Homework 4

1002/M91, Data Mining, Spring 2022

Enjoy your last HW!

Due date: 05.06.2022 23:59

REQUIREMENTS

- File format: For this Homework, you are required to submit both R Markdown and **PDF** files with your answers and codes in it. Make sure that Rmd file works, so that there won't be any errors when it is run and represent the same information as **PDF**. Under each question (not in comments) write the code along with your interpretations. Be sure to put your name at the top of your assignment (in the YAML header in front of the author).
- For installing **tinytex** and having PDF output your can use the following link.
- Rule of thumb: If the number of data points is greater than 50, do not print the whole data. Use subsets. Try to show all outputs (do not just store an object as a variable). Also, try to avoid using the same name for variables in the file.
- Cheating: The purpose of tasks is to check your knowledge (rather than the ability of thinking). Please, try to solve tasks without googling every exercise. Try not to discuss HW with your classmates and work only on your file. Any similarities, which can be considered ass cheated, will not be graded.

  Good luck!

Problem 1 (30 ptoints)

(a) What are the differences among exclusive, overlapping and fuzzy clustering? Bring (create your own) an example of fuzzy clustering with $k = 2$. Use the function `funny()` from library `cluster` and data visualization techniques from package `factoextra` to show your results. Show the membership matrix. Which of your observations belongs to both clusters?

(b) Suppose we have an example of a data set with 20 observations. We need to cluster the data using the K-means algorithm. After clustering using $k = 1, 2, 3, 4$ and 5 we obtained only one non-empty cluster. How is it possible?

(c) Suppose we have an example of a data set consisting of three natural circular clusters. These clusters have the same number of points and have the same distribution. The centers of clusters lie on a line, the clusters are located such that the center of the middle cluster is equally distant from the other two. Why will not *bisecting* K-means find the correct cluster?

Problem 2 (30 points)

Consider the following dataset:

Table 1: Dataset to Perform K-Means

	Variable.1	Variable.2
1	2	5
2	2	3
3	8	3
4	0	4
5	7	5
6	0	7
7	1	5
8	7	3
9	3	7
10	9	5

The goal of this task is to perform K-means clustering via R (manually $\ddot{\smile}$), with $k = 2$, using data with 2 features from the table above. Follow the step above:

(a) Neatly plot the observations using `ggplot`.

(b) **Randomly** assign a cluster label to each observation. You can use the `sample()` command in *R* to do it. Report the cluster labels for each observation.

(c) Define the coordinates of the centroids for each cluster. Show your results.

(d) Assign each observation to the cluster using the closeness of each observation to the centroids, in terms of Euclidean distance. Report the cluster labels for each observation.

(e) Repeat (c) and (d) until the centroids remain the same. You can use loops for this task.

(f) Show the observations on the plot by coloring them according to the clusters obtained. Show centroids on the plot.

Problem 3 (40 points)

For this task you need to download World Value Survey (Wave 6) data and try to understand the disposition of our country among others based on some criterias. The description of the variables and the survey are given with a separate file. Here is the link to obtain more information: . Choose the subset[1] from *Wave 6* data to perform the cluster analysis. Note that you need to use meaningful selections both of variables based on some topic/problem[2] and countries[3].

(a) Describe thoroughly how and why you choose your subset of variables and observations. What is your goal? Hint: You need to prepare data for the next step.

(b) Use all (appropriate) tools/functions from our lecture to cluster the countries (both nested and untested algorithms). Interpret them.

(b1) Is your hierarchical clustering stable regards to between clusters distance measures?

(b2) Compare the results obtained from two different k-means.

(c) Make the conclusion (also based on cluster centers).

[1] The dataset is too large to consider entirely.
[2] For example the level of conservatism of the coutry.
[3] For example post-soviet countries.

Bonus 1 (20 points)

Show that the average pairwise distance between the points in a cluster is equivalent to the SSE of the cluster.