CS24200: Project 2

Due date: Monday Nov 11, 11:59pm

You should use python, together with libraries numpy, maplotlib, sklearn and pandas to complete this assignment. Anything different from that should be asked for permission on Piazza. Submit a PDF with both the code that you used for analysis and your answers to the questions below. Your homework must be typed.

Download the "Small" Movie Lens data from: `https://grouplens.org/datasets/movielens/` This dataset (ml-latest-small) describes 5-star rating and free-text tagging activity from MovieLens, a movie recommendation service. It contains 100,836 ratings across 9724 movies from 610 users.

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 610 rows (one for each user) and 9724 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. What is the most common entry in the constructed matrix? What do we call matrices with such property?

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. Plot the new representations of the movies with a scatter plot.

(c) Determine how much of the variance is explained by each of these first two components. Can you observe the difference in variance between them in the plot?

(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 transposed data from Q1 (rows=movies, columns=users, no mean-centering) and cluster the movies.

(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) What kind of common characteristic would you expect the movies clustered together to have? Is this type of interpretation always possible?

Hint: Think about whether interpretations are explicitly modeled anywhere in the method.

4 Singular Value Decomposition (15 pts)

Apply SVD to the transposed user-movie matrix from Q1 (rows=movies, columns=users, 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`. Plot these values against the corresponding values of k. 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 movies by the cluster memberships you found above. Discuss any patterns you can see and compare them to the previous analysis (either from clustering or PCA).