ASSIGNMENT / TEST BOOKLET

Student's Name : Vishwa M Singh

Class : Final Year B.Tech Division : A (AIML2) Roll No.: 1032170273 Academic Year : 2020-2021

Subject : AML Assignment / Test No. : 1 Date : 18/12/2020

PLEDGE

I solemnly affirm that I have written this Assignment/Test based on my own preparation. I have neither copied it from others nor given it to others for coping. I know that this is to be submitted as a part of my submission at the end of the term.

Vishwa
Signature of the student

Q. No.	1	2	3	4	5	6	7	8	9	10	Total	Name & sign of the faculty Member
Marks/Grade												

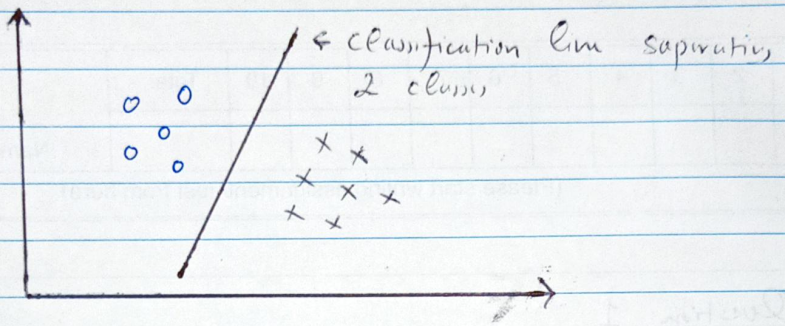
(Please start writing assignment/ test from here)

Question 1

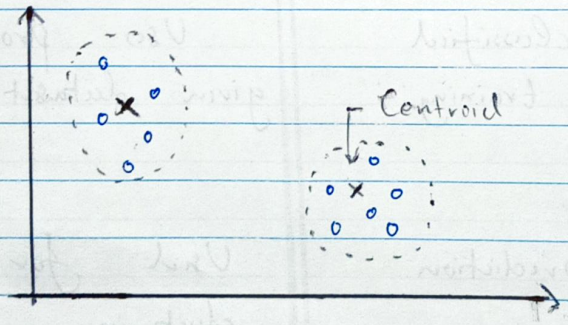
	Supervised	Unsupervised
1>	Input data has labels	Input data is unlabeled
2>	Data is classified based on training dataset.	Uses properties of the given dataset.
3>	Used for prediction	Used for analysis and clustering
4>	Broadly classified into Regression & classification	Broadly divided into clustering & association
5>	Eg: SVM(R, Decision tree, Logistic & Linear Regression, LDA	Eg: PCA, Apriori, K-means Clustering

Working Comparison :

In can of classification, model tries to calculate & reduce the classification error (or loss).



In clustering, we decide on a arbitrary no. of clusters & try to estimate the location of the centroid for each class.



