Assignment 4

Garcia, Jorge A.
Department of Physics
New Mexico State University
(Dated: April 30, 2020)

Course: C S 579

Instructor: Dr. Tuan Le

Question 1.

The resulting cluster labels for each movie are in the *movie_labels.csv* file.

Question 2.

Figure 1 shows the movie clusters when plotted in 2D after using PCA.

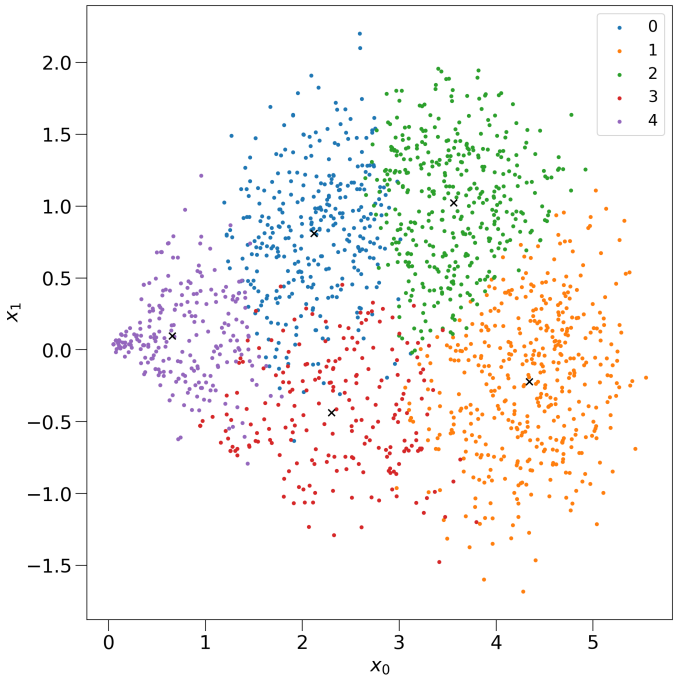

FIG. 1. Visualization of movie clusters in 2D from PCA.

Question 3.

The files movies_cx.txt with $x \in [0-4]$ contain the movie ID, Euclidean distance, movie name and genres for the top 20 movies in each cluster, ordered by increasing distance. Table I shows the genres present in these files, as well as their frequency in each cluster.

There seems to be no straight-forward grouping of genres. One problem when considering genres is the lack of distinction between "major" and "minor" genres. Drama and Action are very broad genres that denote an overall intention, whereas others like Sci-Fi and Film-Noir denote themes. If we disregard, to

some extent at least, the major genres of Drama, Comedy and Action, more insight of the groupings can be extracted.

Cluster 0 could be considered your typical action/shooter movies, given movies I recognize like “Love and a .45”, “Hard Target”, “The Substitute” and “Fair Game”. Cluster 1 are movies with darker themes, as seen by “A Clockwork Orange”, “Face/Off” and “Three Colors: Red”. Cluster 2 is more varied genre-wise, but there is a lot of crime-related action movies like “The Professional”, “A Time to Kill” and the two “Batman” movies. Almost the entirety of Cluster 3 are Drama movies, and are likely very sober and dialogue-based. Cluster 4 is the only one to present horror movies, which is distincitve with the rest, but there seems to be lacking a consistent theme as there are a few romance and children’s movies as well.

Overall, there doesn’t seem to be a great relationship between movie genre and the obtained clusters. This could be a data problem from either poor labelling of movies (for example, I would not label “A Clockwork Orange” as Sci-Fi) or lack of information such as the cast and director, or the themes explored in the movies. Or it could be a model problem, with more clusters being required or using a different clustering approach.

	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Drama	5	10	9	14	9
Comedy	7	7	5	5	5
Action	6	2	7		1
Thriller	4	3	2	2	1
Crime	1		4	2	
Romance	2	1	3	1	3
Adventure	1	1	4		
Children’s			3		2
Mystery		2	2		
Sci-Fi	1	2			
War	1	1	3		
Horror					3
Film-Noir		2			
Documentary			1	1	1
Fantasy			1		
Musical			1		
Animation			1		1

TABLE I. Genre frequency in the top 20 movies of each cluster.