II - ASSIGNMENT

(Start Writing From Here)

1) a) Illustrate the k-means clustering algorithm with an example

b) Compare Agglomerative versus Divisive Hierarchical Clustering approaches.

A) a)

* K-Means clustering is an Unsupervised learning algorithm that is used to solve the clustering problems in data science (or) ML, which groups the unlabeled dataset into different clusters. Here K-defines the number of pre-defined clusters that need to be created in the process.

* It is a centroid based algorithm where each cluster is associated with a centroid. The main aim of this algorithm is to minimize the sum of distances between the datapoint and their corresponding clusters.

* The algorithm takes the unlabeled dataset as input divides the dataset into k-number of clusters and repeat the process until it doesnot find the best clusters.

Example:-

Cluster the following eight points (with (x, y) representing locations) into three clusters:

$A_1(2,10)$ $A_2(2,5)$ $A_3(8,4)$ $A_4(5,8)$ $A_5(7,5)$ $A_6(6,4)$ $A_7(1,2)$ $A_8(4,9)$ Initial cluster centres are $A_1(2,10)$ $A_4(5,8)$ $A_7(1,2)$ The distance function is Euclidian distance. Use k-means algorithm to show only the three cluster centres after the second iteration.

Sol) Euclidian distance

$$d(i,j) = \sqrt{(x_{i1}-x_{j1})^2 + (x_{i2}-x_{j2})^2 + \cdots (x_{in}-x_{jn})^2}$$

Given points	Distance from centre (2,10) of cluster 1	Distance from centre (5,8) of cluster 2	Distance from centre (1,2) of cluster 3	Point belongs to cluster
$A_1(2,10)$	0	$\sqrt{13}$	$\sqrt{50}$	1
$A_2(2,5)$	$\sqrt{5}$	18	$\sqrt{10}$	3
$A_3(8,4)$	$\sqrt{72}$	5	$\sqrt{53}$	2
$A_4(5,8)$	$\sqrt{13}$	0	$\sqrt{52}$	2
$A_5(7,5)$	$\sqrt{50}$	$\sqrt{13}$	$\sqrt{45}$	2
$A_6(6,4)$	$\sqrt{52}$	$\sqrt{17}$	$\sqrt{29}$	2
$A_7(1,2)$	$\sqrt{65}$	$\sqrt{52}$	0	2
$A_8(4,9)$	$\sqrt{5}$	$\sqrt{2}$	$\sqrt{58}$	2

Cluster-1 ⇒ A_1 → we have only one point cluster centre remains same

Cluster-2 ⇒ B_1, A_3, B_2, B_3, C_2 ⇒ centre for cluster-2

$((8+5+7+6+4)/5, (4+8+5+4+9)/5)$

$(6, 6)$

Cluster-3 ⇒ C_1, A_2

↳ centre for cluster-3

$((2+1)/2, (5+2)/2) = (1.5, 3.5)$

Iteration 2

Given Points	Distance from centre (2,10) of cluster 1	Distance from centre (6,6) of cluster 2	Distance from centre (1.5,3.5) of cluster 3	Point belongs to cluster
$A_1(2,10)$	0	$4\sqrt{2}$	6.51	C_1
$A_2(2,5)$	5	$\sqrt{17}$	1.58	C_3
$A_3(8,4)$	$6\sqrt{2}$	$2\sqrt{2}$	6.51	C_2
$A_4(5,8)$	$\sqrt{13}$	$\sqrt{5}$	5.7	C_2
$A_5(7,5)$	$5\sqrt{2}$	$\sqrt{2}$	5.7	C_2
$A_6(6,4)$	$2\sqrt{13}$	2	4.52	C_2
$A_7(1,2)$	$\sqrt{65}$	$\sqrt{41}$	1.58	C_3
$A_8(4,9)$	$\sqrt{5}$	$\sqrt{13}$	6.04	C_1

Cluster-1 ⇒ A_1, A_8 → centre for cluster-1 ⇒ $((2+4)/2, (10+9)/2) = (3, 9.5)$

Cluster 2 ⇒ A_3, A_4, A_5, A_6 → centre for cluster 2

$((8+5+7+6)/4, (4+8+5+4)/4) = (6.5, 5.25)$

Cluster 3 ⇒ A_2, A_7

↳ centre for cluster 3

$((2+1)/2, (5+2)/2) = (1.5, 3.5)$

After second iteration the centre of three clusters are
C1(3, 9.7) C2(6.5, 5.25) C3(1.5, 3.5)

b)

* Divise clustering is more complex as compared to agglomerative clustering, as in the case of divise clustering we need a flat clustering method as "subrouting" to split each cluster until we have each data having its own singleton cluster.

* A Divisive Algorithm is more accurate. Agglomerative clustering makes decision by considering the local patterns or neighbor points without initially taking into account the global distribution of data. These early decisions cannot be undone. where as divise clustering takes into consideration the global distribution of data when making top level partioning decisions.

* Divise Clustering is more efficient if we don't generate a complex heirarchy all the way down to individual data leaves.

* Agglomerative is bottom-up approach. Each observation starts in its own cluster and pairs of cluster are merged as move up the heirarchy

* Divislue is top-down approach. All observations start in one cluster

and splits are perform recursively as move down the hierarchy.

