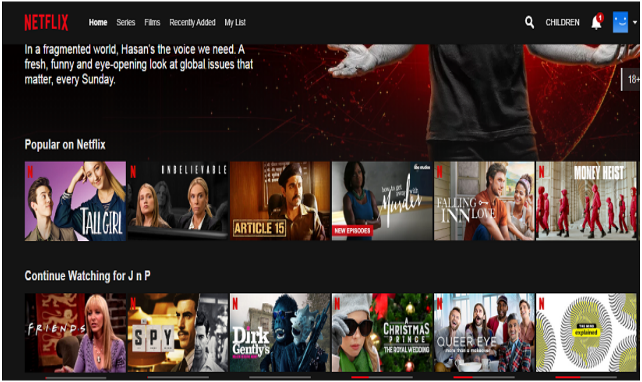
Non-negative Matrix Factorization and its Applications

Jaiprakash R

Research Scholar, Department of Mathematics
IIT Kharagpur

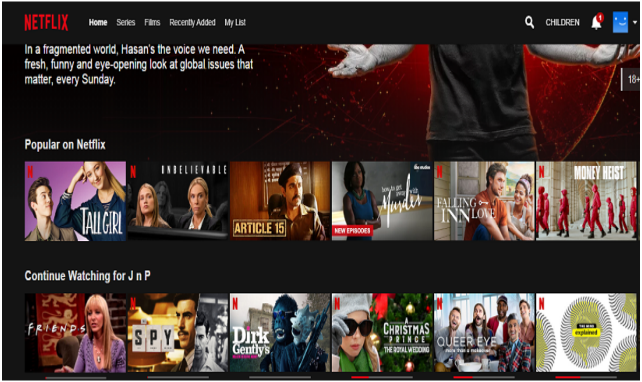
19 September 2019

Why?



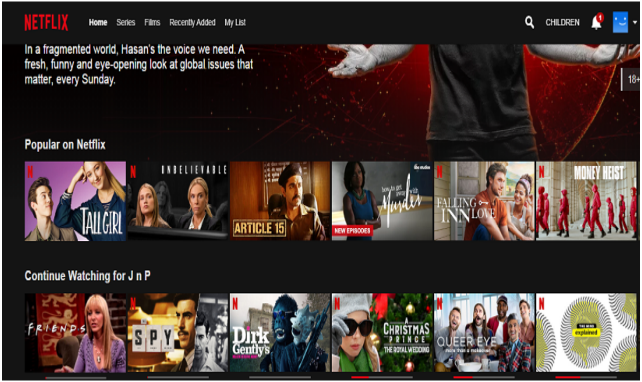
- Why does Netflix recommend these movies to me?

Why?



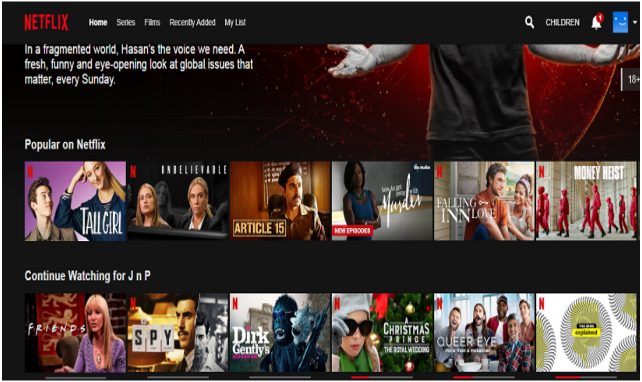
- Why does Netflix recommend these movies to me?
- Strangely enough, these happen to match my taste.

Why?