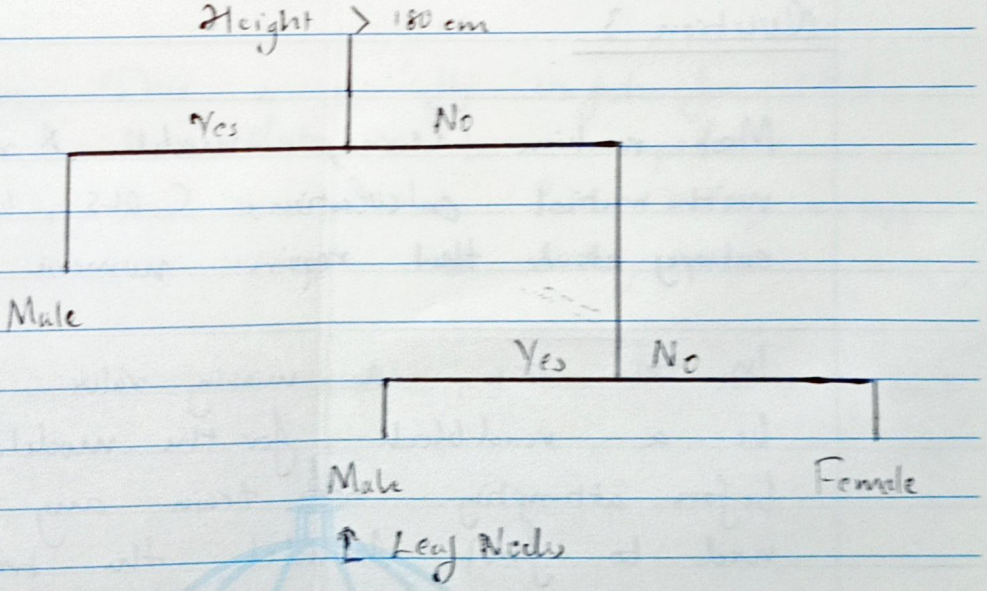
Question 2

CART : Classification and Regression Trees are decision tree based supervised learning algorithm.

CART model representation :

A ~~CART~~ CART model is a binary-tree. ~~from algorithm~~. Each Root represents a single input variable (x) and a split point on that variable.

The leaf nodes of the tree contain an output variable (y) which is used to make a prediction.



Notes :

- The splitting criterion is the entropy i.e. the measure of randomness. The splitting should minimize the entropy.

- Using a threshold minimum count on the number of training instances assigned to each leaf node.

- The performance can be boosted using ensembling (i.e. bagging) multiple decision tree. This can reduce variance drastically.

Question 3

Most machine learning model relies on mathematical calculations (OLS, back propagation entropy etc.) that require numeric inputs.

In this case, a missing value or 'NAN' can be a roadblock for the model. Therefore, before attempting to train any model, we need to first handel the ~~training etc~~ missing parts of the dataset.

Strategies:

1) **Droping ~~columns~~ rows with missing value:** This is the most common method. If the number of missing values are not too many, some rows can be droppied. However, this will not work in can of large count of missing values of or time series data.

2) **Imputer:** This refers to replacing the missing values with either mean, median or any other central tending including 0.

3) **Interpolation:** This is specially useful for time series data. In interpolation, we use methods like moving average to estimate the missing value.

Question 4

~~Naiv~~ Naive Bayes Classification: This is a supervised classification method which uses the Bayes theorem.

Bayes Theorem:

Likelyhood ↓ Prior Probability ↓

$$P(A|B) = P(B|A) \times P(A) / P(x)$$

↑ Posterior Probability (~~Poster~~)

↳ Marginal Likelyhood

In case of ~~the~~ classification, 'A' can be the class and 'B' is the record / features.

=> P(A|B) is the probability of something belonging to class A given it's features B.

Likelyhood calculation: This is done using the chain rule of probability

i.e. $P(A, B, C|Y) = P(A|B, C, Y) \times P(B|C, Y) \times P(C|Y)$

If we assume features A, B & C are independent

$P(A, B, C \mid Y) = P(A \mid Y) * P(B \mid Y) * P(C \mid Y)$

However, this assumption of independence is not true in majority cases due to the presence of some degree of multi-colinearity. Therefore the algorithm is called "naive"

Types:

* **Gaussian**: It is used in classification & it assumes that features follow a normal distribution

* **Multinomial**: It is used for discrete counts. Eg, in bernoulli trials which is one step further and instead of "word occuring in the document".

* **Bernoulli**: They are useful for binary feature ~~vector~~ vector (0 or 1). Example of this is the "bag of words" model

Question 5

=> **Feature Engineering**: It is the process of transforming raw data into features that better represents the underlying problem to the predictive models, resulting in improved model accuracy on unseen data.

Feature Engineering Leads to:
- Flexibility in model selection
- Simpler models
- Better results.

Techniques:

- **Imputation**: To handle missing data
- **Handling Outliers**: To remove any data that can harm the model.
- **Binning**: To convert continuous data to categorical features
- **One Hot Encoding**: To convert categorical features to numeric data
- **Feature Split**: To divide the data set into testing & training set
- **Scaling**: To bring the columns in the same range.