ADS4293 LAB9

Chapter 14: Unsupervised Data Mining

TA: Malak Al-Hanafi

November 13 , 2025

LAB SETUP

- Excel file **Ch14** dataset
- Submission:
  - Submit your updated **Excel file** on Brightspace (Owl)

QUESTION 1

3. FILE ***Exercise_14.3.*** Perform agglomerative clustering on the accompanying data file. Include all seven variables, first standardized to *z*-scores, for the analysis. Choose Euclidean for the distance between observations and complete linkage for the distance between clusters. Plot and inspect the dendrogram. How many clusters are generated if the minimum distance between clusters is 5? How many observations are in the largest cluster? What is the average value of x_1 of the largest cluster?

A.1 IN R

1. **Install & Load packages:**
   library(readxl)
   library(cluster)

2. **Import the data and save it into *“myData”***

3. **Standardize (z-scores) using *“scale”* and save it into *“myData1”***

4. **Compute distances using *“myData1”* , method = "euclidean" and save it into *“d”*.** * hint: use dist() function

5. **Run agglomerative clustering on *“d”* and save it into *“aResult”***
   **hint: use agnes() function*
   diss = TRUE tells R that d is a distance object, method = "complete" uses complete linkage

6. **Print *“aResult”*,** you will get an Agglomerative coefficient, comment on that number
   **hint: (closer to 1 = stronger clusters)*

7. ***Plot “aResult” (banner + dendrogram)***
   ** R will show two plots when you hit Enter: Dendrogram is the important one, it is a Tree diagram that shows how clusters are merged step-by-step as the distance threshold increases.*

to be continued on next slide...

A.1 IN R

8. ***Choose k (number of clusters)***

** The question says: "minimum distance between clusters is 5". So we draw an imaginary horizontal line at height=5 → this intersects the dendrogram in three places, so k = ?? clusters.*

** use cutree function (takes a hierarchical tree and assigns each observation to one of k clusters)*

aClusters <- cutree(aResult, k = ??)

9. ***Attach clusters to data myData <- data.frame(myData, aClusters)***

** This adds a new column aClusters to myData so each row knows which cluster it belongs to.*

10. ***Check cluster sizes: summary(as.factor(aClusters))***

** Which cluster is the largest (has the most observations)? 1, 2, or 3? Comment on this*

11. **Look at the largest cluster** (step 9) **and compute the mean of x1: summary(subset(myData, aClusters == ??))**

Using all these steps, make sure you answered:

1. How many clusters when minimum distance between clusters is 5?
2. How many observations in the largest cluster?
3. Average value of x1 in the largest cluster?

QUESTION 2

21. FILE *Internet_Addiction2.*  ,  Internet addiction has been found to be a widespread problem among university students. A small liberal arts college in Colorado conducted a survey of Internet addiction among its students using the Internet Addiction Test (IAT) developed by Dr. Kimberly Young. The IAT measures three underlying psychometric factors: emotional/psychological conflicts (Factor1), time management issues (Factor2), and mood modification (Factor3). Individuals who score high on the IAT are more likely to have problematic Internet use. The accompanying data file contains the scores for the three psychometric factors for 350 students. The respondents also provided a couple of pieces of demographic data including sex (equals 1 if female, 0 otherwise) and level of study (Graduate, equals 1 if graduate study, 0 otherwise).

    a. Perform hierarchical clustering to group the 350 university students using the three psychometric factors as well as the other demographic variables. Use the Gower's similarity coefficient for the distance measure and Ward's clustering method to cluster the students into three clusters. How many students are in the largest cluster?

    b. What are the average scores for emotional/psychological conflicts, time management issues, and mood modification for students in the largest cluster?

A.2 IN R

1. **Install & Load packages:**
   library(readxl)
   library(cluster)

2. **Import the data and save it into *“myData”***

3. **Because this dataset has mixed variables (numeric + categorical), we compute distances using Gower similarity**
   ** hint: use daisy() function with metric = "gower"*

4. **Run agglomerative clustering on “d” and save it into “aResult”** * hint: use agnes() function
   ** method = "ward" performs Ward’s minimum variance clustering (as stated in question)*

5. ***Print “aResult”, you will get an Agglomerative coefficient,*** *comment on that number* **hint: (closer to 1 = stronger clusters)*

6. **Plot “aResult” (banner + dendrogram)**
   ** R will show two plots when you hit Enter: Dendrogram is the important one, it is a Tree diagram that shows how clusters are merged step-by-step as the distance threshold increases.*

7. ***Choose k=3 (from question): aClusters <- cutree(aResult, k = 3)***

to be continued on next slide...

A.2 IN R

8. ***Attach clusters to data myData <- data.frame(myData, aClusters)***
   ** This adds a new column aClusters to myData so each row knows which cluster it belongs to.*

9. ***Check cluster sizes: summary(as.factor(aClusters))***
   ** Which cluster is the largest (has the most observations)? 1, 2, or 3? Comment on this*

10. **Look at the largest cluster** (step 9) **and compute Factor1, Factor2, Factor3 means:**
    summary(subset(myData, aClusters == ??))

Using all these steps, make sure you answered:

a. How many students are in the largest cluster?
b. What are the mean values of Factor1, Factor2, Factor3 in that cluster?

Thank you!