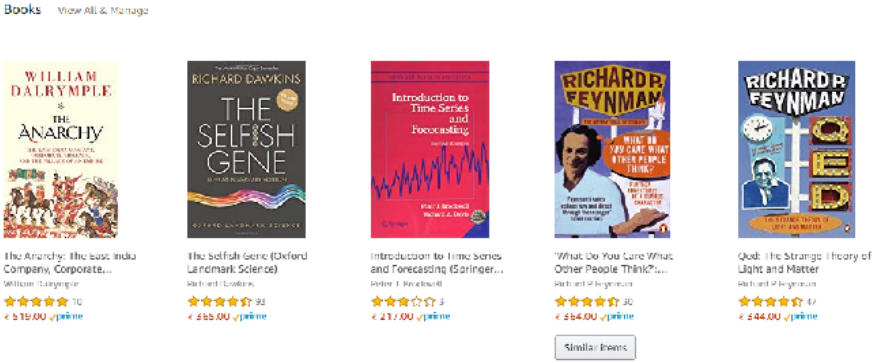
- Why does Netflix recommend these movies to me?
- Strangely enough, these happen to match my taste.
- How?

Why?



- Why does Netflix recommend these movies to me?
- Strangely enough, these happen to match my taste.
- How?
- Linear Algebra. That's how!

Why?


Books
View All & Manage
The Anarchy: The East India Company, Corporate...
The Selfish Gene (Oxford Landmark Science)
Introduction to Time Series and Forecasting (Springer...
"What Do You Care What Other People Think?":...
Qed: The Strange Theory of Light and Matter
Similar Items

Toy Example

Let's construct a toy example which would help us understand *NMF* better.
Suppose we have the following data of ratings by certain users for certain TV shows. The ratings are on a scale of 1-5. A '0' entry denotes that the rating is not available.

X	**F.R.I.E.N.D.S**	**TBBT**	**Black Mirror**	**Modern Family**	**Sacred Games**	**GoT**	**Quantico**
Jai	4.8	4.7	0	4.6	0	0	2.1
PB	4	4.1	4.8	0	4.9	4.7	1.5
Arnab	0	3.5	4	0	0	4.6	2.5
Abhi	4.9	4	0	0	0	4.4	0
Diya	2.1	4.4	4.7	0	0	4.9	0

Question: Based on this data, for a given TV show, can we predict the rating of a user who hasn't rated it yet? Thereby, create a scheme of recommending shows to users which they haven't watched yet.

Toy Example

Let X be a matrix whose rows correspond to the TV shows and columns correspond to the users in the given data.

$$X = \begin{bmatrix} 4.8 & 4 & 0 & 4.9 & 2.1 \\ 4.7 & 4.1 & 3.5 & 4 & 4.4 \\ 0 & 4.8 & 4 & 0 & 4.7 \\ 4.6 & 0 & 0 & 0 & 0 \\ 0 & 4.9 & 0 & 0 & 0 \\ 0 & 4.7 & 4.6 & 4.4 & 4.9 \\ 2.1 & 1.5 & 2.5 & 0 & 0 \end{bmatrix}$$

What is NMF?

- Let $A \in \mathbb{R}^{m \times n}$. Then it 'might' admit various decompositions such as *LU*, *Cholesky*, *QR*, *Spectral*, *SVD* etc...

What is NMF?

- Let $A \in \mathbb{R}^{m \times n}$. Then it 'might' admit various decompositions such as *LU*, *Cholesky*, *QR*, *Spectral*, *SVD* etc...
- *NMF* is one such factorization with some caveats.

What is NMF?

- Let $A \in \mathbb{R}^{m \times n}$. Then it 'might' admit various decompositions such as *LU*, *Cholesky*, *QR*, *Spectral*, *SVD* etc...
- *NMF* is one such factorization with some caveats.
- As the name suggests, both A and its factors must be non-negative i.e. each of their entries must be non-negative.

Formal Definition

Definition

Let $X \in \mathbb{R}^{m \times n}$ and $X \geq 0$ (i.e. $\forall i, j; x_{ij} \geq 0$). It is said to admit *NMF* if $\exists\ 0 \leq W \in \mathbb{R}^{m \times r}, 0 \leq H \in \mathbb{R}^{r \times n},$ for some $r < \min(m, n),$ such that

$$X = WH$$

.

