CS24200: Project 2

Due date: Sunday Apr 11, 11:59pm

You can use python's sklean library for PCA, clustering and eigendecomposition. Submit a Jupyter notebook with both the code that you used for analysis and your answers to the questions below.

Download the "Small" Movie Lens data from: `https://github.com/srviest/movie-recommender/tree/master/dataset/movielens/small/`
This dataset (ml-latest-small) describes 5-star rating and free-text tagging activity from MovieLens, a movie recommendation service. It contains 100,004 ratings across 9125 movies from 671 users. These data were created by users between January 09, 1995 and October 16, 2016.

In this assignment, you will parse, transform, and cluster this dataset. You will evaluate and visualize the results.

1 Transforming Data (5 pts)

The `ratings.csv` file contains user ratings, one movie per line. See the README file for more information.

Transform the data into a user-movie ratings matrix. There should be 671 rows (one for each user) and 9125 columns (one for each movie). Each cell should contain the users rating for that movie. Note that not every user has rated every movie. Assign a value of 0 for any missing values.

2 Principle Component Analysis (15 pts)

Apply PCA to the reduce the dimensionality of the movies.

(a) Transpose the matrix from Q1 so that rows refer to movies and columns refer to users. Mean center the data. *Note that you will only use this transformed, mean-centered data for this question.*

(b) Apply PCA with number of components $k = 2$ to reduce the dimensionality of the movies.

(c) Plot the results and color each movies by its genre. Genres for each movie are listed in `ratings.csv`. Since each movie may have more than one genre, to simplify just use the first genre in the list. Discuss what patterns you see in the visualization.

(d) Determine the "intrinsic" dimensionality of the movies, by finding the number of principle components that are needed to explain 80% of the variance of the data. Discuss how this compares to $k = 2$ and how this may impact the quality of the visualization above.

3 Clustering (15 pts)

Apply k-means clustering to the data from Q1 (rows=users, columns=movies, no mean-centering) and cluster the users.

(a) For values of $k = [2, 4, 8, 16, 32]$, apply k-means and measure the `inertia` for each value of k. Plot the resulting inertia scores for each choice of k.

(b) From the above results, choose an appropriate value of k from the plot and support your choice.

(c) Cluster the data again with your chosen value of k. For each of the resulting clusters, find the top three movies that are highest rated (on average) by the users in the cluster. Report the movie titles and discuss whether the results seem reasonable (i.e., do the top-rated movies in each cluster seem to correspond to recognizable groups).

4 Singular Value Decomposition (15 pts)

Apply SVD to the user-movie matrix from Q1 (rrows=users, columns=movies, no mean-centering).

(a) Apply SVD with number of components $k = 32$. Plot the resulting `singular_values`.

(b) For each of the values of $k = [2, 4, 8, 16, 32]$ considered above, report the sum of the `explained_variance_ratio`. Discuss how the results compare to the `inertia` values above and whether it supports your choice of k.

(c) Apply SVD with $k = 2$ and transform the data.

(d) Plot the results (for $k = 2$) and color the users by the cluster memberships you found above. Discuss any patterns you can see and compare them to the previous analysis (from clustering and PCA).