REI602M **Quiz 19.2.2021**

Hand in on Gradescope before 22:00 on Feb. 20 (Saturday). Each question will be given 1, 0.5 or 0 points as follows. If the question is more or less correct it gets 1 point. If it is partly correct it gets 0.5, and if it is missing or completely wrong it gets 0 points.

1) [Exam 2019]

The matrix X below shows the number of products purchased by six customers over a short period.

	John	Alice	Mary	Greg	Peter	Jennifer
Vegetables	0	1	0	1	2	2
Fruits	2	3	1	1	2	2
Sweets	1	1	1	0	1	1
Bread	0	2	3	4	1	1
Coffee	0	0	0	0	1	0

In order to analyze purchasing behavior, non-negative matrix factorization $X \approx WH$ was performed with $k = 3$ components, resulting in

$$W = \begin{bmatrix} 0 & 0.04 & 2.74 \\ 1.93 & 0.15 & 0.47 \\ 0.97 & 0 & 0 \\ 0 & 2.66 & 1.18 \\ 0 & 0 & 0.59 \end{bmatrix}, \qquad H = \begin{bmatrix} 1.04 & 1.34 & 0.55 & 0.26 & 0.89 & 0.9 \\ 0 & 0.6 & 1.12 & 1.36 & 0.03 & 0.07 \\ 0 & 0.35 & 0 & 0.34 & 0.77 & 0.69 \end{bmatrix}$$

a) Provide a qualitative interpretation of the columns of W.

If we had K = 6 we would have each column for each person, Each line in each column in W is a representation of each food purchase.

b) Provide a qualitative interpretation of the first two columns of H.

Column 1 is a reference to John and column 2 is a reference to Alice.

2)

Building a recommendation engine using k-means. A set of N users of a music-streaming app listens to songs from a library of n songs over some period (say, a month). We describe user i's listening habits by her playlist vector, which is the n-vector p_i defined as

$$(p_i)_j = \begin{cases} 1 & \text{user } i \text{ has played song } j \\ 0 & \text{user } i \text{ has not played song } j, \end{cases}$$

for $j = 1, \ldots, n$. (Note that p_i is an n-vector, while $(p_i)_j$ is a number.) You can assume that if a user listens to a song, she likes it.

Your job (say, during a summer internship) is to design an algorithm that recommends to each user 10 songs that she has not listened to, but might like. (You can assume that for each user, there are at least 10 songs that she has not listened to.)

To do this, you start by running k-means on the set of playlist vectors $p_1, \ldots, p_N$. (It's not relevant here, but a reasonable choice of k might be 100 or so.) This gives the centroids $z_1, \ldots, z_k$, which are n-vectors.

Now what do you do? You can explain in words; you do not need to give a formula to explain how you make the recommendations for each user.

First thing to note is that P(i)j = User likes song.
(1)We start by placing each play'd song on a dataset. If we have k = 100, therefore 100 centroids.
(2)We start by placing playlists in each cluster and (3)calculate the mean of all datapoints in each cluster and use the mean as a new centroid. Repeat from step 2 until the clusters do not change anymore.
Now we can recommend songs to user according to his preference of songs.

3)

Pre-assigned vectors. Suppose that some of the vectors $x_1, \ldots, x_N$ are assigned to specific groups. For example, we might insist that x_{27} be assigned to group 5. Suggest a simple modification of the k-means algorithm that respects this requirement. Describe a practical example where this might arise, when each vector represents n features of a medical patient.

We could use weighted k-means by using the pre-assigned points as a centroid or with more weight so the centroid of the group that it was defined in does not move away from the point in the iteration process.

Example for this is medication that works for most people but we need to place it in a "not useable for this patent cluster" because the medication contains a chemical that the patent this patent is allergic to.

4) Consider ratings data for a group of products in the form of like/dislike. Is non-negative matrix factorization (NMF) a suitable method for this type of data? Explain your answer briefly.

NMF should not be considered as a method for this kind of data, the data with two groups of

(like and dislike) are really binary data.