Worksheet-1

Machine Learning

Find –S Algorithm

Find a maximally specific hypothesis?

How many concepts are possible for this instance space?

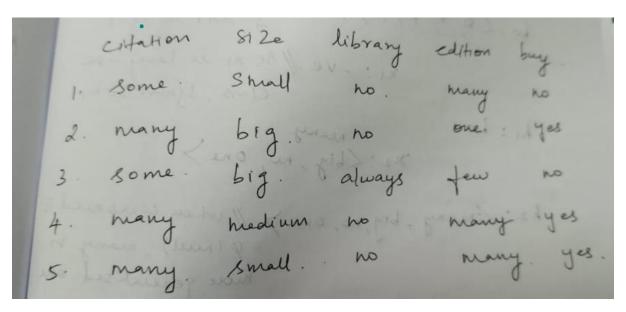How many hypotheses can be expressed by the hypothesis language?

How many semantically distinct hypothesis?

Question: 1

For practise (Eg.2)

Sky	Temp	humidity	forecast	Enjoy
	Warm	normal	same	Yes
Sunny	Warm	high	same	Yes
Sunny	cold	high	change	no
Rainy	cold	high	change	Yes
Sunny	warm	high	?	?

Question: 2

	citation	Size	library	edition	buy
1.	some	small	no	many	no
2.	many	big	no	one	yes
3.	some	big	always	few	no
4.	many	medium	no	many	yes
5.	many	small	no	many	yes

Candidate Elimination Algorithm

Find the General and Specific hypothesis for the given dataset

Question: 3

Size	Color	Shape	Class
Big	Red	Circle	No
Small	Red	Triangle	No
Small	Red	Circle	Yes
Big	Blue	Circle	No
Small	Blue	Circle	Yes

Question: 4

Citation	Size	In library	Price	Edition	Buy
Some	Small	No	Affordable	one	No
many	big	No	Expensive	many	Yes
many	medium	No	"	few	Yes
many	Small	No	Affordable	many	Yes

Question: 5

Shape	Size	Color	Surface	Thick	Target
Circular	large	light	Smooth	Thick	Malignant (+)
Circular	large	light	Irregular	Thick	"
Oval	large	dark	Smooth	Thin	Benign (-ve)
Oval	large	light	Irregular	Thick	Malignant (+ve)

PAC Learning

Question: 6

Find the hypothesis that is a good approximation of target concept f(x) using PAC learning to classify the given dataset as even and odd numbers. Train the model using the dataset,

$$(2, 1), (3, 0), (6, 1), (7, 0), (8, 1), (9, 0).$$

And test the model using the dataset

$$(4, 1), (5, 0), (10, 1), (11, 0).$$

Question: 7

You are tasked with learning a classifier for emails. The goal is to learn a function that predicts whether an email is spam or not spam based on various features of the email. These features could include the presence of certain words, the length of the email, the number of links in the email, etc. Calculate the sample complexity using PAC learning method. Assume the following values,

- **Error tolerance** $\epsilon = 0.05$ (we want the error rate on unseen data to be at most 5%).
- **Confidence level** $\delta = 0.01$ (we want 99% confidence).
- **Number of features** $k = 10$ (let's assume there are 10 features such as the frequency of certain keywords, the number of links, etc.).
- **VC Dimension** $d = 10$ (since we have 10 features, the VC dimension is 10).

Linear Regression

Formulate the line of equation for the following questions and find the following

(i)   Mean Squared Error(MSE)
(ii)   Mean Absolute Error (MAE)
(iii)   Root Mean Squared Error(RMSE)
(iv)   (iv) Standard Error

Question: 8

Problem Statement

A researcher collects data to study the relationship between the number of hours a student studies (X) and their test score (Y). The goal is to create a regression model to predict test scores based on study hours and validate the model.

Dataset

Study Hours (X)	Test Score (Y)
2	50
4	55
6	60
8	70
10	85

Question: 9

Question: Find linear regression equation for the following two sets of data:

x	2	4	6	8
y	3	7	5	10

Question: 10

Sales data for five weeks is given below.

X_i (item)	Y_i (actual sales in 1000s)
I_1	80
I_2	90
I_3	100
I_4	110
I_5	120

Multilinear Regression

Question: 11

ex:-

x_1	x_2	y
1	2	6
2	3	7
3	6	8

Question: 12

TV Ad x_1	Newspaper x_2 Ad	y (Sales)
60	22	140
62	25	155
67	24	159
70	20	179
71	15	192
72	14	200
75	14	212
78	11	215

Logistic Regression

Formulate Confusion matrix and calculate the following ,

(i)　　　Accuracy (ii) Precision (iii) Recall (iv) F1 score (v) Log-loss

Question: 13

We want to predict if a person has a disease ($y = 1$) or not ($y = 0$) based on the following numerical features:

1. Age (x_1)

2. Blood Pressure (x_2)

3. Cholesterol Level (x_3)

Person	Age (x_1)	Blood Pressure (x_2)	Cholesterol (x_3)	Disease (y)
1	25	120	180	0
2	45	140	200	1
3	35	130	190	0
4	50	150	220	1
5	60	160	240	1

Question: 14

We want to predict whether a stock price will go up ($y = 1$) or down ($y = 0$) based on three numerical features:

1. **Market Sentiment Score (x_1)**: Ranges from -1 (negative sentiment) to +1 (positive sentiment).

2. **Trading Volume (x_2)**: Number of shares traded in thousands.

3. **Previous Day's Price Change (x_3)**: Percentage change in stock price from the previous day.

Day	Sentiment Score (x_1)	Trading Volume (x_2)	Price Change (x_3)	Stock Up (y)
1	0.5	200	1.2	1
2	-0.3	180	-0.5	0
3	0.7	220	1.8	1
4	-0.1	150	-1.0	0
5	0.4	210	1.0	1

Question : 15

A bank wants to predict whether a loan application will be **approved** $(y = 1)$ or **denied** $(y = 0)$ based on applicant details.

1. **Income** (x_1): Monthly income of the applicant (in $).
2. **Credit Score** (x_2): Creditworthiness score (ranges from 300 to 850).
3. **Loan Amount** (x_3): Requested loan amount (in $).
4. **Loan Approved** (y): Target variable; 1 if approved, 0 if denied.

Applicant	Income (x_1)	Credit Score (x_2)	Loan Amount (x_3)	Loan Approved (y)
1	5000	720	20000	1
2	3000	650	15000	0
3	8000	800	25000	1
4	2500	600	12000	0
5	4000	680	18000	1