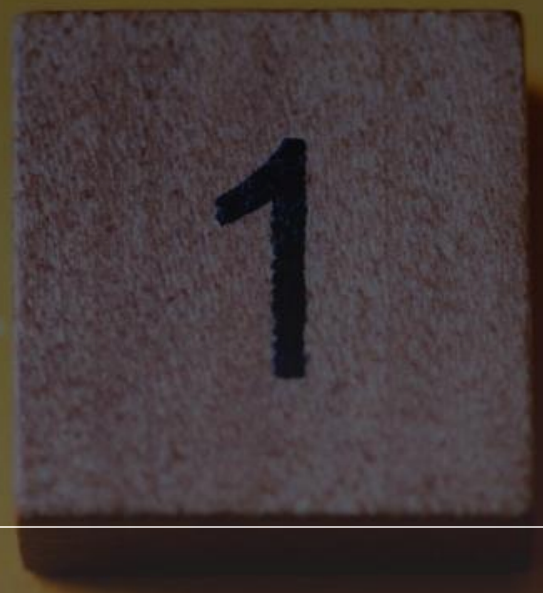
Week 8



ASSIGNMENT 8

Consider the dataset "USArrests.csv". Answer questions 1 to 4 based on the information given below

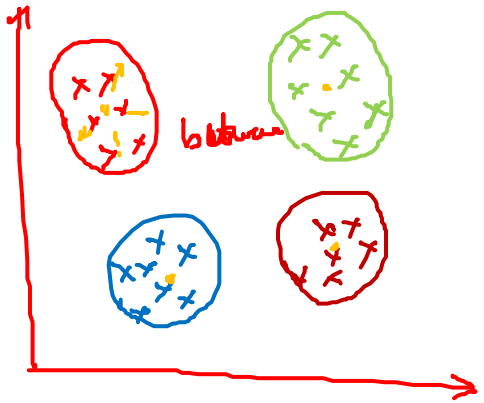
This data set contains statistics, in arrests per 100,000 residents for assault, murder, and rape in each of the 50 US states in 1973. Also given is the percent of the population living in urban areas.

Variables	Description
States	The state where the incident occurred
Murder	No. of arrests for murder (per 100,000 residents)
Assault	No. of arrests for assault (per 100,000 residents)
UrbanPop	Percentage of urban population
Rape	Rape arrests (per 100,000 residents)

- Set the column "States" as index of the data frame while reading the data
- Set the random number generator to set.seed(123)
- Normalize the data using scale function and build the K-means algorithm with the given conditions:
    - o number of clusters = 4
    - o nstart=20

1) According to the built model, the within cluster sum of squares for each cluster is _____(the order of values in each option could be different):-

- [x] 8.316061 11.952463 16.212213 19.922437
- [ ] 7.453059 12.158682 13.212213 21.158766
- [ ] 8.316061 13.952463 15.212213 19.922437
- [ ] None of the above



Answer: a

Solution:

```
> data=read.csv("USArrests.csv",header = T,row.names  ="States")
> df <- scale(data)
> set.seed(123)
> fit<-kmeans(df,centers = 4,nstart=20)
> print(fit$withinss)
[1]  8.316061 11.952463 16.212213 19.922437
```

2) According to the built model, the size of each cluster is _____ (the order of values in each option could be different):-

- 13 13 7 14
- 11 18 25 24
- 8 13 16 13
- None of the above

Answer: c

Solution:

```
> print(fit$size)
[1]  8 13 16 13
```

3) The Between Cluster Sum-of-Squares (BCSS) value of the built K-means model is _____(Choose the appropriate range)

- 100 - 200
- 200 - 300
- 300 – 350
- None of the above

Answer: a

Solution:

```
> print(fit$betweenss)
[1] 139.5968
```

4) The Total Sum-of-Squares value of the built k-means model is_____ (Choose the appropriate range)

- 100 - 200
- 200 - 300
- 300 – 350
- None of the above

Answer: a

```
> print(fit$totss)
[1] 196
```

5. Which of the statement is INCORRECT about *KNN* algorithm?
    a. KNN works ONLY for binary classification problems
    b. If k=1, then the algorithm is simply called the nearest neighbour algorithm
    c. Number of neighbours (K) will influence classification output
    d. None of the above

Answer: a

Solution:

The statement in option A is incorrect because, **KNN works for multi label classification problems** as well.

6. K means clustering algorithm clusters the data points based on:-
    a. dependent and independent variables
    b. the eigen values
    c. distance between the points and a cluster centre
    d. None of the above

Answer: c

Solution:

K means **clustering** classifies variables based on the distance between the points and a cluster centre

7) The method / metric which is NOT useful to determine the optimal number of clusters in unsupervised clustering algorithms is

- Scatter plot
- Elbow method
- Dendrogram
- None of the above

Answer: a

Solution:

The method used to determine the optimal number of clusters in unsupervised clustering algorithms are Dendrogram and Elbow method

8) The unsupervised learning algorithm which aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest centroid is

- Hierarchical clustering
- K-means clustering
- KNN
- None of the above

Answer: b

Solution:

K-means algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible

Practice Questions

1) The most commonly used distance metric to calculate distance between centroid of each cluster and data points in K-means algorithm is

- ○ Chebyshev distance
- ○ Manhattan
- ○ Euclidean distance
- ○ None of the above

Ans: Euclidean Distance

2) K means clustering classifies variables based on :-

- ○ dependent and independent variables
- ○ the eigen values
- ○ distance between the points and a cluster centre
- ○ None of the above

Ans: Distance between the points and a cluster centre

3) Which of the following statement is **NOT TRUE** about kNN algorithm?

- ○ It is a non-parametric algorithm
- ○ It is an instance based learning algorithm
- ○ Explicit training and testing phases are involved while implementing kNN
- ○ None of the above

Ans: Explicit training and testing phases are involved while implementing kNN

4) Scaling is important in distance-based algorithms because

- ○ variables with higher magnitude will influence the results more
- ○ distance calculations are affected by the magnitude of the variables
- ○ The data is large and has many features
- ○ Both A and B

Ans: Variables with higher magnitude will influence the results more

5) Which of the following is TRUE with respect to K means clustering?

- ○ The inter cluster distance is minimized and intra cluster distance is maximized
- ○ The inter cluster distance is maximized and intra cluster distance is minimized
- ○ Both inter cluster and intra cluster distance are minimized
- ○ Both inter cluster and intra cluster distance are maximized

Ans: The inter cluster distance is maximized and intra cluster distance is minimized