Big Data Homework 4

Fabio Fraschetti 1834942

December 2023

[Fabio Fraschetti]

1 Exercise

I started this first exercise by importing the datasets from the links that you provided, and stored them into a dataframe. I used a particular method for transforming the dataframe in a way where I can take the movieIds in a specific order, that is CrossTab which set on the rows the movieids and on the columns the userids. Then from this I created a matrix and then converted it into a sparse matrix, that I think is the better structure to use for our purpose. I compute the Elbow method on this matrix and find that the best number of clusters to choose is 3 because here the curve start to decrease lower. Once I choose the number of clusters I run the KMeans Function and store the results. I aggregate to the previous dataframe the clustering results and the genres of the movie for based on the movieId. Finally I count the genres for each cluster and plot them. All the plots can be seen in the colab program. For the second part of the exercise I applied the function Truncated SVD and I found that starting from 20 number of components the variance is increasing slower and so I choose 20 as the total userId components I tried also other numbers like 40, 60 and 100 but in the end the results were almost the same. Now the sparse matrix pass from a shape of (65107, 6724) to a shape of (65107, 20) and then I applied the same process of before and found with the elbow method that as before the clusters to be produced are 3. Then I plot the counts of the genres for each plot.

Now we can analyze the plots and compare the two results. In general, producing a result in the matrix with SVD is faster. We can find the clusters with 1.7 seconds compered to the one without SVD that find the clusters in 7.8 seconds. To answer the question: Does the clustering based on the above matrix U reflect a clustering of the movies into different genres? Yes, based on the plots that I made I have a quite good division of the genres into the 3 clusters that identifies the groups of user that likes some particular genre for example in the plots where I applied only KMeans the first cluster represent the group of user that likes Romance and Horror while in the second cluster the most liked genres are Sci-Fi and Crime and in the last cluster the most liked are Sci-Fi and Romance. I tried also with 4 clusters but 2 of them were almost equals for this reason I decided to take only 3 clusters. The clusters generated with SVD are more or less the same as the one with only KMeans that is a good result considering that the used matrix was with less columns. I noticed also that a lot of films haven't a ratings, maybe with more reviews we could have given more importance to the other genres that aren't predominant in the clusters, such as War, Western, Documentary ecc...

In the end I can say that I found a quite good clustering of the genres of the movies based on the ratings of the user.

2 Exercise

2.1

First of all I start applying some definitions given by the problem. So we can write this:

$$E[f(\boldsymbol{x})^T f(\boldsymbol{y})] = E[(S\boldsymbol{x})^T S\boldsymbol{y}] = E\left[\left(\frac{1}{\sqrt{k}}U\boldsymbol{x}\right)^T \left(\frac{1}{\sqrt{k}}U\boldsymbol{y}\right)\right] = E\left[\frac{1}{k}\left(U\boldsymbol{x}\right)^T \left(U\boldsymbol{y}\right)\right]$$

Now I can rewrite the $i-th$ component of the matrix Uy and $(Ux)^T$ as

$$(U\boldsymbol{y})_i = \sum_{j=1}^{d} U_{ij}\boldsymbol{y}_j$$

$$(U\boldsymbol{x})_i^T = \left(\sum_{h=1}^{d} U_{ih}\boldsymbol{x}_h\right)^T = \sum_{h=1}^{d} U_{ih}\boldsymbol{x}_h$$

So replacing this result into the previous equation, it become

$$\frac{1}{k}E\left[\sum_{i=1}^{k}\sum_{j,h=1}^{d} U_{ih}U_{ij}\boldsymbol{x}_h\boldsymbol{y}_j\right] = \frac{1}{k}\sum_{i=1}^{k}\sum_{j,h=1}^{d} E\left[U_{ih}U_{ij}\right]\boldsymbol{x}_h\boldsymbol{y}_j$$

I can make the expected value of $U_{ih}U_{ij}$ that is 1 when $h = j$ because a squared normal distribution has expected value 1 while when $h \neq j$ the expected value is 0 because the expected value of the multiplication of 2 independent normal distribution is 0. So we can write:

$$\frac{1}{k}\sum_{i=1}^{k}\sum_{j,h=1}^{d} \delta_{hj}\boldsymbol{x}_h\boldsymbol{y}_j$$

The last sum is the scalar product of $\boldsymbol{xy}$ in the end I can write

$$\frac{1}{k}\sum_{i=1}^{k}\boldsymbol{x}^T\boldsymbol{y} = \frac{k}{k}\boldsymbol{x}^T\boldsymbol{y} = \boldsymbol{x}^T\boldsymbol{y}$$

2.2

I choose as unbiased estimator $f(\boldsymbol{x})^T f(\boldsymbol{y}) = f(\boldsymbol{x}) \cdot f(\boldsymbol{y})$ because $E[f(\boldsymbol{x}) \cdot f(\boldsymbol{y})] = \boldsymbol{x} \cdot \boldsymbol{y}$ demonstrated in the previous exercise and from the first assumption ($||\boldsymbol{x}||_2 = ||\boldsymbol{y}||_2 = 1$) we can derive that $\frac{\boldsymbol{x}\cdot\boldsymbol{y}}{||\boldsymbol{x}||_2||\boldsymbol{y}||_2} = \boldsymbol{x} \cdot \boldsymbol{y} = \cos\theta$ by definition of scalar product.