Formal Definition

Definition

Let $X \in \mathbb{R}^{m \times n}$ and $X \geq 0$ (i.e. $\forall i, j; x_{ij} \geq 0$). It is said to admit *NMF* if $\exists\ 0 \leq W \in \mathbb{R}^{m \times r}, 0 \leq H \in \mathbb{R}^{r \times n},$ for some $r < \min(m, n),$ such that

$$X = WH$$

.

- The smallest r for which it is possible is called as the non-negative rank, say r^+.
- r^+ is the smallest number such that the matrix can decomposed into a sum of non-negative rank-1 matrices.
- For our purpose, we interpret X to be a data matrix whose rows represent features, and columns represent observations.

Low Rank Approximation

- Given a choice of r, finding the exact *NMF* for a non-negative matrix has been shown to be infeasible.
- Therefore, we try to find the nearest factorization, in some sense, by numerically solving the following optimization problem.

$$(W^*, H^*) = \arg \min_{W \geq 0, H \geq 0} \|X - WH\|$$

- The minimization could be w.r.t. any norm that is appropriate for the particular application.
- Note that *NMF* is posed as a *constrained low rank approximation* problem. This indicates that it could be used for dimensionality reduction in data with non-negative values.

Closer look at NMF

- Suppose $V = [v_1, v_2, \cdots, v_n] = WH$ where $W = [w_1, w_2, \cdots, w_r]$ and $H = [h_1, h_2, \cdots, h_n]$.
- Then, it's clear that $v_i = Wh_i$ for $i = 1, 2, \cdots, n$.

Closer look at NMF

- Suppose $V = [v_1, v_2, \cdots, v_n] = WH$ where $W = [w_1, w_2, \cdots, w_r]$ and $H = [h_1, h_2, \cdots, h_n]$.
- Then, it's clear that $v_i = Wh_i$ for $i = 1, 2, \cdots, n$.
- Thus, in *NMF* the columns of the transformed matrix (the approximation) is the non-negative linear combinations of the columns of W.
- Therefore, the columns of W represent a set of basis functions for the transformed data, and H contains some sort of encoding.

Geometric View

- Since each of our observations x_i, the i^{th}*column* of X, is non-negative, it follows that our data lies in the non-negative orthant of $\mathbb{R}^m$ which is a cone.
- Through *NMF* we essentially say that these data points actually lies in lower dimensional cone generated by $\{w_1, \cdots, w_r\}$.
- However, the factors obtained in *NMF* is not unique, in general. Thus, every time we run the algorithm, we might end up with a different lower dimensional cone. Nevertheless, a unique solution for *NMF* does exist under certain regularity conditions.

NMF for the Toy Example

- Although, we have 7 TV shows (features) in our data, they seem to belong to two categories viz. thrillers and sitcoms. Therefore, we can choose r to be 2.
- On running an *NMF* algorithm, we obtained the following factors.

$$W = \begin{bmatrix} 0.971 & 1.929 \\ 1.699 & 1.626 \\ 1.872 & 0. \\ 0. & 1.317 \\ 0.77 & 0. \\ 2.219 & 0.232 \\ 0.452 & 0.543 \end{bmatrix} \qquad H = \begin{bmatrix} 0. & 2.53 & 1.898 & 1.09 & 2.217 \\ 2.853 & 0.312 & 0. & 1.314 & 0.045 \end{bmatrix}$$

Interpretation of W

- The rows represent the shows, and the columns represent the new features (categories, in this case). What do the entries signify?

Interpretation of W

- The rows represent the shows, and the columns represent the new features (categories, in this case). What do the entries signify?
- Looking from the perspective of rows, they give the degree to which they belong to each of the category.

