CAP 5768: Intro to Data Science, Fall 2023

# Homework 2: Python and R

**Due Sep 28 via Canvas by 6:00 PM. Submit file named:** YourLastName-HW2.pdf

1. (15 points) Carefully study the Python notebook called MovieLens1M.ipynb. The data file required to run this notebook is in the ml-1m directory under data provided to you on Canvas. After you have understood the code in the notebook, modify the notebook to answer the following questions. You can do the work in Python or R. Submit the modified Python/R notebook.
2. Each movie in the database has a genre (comedy, animations, etc.) associated with it. Show the top 20 genres with the highest number of responses from users.

A screen shot of a computer

Description automatically generated

1. Show the top 20 genres sorted by average ratings.

A screenshot of a computer

Description automatically generated

1. Show the top 20 movies sorted by descending mean female ratings for a specific genre (say Drama).

A screenshot of a computer

Description automatically generated

1. (**Extra Credit**) If you submitted a Python solution for the above problem, then rewrite the code in R and submit the R notebook. (Obviously, if you submitted an R notebook for the above problem, then rewrite and submit a Python notebook.) Make sure your program runs properly and gives the same answer as the Python program it emulates. Submit the R notebook as a separate file. The name of all files you submit should include your name. For example, my files would be named GiriNarasimhan HW2.pdf and GiriNarasimhan HW2.Rmd
2. (10 points) In class, we discussed the MDCPS data set (see 5.3-MDCPS-Grades-2017.ipynb).

In class, we discussed clustering of the data set based on 4 features English Language Arts Achievement, Social Studies Achievement, Mathematics Achievement, Science Achievement (see 7.2-Clustering.ipynb). Study K-Means options from:

https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans. html. Redo the K-Means clustering by choosing (a) n init = 8, and (b) elkan algorithm instead of lloyd. Remember to set the seed to achieve reproducibility of your results.

A screenshot of a computer screen

Description automatically generated

I’n not sure if I was supposed to have the same clusters when changing the algorithm and n\_init but the only difference I can see is in execution time.

1. (10 points) For the above problem, implement a reasonable quality measure to compare the different clusterings above. Display your results.

As mentioned before, the only difference the output gives is in execution time and shows that the Elkan algorithm is more efficient when finding the clusters. I tried silhouette score but it still gave the same average score.

Silhouette:

Default:

A computer screen shot of a program code

Description automatically generated

Elkan and n\_init=8: (even separated as elkan in one run and n\_init=8 in another still showed same results)

A screenshot of a computer program

Description automatically generated

Timed comparison:

For Default:



For elkan algorithm and n\_init=8



1. (10 points) In class, we discussed 6.1-PCA.ipynb, where we ran PCA on the MDCPS data set. Follow it up by doing K-Means clustering on the points transformed by PCA (i.e., using pca 0 and pca 1). State and explain your parameter choices. Display your results by coloring points according to their cluster.

Parameters Chosen:

Components for PCA = 2 as permutation test showed only the first two components are relevant

Clusters = 3 as elbow method showed either 3 or 4 should be good enough clusters

Algorithm = 'LLoyd' as suggested in sklearn Elkan algorithm is efficient only with datasets with well-defined clusters and as seen on the first graph of PCA it looks like a cloud with not easy to locate clusters.

A screen shot of a computer

Description automatically generated

1. (5 points) Redo the PCA experiment with the MDCPS data, but color the points according to School Type. Try it again, but coloring the points according to Percent of Economically Disadvantaged Students. Discuss your conclusions from what you observe with these different color schemes. Note School Type is a categorical variable, but the other one is a real number.

We can see that school type is all over the place after the PCA Analysis not showing a very clear pattern. Nevertheless, we can see a pattern with economically disadvantaged students. We can practically imagine a set of clusters that differentiate the economic status of a student.

A screen shot of a graph

Description automatically generated

A screen shot of a graph

Description automatically generated