Exam 1.

Your Name

2022-09-21

This is an exam. You cannot discuss it with anybody. You should do the work yourself. Questions 1, 2 and 3 can be done with pen and paper. If you choose so, you can submit the scans along with other material (you also can embed them to your knitted pdf).

The submitted files must include pdf-s with your answers along with all R scripts. For example:

- Student A submitted:
  - Exam_1.pdf - final report containing all answers
  - Exam_1.Rmd - R-markdown files with student solutions
- Student B submitted:
  - Exam_1_Q1_Q2.pdf - scanned paper answers to questions 1 and 2
  - Exam_1_Q3_Q4.pdf - answers to questions 3 and 4
  - Exam_1.Rmd - R-markdown files with student solutions

No pdf report - no grade. If you experience difficulties with knitting, combine your answers in Word and any other editor and produce pdf-file for grading.

No R scripts - 50 % reduction in grade if relative code present in pdf- report, 100% reduction if no such code present.

Late submissions will result in a points reduction: 10% first hour, 50% second hour, 100% six hours. Only valid and documented reason to miss the exam or be late is acceptable. Sleeping in, lack of preparation, ennui, grogginess, inability to knit, etc. are not acceptable excuses. University policy allows multiple midterm exams on same day.

Reports longer than 25 pages are not going to be graded.

Exam must be submitted through UBLearns by 11:59pm 2022 October 14th.

Question 1. Clustering with Pen and Paper (25 Points).

Consider following 6 points:

	1	2	3	4	5	6
x	1	4	3	4	3	5
y	1	1	4	5	7	7

Q1.1 Perform single K-Means clustering with **Manhattan distance** using pen and paper. Show you work in readable manner. (20 points)

You can use rmarkdown instead of paper, simple calculator or simple vector calculations in R.

> You can set points as p1 <- c(1,1), p2 <- c(4,1) and use regular math operation on them (+-*/).
> You also can use math function like sum, abs and so on and output function like print and cat

You will need a random sample without replacement, either of size two or six, use one generated below:

5	6	3	4	1	2

```
hint:
If you used pen and paper you can take a photo of it and add to Rmd using
following markdown command:
![Note](filename.jpeg)
```

Q1.2 One of the students decided to check himself and make following clustering with R (5 points):

```
## [1] 1 1 2 2 2 2
```

That is the student got that first two points belong to first cluster and 4 other points to second cluster. But the students results in Q1.1 are different. Can you explain why?

> hits: in general there are two reasons besides that student mess-up in Q1.1. assume that student is correct in Q1.1. Out of thouse two reasons only one actually have effect in this small example. Though if it is done not properly can play too.

Question 2 (25 Points)

Q2.1 Just like in Q.1.1 but this time perform hierarchical clustering with complete linkage on same data. (20 points)

> After building whole tree, don't forget to cut it to get two clusters. Draw dendogram by hand (show proper hights) or using ascii graphics

Q2.1 Do you expect hclust provides same answer as yours? Check it. (5 points)

Question 3 (10 points)

Compare and contrast k-means and hierarchical clustering in terms of input, output, speed, cluster characteristics/ability to separate manifold structures.

Question 4 (40 points)

Clustering is often used to reexamine existing classifications. Sometime misclassification can occur in the original design. For example, one class might consist of two sufficiently different groups, or two classes are essentially the same.

In this exercise, you are presented with a seed dataset containing measurements from multiple wheat subspecies. Your task is to identify a number of subspecies.

`seeds.csv` constists of 200 records and reports 7 measurements for each seed:

- `length` - length of kernel
- `width` - width of kernel
- `asymmetry` - asymmetry coefficient
- `groove` - length of kernel groove
- `area` - area A
- `perimeter` - perimeter P
- `compactness` - compactness C = $4pi$A/P^2

Q4.1 Read and Visualize Dataset (5 points)

```
# read dataset
```

```
# Examine dataset, for example by plotting it with GGally:ggpairs
```

Comment on observation, any distinct clustering?

Q4.2 Run PCA (8 Points)

Q4.2.1 Run PCA (don't forget to scaled data), save as `pc_out`, we will use `pc_out$x[,1]` and `pc_out$x[,2]` later for plotting

Q4.2.2 Make scree plot/percentage variance explained plot. Comment on percentage variance explained (will two first components cover enough varience in dataset)

Q4.2.3 Make biplot. Comment on biplots (clusters?, possible meaning of PC?)

Q4.3 Perform Clustering of your choise. Select number of clusters. (20 points)

looks like k-means should do the job, biplot has nice packed shape.

Q4.4 Characterize clusters. (5 points)

Q4.5 How many subspecies you think are in the set? (does it matches conclussion from Q4.3?) (2 points)