Sapienza University of Rome

Big Data Computing Second Homework

Andrea Fioraldi 1692419

December 11, 2019

Exercise 1

The dataset can be easily converted to a list of tuples of this form:
`[ (x, [friends of x ...]) ...]`
With flatMap, we execute a function that generates a list in this format:
`[ ((x, y), [a common friend]) ...]`
`x` and `y` are nodes and of course the key `(x, y)` is not unique.
Reducing by this key we sum all the common friends related to each `x`,`y` pair.
The final result is in this form: `[ ((x, y), [common friends ...]) ...]`
With `(x, y)` as unique key.

Exercise 2

The topics are randomly chosen in a smart way to have a balanced number for each of the 6 macro-topics.
I've tested all the 3 approcahes proposed but PCA and $i = 20$ finishes the avaiable RAM and so I used $i = 10$.
The results, in term of seconds nedded to produce the clusters, are:

	KMeans	Dim. Reduction	KMeans after D.M.
SVD $i = 6$	228.273s	0.575402s	0.413s
SVD $i = 20$	1019.745s	3.378942s	4.001s
PCA $i = 6$	211.559s	10.296523s	0.494s
PCA $i = 10$	405.742s	31.326672s	1.102s

As you can see, SVD seems faster than PCA. Looking in the attached `ex2_output` file you can also see that KMeans after PCA improve the Homogeneity respect than KMeans after

SVD (compare the runs with $i = 6$, 0.586 vs. 0.450). In general, KMens after SVD/PCA improve the Homogeneity respect plain KMeans.

Exercise 3

3.1

$$A^{-1} = (U\Sigma V^T)^{-1} =^{(1)} (V^T)^{-1}\Sigma^{-1}U^{-1} =^{(2)} V\Sigma^{-1}U^T = \sum_{i=1}^{r} \frac{1}{\sigma_i} v_i u_i^T$$

(1) Due to [1].
(2) We know that U and V are orthonormal and so the transpose is equal to the inverse.

3.2

$$\begin{aligned} BAx &= (V\Sigma^{-1}U^T) * (U\Sigma V^T) * x = \Big(\sum_{i=1}^{r} \frac{1}{\sigma_i} v_i u_i^T\Big) * \Big(\sum_{i=1}^{r} \sigma_i u_i v_i^T\Big) * x \\ &= \Big(\sum_{i=1}^{r} v_i u_i^T u_i v_i^T + \sum_{i=1}^{r}\sum_{j \neq i}^{r} \frac{\sigma_i}{\sigma_j} v_i u_i^T u_j v_j^T\Big) * x \end{aligned}$$

We know that U and V are orthonormal, so a vector k of U or V has the following property:

$$(1)\ k_i^T k_j = \begin{cases} 0 & i \neq j \\ 1 & i = j \end{cases}$$

So

$$BAx = \Big(\sum_{i=1}^{r} v_i * 1 * v_i^T + \sum_{i=1}^{r}\sum_{j \neq i}^{r} \frac{\sigma_i}{\sigma_j} v_i * 0 * v_i^T\Big) * x = \Big(\sum_{i=1}^{r} v_i v_i^T\Big) * x$$

Knowing that $x = \sum_{i=1}^{r} \alpha_i v_i$ and that we want $BAx = x$:

$$\begin{gathered} \Big(\sum_{i=1}^{r} v_i v_i^T\Big) * \Big(\sum_{i=1}^{r} \alpha_i v_i\Big) = \sum_{i=1}^{r} \alpha_i v_i \\ \to \sum_{i=1}^{r} v_i v_i^T \alpha_i v_i + \sum_{i=1}^{r}\sum_{j \neq i}^{r} v_i v_i^T \alpha_j v_j = \sum_{i=1}^{r} \alpha_i v_i \\ \to^{(1)} \sum_{i=1}^{r} v_i * 1 * \alpha_i + \sum_{i=1}^{r}\sum_{j \neq i}^{r} v_i * 0 * \alpha_j = \sum_{i=1}^{r} \alpha_i v_i \\ \to \sum_{i=1}^{r} \alpha_i v_i = \sum_{i=1}^{r} \alpha_i v_i \end{gathered}$$

We proved that $BAx = x$.

References

[1] "linear algebra - Product of inverse matrices $(AB)^{-1}$ - Mathematics Stack Exchange." `https://math.stackexchange.com/questions/688339/product-of-inverse-matrices-ab-1`. Accessed: 2019-12-11.