University of Asia Pacific

Department of CSE

Mid-Semester Examination, Fall 2020

Name: Rashik Rahman

Reg ID: 17201012

Year: 4th

Semester: 1st

Course Code: CSE 427

Course Title: Machine Learning

Date: 23.02.2021

"During Examination and upload time I will not take any help from anyone. I will give my exam all by myself."

University of Asia Pacific

Admit Card

Mid-Term Examination of Fall, 2020

Financial Clearance	PAID

Registration No : 17201012
Student Name : Rashik Rahman
Program : Bachelor of Science in Computer Science and Engineering

SI.NO.	COURSE CODE	COURSE TITLE	CR.HR.	EXAM. SCHEDULE
1	CSE 400	Project / Thesis	3.00	
2	CSE 330	Industrial Training	1.50	
3	CSE 401	Mathematics for computer Science	3.00	
4	CSE 403	Artificial Intelligence and Expert Systems	3.00	
5	CSE 404	Artificial Intelligence and Expert Systems Lab	1.50	
6	CSE 405	Operating Systems	3.00	
7	CSE 406	Operating Systems Lab	1.50	
8	CSE 407	ICTLaw, Policy and Ethics	2.00	
9	CSE 410	Software Development	1.50	
10	CSE 427	Topics of Current Interest	3.00	

Total Credit: 23.00

1. Examinees are not allowed to enter the examination hall after 30 minutes of commencement of examination for mid semester examinations and 60 minutes for semester final examinations.

2. No examinees shall be allowed to submit their answer scripts before 50% of the allocated time of examination has elapsed.

3. No examinees would be allowed to go to washroom within the first 60 minutes of final examinations.

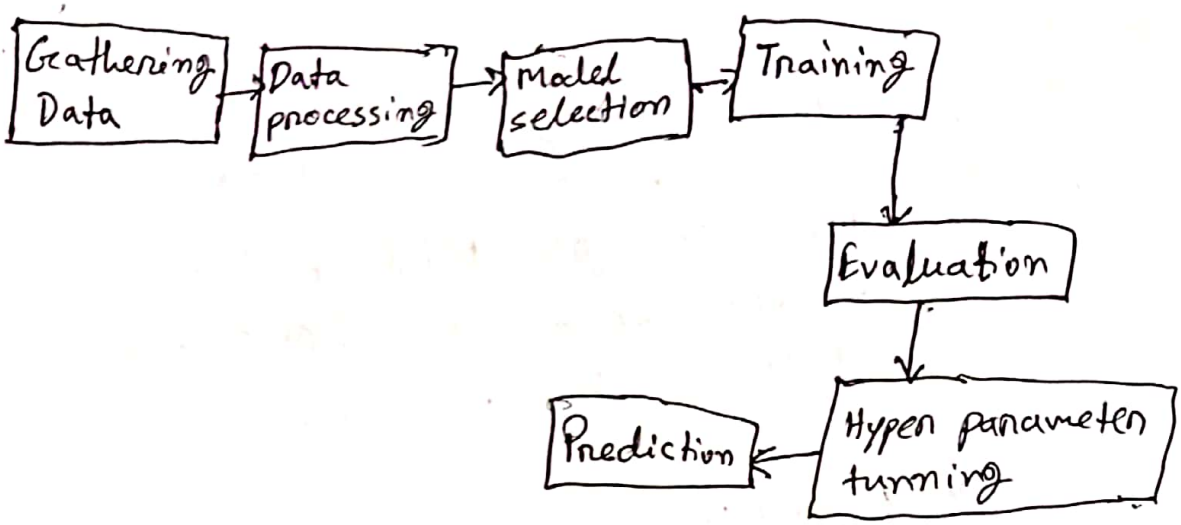
4. No student will be allowed to carry any books, bags, extra paper or cellular phone or objectionable items/incriminating paper in the examination hall.
Violators will be subjects to disciplinary action.

This is a system generated Admit Card. No signature is required.

Admit Card Generation Time: 21-Feb-2021 10:23 PM

Scanned with CamScanner

Answer to the Q.NO.1 (a)



i) Gathering Data: To train a model we need data and this data can be of any kind and needs to be collected. Mostly data are gathered by the means of web scrapping.

ii) Data processing: In this part it is said that 70% of a machine learning task's time is consumed. Here we need to clean and structure the unstructure data. We structure the data in such a manner that is a good input for model. We also do EDA here.

iii) Model selection: There are various models out there. No one knows which model performs well for a certain dataset. So we have to select a model that performs well with the given data. We can do this with the help of K fold cross validation, thus pick the best model among all the models.

Scanned with CamScanner

17201012

(2)

iv) Training: After selecting the model we train it with provided data.

v) Evaluation: We use different means like accuracy, logloss, RSME, of1 score, confusion matrix ~~to etc~~, ROC & AUC curve to evaluate the model.

vi) Hyper parameter tunning: We can use grid search CV to tune the parameters of the selected model thus achieve higher accuracy as we evaluate it again.

vii) Prediction: Finaly when the model is finished training & tunning then it is ready to predict for a fresh data.