2) Define Hypothesis? Explain the Null and Alternative Hypothesis with an example?

A) Hypothesis:-

* Hypothesis is defined as a formal statement, which gives the explaination about the relationship between the two (or) more variables of the specified population.

* It helps the researcher to translate the given problem to a clear explaination for the outcome of the study.

Null hypothesis

* In the null hypothesis there is no significant difference blw the populations specified in the experiments due to any experimental (or) sampling error. It is denoted by H_0.

Alternative Hypothesis:

The simple observations are easily influenced by some random cause.

* It it denoted by H_1.

Eg: Determine about chi-square test with following example

$\alpha = 0.05$ (tabular value = 15.0861)

Coin	1	2	3	4	5	6	Total
Freq	22	24	38	30	46	44	204

Observed table

Sol) Degree of freedom = (rows − 1)(columns − 1) = (6 − 1)(2 − 1) = 5

$E(1) = 204 \times \frac{1}{6} = 34$

# Coin	1	2	3	4	5	6
Expected	34	34	34	34	34	34

Expected table

$\chi^2 = \sum \frac{(O-E)^2}{E}$

Expected table

O	E	O − E	$(O-E)^2$	$(O-E)^2/E$
22	34	−12	144	4.23
24	34	−10	100	2.94
38	34	4	16	0.47
30	34	−4	16	0.47
46	34	12	144	4.23
44	34	10	100	2.94
				15.29

$\chi^2_{cal} = 15.29$ $\chi^2_{tab} = 11.07$

$15.29 > 11.07$ $(\alpha = 0.05)$

$15.29 > 11.07$ $(\alpha = 0.01)$

Hence H_1 is accepted

3) list out the type of variables in R

A1 * A variable is a name given to a memory location which is used to store values in a computer program.

* Variables in R programming can be used to store numbers, words, matrices and Tables.

* For a variable to be valid
  i) It should contain letters, numbers and only dot (or) underscore characters.
  ii) It should not start with a number, start with a dot followed by a number.
  iii) It should not start with an underscore and not be a reserved keyword.

* R doesnot have a command for declaring a variable A variable is created the moment you first assign a value to it. To assign a value to a variable, use the ← sign. To output (or print) the variable value, just type the variable name.

Eg: name <- "John"
age <- 40
name #output "John"
age #output 40

name <- "John Doe"
print(name) #print the value of the name variable
O/p: John Doe

Multiple variable,
var1 <- var2 <- var3 <- "Orgoange"

4) Explain Maximum likelihood test in R with example.

A) Maximum likelihood

* The goal of maximum likelihood is to fit an optimal statistical distribution to some data.

* This makes the data easier to work with, makes it more general, allows us to see if new data follows the same distribution as the previous data and lastly it allows us to tell if (new) data classify unlabelled data points.

* likelihood is defined as the probability, giving a model and a set of parameter values of obtaining a particular set of data.

* To calculate a likelihood we have to consider a particular model that may have generated the data That model will almost always have parameter values that need to be specified. We can refer to this specific model as a hypothesis H, The likelihood is then

$$L(H/D) = Pr(D/H)$$

Here L → Likelihood

Pr → Probability

D → Data

H → hypothesis

Example: We need to calculate the likelihood as the probability of obtaining heads 63 out of 1000 lizard flips, given some model of lizard flipping.

We can write the likelihood for any combination of H "successes" (flips that gives heads) out of n trails. We will also have one parameter p_H which will represent the probability of "success" that is the probability that any one flip comes up heads. We can calculate the likelihood of our data using the binomial theorem

$$L(H/D) = Pr(D/p) = \binom{n}{H} p_H^H (1-p_H)^{n-H}$$

$$L(H/D) = \binom{100}{63} p_H^{63} (1-p_H)^{37}$$

5) Define Regression & Poisson Regression in R with example.

A) Regression :-

Regression is a technique for investigating the relationship between the independent variables (or) features and a dependent variable (or) outcome.

* Regression is a key element of predictive modelling, so can be found with in many different applications of machine learning.

Poisson Regression

* These models are used for modeling events where the outcomes are counts

* Count data is a discrete data with non-negative integer values that count things such as no of people in line at the grocery store

* Poisson regression allows us to determine which explanatory variable (x values) influence a given response variable (y value, count, or a rate).

Eg:- Poisson regression can be implemented by a grocery store to understand better and predict the no of people in a row

There is following general Mathematical equation for poisson Regression

Sno	Parameter	Description
1	y	It is the response variable
2	a and b	These are the numeric coefficients
3	y	x is the predictor variable

* The poisson regression model is created with the help of the familiar function glm().

Eg:- In this example we have considered an in-built dataset "warpbreaks" that describe the tension (low, medium, high) and the effect of wool type (A and B) on the number of warp breaks per loom. We will consider wool "type" and "tension" as the predictor variables, and "breaks" is taken as the response variable.

```
reg_data <- warpbreaks  ---print(head(reg_data))
glm(formula = breaks ~ wool + tension, data = warpbreaks,
    family = poisson)
output <-
```

print(summary(output))

O/P :-

	break	wool	tension
1	26	A	L
2	36	A	L
3	54	A	L
4	25	A	L
5	70	A	L
6	72	A	L