2}$$
$$= 1$$

$d_6 = (x_1, y_1) = (1,1)$ & $(x, y) = (1,2)$

$= \sqrt{(x_1 - x)^2 + (y_1 - y)^2}$

$= \sqrt{(1-1)^2 + (1-2)^2}$

$= 1$

$d_7 = (x_1, y_1) = (1,1)$ & $(x, y) = (2,2)$

$= \sqrt{(x_1 - x)^2 + (y_1 - y)^2}$

$= \sqrt{(1-2)^2 + (1-2)^2}$

$= \sqrt{2}$

$d_8 = (x_1, y_1) = (1,1)$ & $(x, y) = (2,2)$

$= \sqrt{(x_1 - x)^2 + (y_1 - y)^2}$

$= \sqrt{(1-2)^2 + (1-3)^2}$

$= 2.4.$

d_1	2
→ d_2	1
d_3	1.414
d_4	2
→ d_5	1
→ d_6	1
d_7	1.414
d_8	2.414

predicted ~~value~~ class is positive all d_2, d_5 & d_6 are positive.

Q5) What is hierarchi clustering ? consider following distance matrix and applay hierarchical clustering to cluster u, v, w, x, y

	u	v	w	x	y
u	0	1	2	2	3
v	1	0	2	4	3
w	2	2	0	1	5
x	2	4	1	0	3
y	3	3	5	3	0

	uv	w	x	y
uv	0	2	2	3
w	2	0	1	5
x	2	1	0	3
y	3	5	3	0

d (w, (uv))
mind (d [wu], d [wv])
min (2 , 2)

d (x, uv)
mind (d [xu], d [xv])
min (2, 4)

d (y, uv)
mind (d [yu], d [yv])
(3, 3)

	u	v w	x	y
u v w	0		1	3
x	1		0	3
y	3		3	0

d (x, (u v w))

min d [d(xu), d(xv), d(xw))

min d (2, 4, 1)

d (y, (u v w))

min d [d (yu), d(yv) & d(yw)]

(3, 3, 5)

	u v w x	y
u v w x	0	
y		0

d (y, u v w x)

min d (d(yu), d(yv), d(yw) & d(yx))

min d (3, 3, 5, 3).