Scanned with CamScanner

Answer to the Q.NO.1(b)

Machine learning problem are typicaly two types those are supervised and unsupervised problems. Supervised problem can be further specified in two types these are regression and classification problem. In classification problem the dependent attribute or we can say ~~labels has values~~ predicted values are discrete and we call them labels. Ex:

score1	score2	result
29	43	Pass
22	29	Fail
10	47	Fail
31	55	Pass
17	18	Fail
33	54	Pass
32	40	Pas
20	41	Pass

Here we can see that the dependent attribute result has discrete values of jus pass and fail. So in a classification problem a model learns from independent attributes the predicts a discrete value.

Scanned with CamScanner

17201012

(9)

Answer to the Q. NO. 2(a)

The basic ~~ba~~ of naive bayes ~~algorithm~~ algorithm is:

$$P(c|x) = \left(P(x_1|c) \cdot P(x_2|c) \cdot P(x_3|c) \cdots P(x_n|c)\right) \times \frac{P(c)}{P(x)} \quad \text{(i)}$$

Calculation of probabilities are;

if a condition B is given and we need to find probability of a condition A happening then we can say,

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)}$$

If we breakdown eqn (i) then we can understand the followings:

$P(c|x)$ = Posterior probability

$P(x|c)$ = Likelihood

$P(c)$ = ~~Prediction~~ Close prior probability

$P(x)$ = Prediction prior probability.

Scanned with CamScanner

Answer to the Q. NO. 2(b)

Freq table

Chills		
	Y	N
Y	3	1
N	2	2
Σ	5	5

running nose		
	Y	N
Y	4	1
N	1	2
	5	3

headache		
	Y	N
no	1	1
mild	2	2
strong	2	1
	5	3

feven		
	Y	N
Y	4	1
N	1	2
	5	3

Likelihood table

chills		
	Y	N
Y	3/5	1/3
N	2/5	2/3

running nose		
	Y	N
Y	4/5	1/3
N	1/5	2/3

headache		
	Y	N
no	1/5	1/3
mild	2/5	1/3
strong	2/5	1/3

feven		
	Y	N
Y	4/5	1/3
N	1/5	2/3

Now for

chills = Y

feven = Y

headache = midd

nose running = N

$P(Yes) = 3/5 \times 4/5 \times 2/5 \times 1/5 = 0.0384$

$P(NO) = \frac{1}{3} \times \frac{1}{3} \times \frac{1}{3} \times \frac{2}{3} = 0.025$

Overall probability:

$P(Yes) = 0.0384 \times 5/8 = 0.024$

$P(NO) = 0.025 \times 3/8 = 0.0094$

Scanned with CamScanner

17201012

(6)

So the answer will be the patient has high probability to have flu.

Answer to the Q. No. 7

Here we use; $-P_1 \log_2(P_1) - P_2 \log_2 P_2 - P_3 \log_2(P_3)$ to calculate gain;

Gain again of whole dataset,

$$\text{Gain(dataset)} = -\frac{4}{10}\log_2(4/10) - \frac{3}{10}\log_2(3/10) - \frac{3}{10}\log_2(3/10)$$

$$= -0.4(-1.32) - 0.3(-1.74) - 0.3(-1.737)$$

$$= 1.571$$

For Gender:

	Bus	Train	Car	total	$-P_1\log_2P_1 - P_2\log_2P_2 - P_3\log_2P_3$
Male	3	1	1	5	1.371
Female	1	2	2	5	1.522

Scanned with CamScanner

Info gain of Geende:

$= Gain(dataset) - \{\frac{5}{10} gain(Male) + \frac{5}{10} gain(Femal)\}$

$= 1.571 - \{0.5 \times 1.371 + 0.5 \times 1.522\}$

$= 1.571 - 1.447 = 0.125$

For canownenship

	Bns	Train	Cane	total	$-P_1 \log_2 P_1 - P_2 \log_2 P_2 - P_3 \log_2 P_3$
0	2	1	0	3	0.918
1	2	2	1	5	1.522
2	0	0	2	2	0

Info gain of canownenship

$= Gain(dataset) - \{\frac{3}{10} gain(0) + \frac{5}{10} gain(1) + \frac{2}{5} gain(2)\}$

$= 1.571 - \{0.3 \times 0.918 + 0.761\}$

$= 0.535$

Scanned with CamScanner

17201012

(8)

For travel cost

	Bus	train	Car	total	$-P_1\log_2P_1 - P_2\log_2P_2 - P_3\log_2P_3$
cheap	4	1	0	5	0.722
Standard	0	2	0	2	0
expensive	0	0	3	3	0

∴ Info gain of travel cost

$= Gain(dataset) - \{\frac{5}{10} gain(cheap) + \frac{2}{10} gain(standard) + \frac{3}{10} gain(expesive)$

$= 1.571 - 0.5 \times 0.722$

$= 1.21$

For Income level:

	Bus	train	car	total	$-P_1\log_2P_1 - P_2\log_2P_2 - P_3\log_2P_3$
Low	2	0	0	2	0
Medium	2	3	1	6	1.459
High	0	0	2	2	0

Scanned with CamScanner

∴ Info gain of Income level:

$\Rightarrow$ Gain(dataset) $- \{ \frac{2}{10}$ gain(Low) $+ \frac{6}{10}$ gain(medium) $+ \frac{2}{10}$ gain(High) $\}$

$= 1.571 - 0.2 \times 1.459$

$= 1.279$

Summary:

Info gain (Gender) = 0.125

Info gain (carownership) = 0.535

Info gain (travel cost) = 1.29

Info gain (Income level) = 1.279

Scanned with CamScanner