Interpretation of W

- The rows represent the shows, and the columns represent the new features (categories, in this case). What do the entries signify?
- Looking from the perspective of rows, they give the degree to which they belong to each of the category.
- For instance, TBBT (1^{st} row) leans more towards the second category. On the other hand, Black Mirror(3^{rd} row) leans more towards the first. They make sense, right?

Interpretation of W

- The rows represent the shows, and the columns represent the new features (categories, in this case). What do the entries signify?
- Looking from the perspective of rows, they give the degree to which they belong to each of the category.
- For instance, TBBT (1^{st} row) leans more towards the second category. On the other hand, Black Mirror(3^{rd} row) leans more towards the first. They make sense, right?
- On the other hand, looking from the perspective of columns (the new basis vectors), each entry corresponds to the contribution of the corresponding show to the respective column.

Interpretation of W

- The rows represent the shows, and the columns represent the new features (categories, in this case). What do the entries signify?
- Looking from the perspective of rows, they give the degree to which they belong to each of the category.
- For instance, TBBT (1^{st} row) leans more towards the second category. On the other hand, Black Mirror(3^{rd} row) leans more towards the first. They make sense, right?
- On the other hand, looking from the perspective of columns (the new basis vectors), each entry corresponds to the contribution of the corresponding show to the respective column.
- Black Mirror and Sacred Games contribute nothing, based on this data, to the sitcom category. Good job, *NMF*!

Interpretation of H

- The rows correspond to the categories, and the columns correspond to the users. So?

Interpretation of H

- The rows correspond to the categories, and the columns correspond to the users. So?
- Each entry represents indicates the leaning of a user to a particular category.

Interpretation of H

- The rows correspond to the categories, and the columns correspond to the users. So?
- Each entry represents indicates the leaning of a user to a particular category.
- Through this natural interpretation of the results, Linear Algebra, and hence math itself, has established its supremacy again.

Interpretation of H

- The rows correspond to the categories, and the columns correspond to the users. So?
- Each entry represents indicates the leaning of a user to a particular category.
- Through this natural interpretation of the results, Linear Algebra, and hence math itself, has established its supremacy again.
- The interpretation made in this TV show-users setting can be naturally extended to any other setting where *NMF* can be employed.

Application 1 : Collaborative Filtering

- Collaborative filtering is a recommender system which recommends products to users based on 'similarity' of products/users.

Application 1 : Collaborative Filtering

- Collaborative filtering is a recommender system which recommends products to users based on 'similarity' of products/users.
- The most commonly used similarity measure is the cosine distance.

$$d_{cos}(x, y) = \frac{< x, y >}{\|x\|_2 \|y\|_2}$$

Application 1 : Collaborative Filtering

- Collaborative filtering is a recommender system which recommends products to users based on 'similarity' of products/users.
- The most commonly used similarity measure is the cosine distance.

$$d_{cos}(x, y) = \frac{< x, y >}{\|x\|_2 \|y\|_2}$$

- In the dense matrix W each of the features are essentially points in $\mathbb{R}^r$.

Application 1 : Collaborative Filtering

- Collaborative filtering is a recommender system which recommends products to users based on 'similarity' of products/users.
- The most commonly used similarity measure is the cosine distance.

$$d_{cos}(x, y) = \frac{< x, y >}{\|x\|_2 \|y\|_2}$$

- In the dense matrix W each of the features are essentially points in $\mathbb{R}^r$.
- If the rating of user j for the item i is not available, it's estimated as the weighted average of the ratings of some number of nearest neighbors by the user i.

Application 1 : Collaborative Filtering

The similarity matrix for the shows in our example is as below:

$$S = \begin{bmatrix} 1. & 0.942 & 0.45 & 0.893 & 0.45 & 0.54 & 0.974 \\ 0.942 & 1. & 0.722 & 0.691 & 0.722 & 0.79 & 0.994 \\ 0.45 & 0.722 & 1. & 0. & 1. & 0.995 & 0.64 \\ 0.893 & 0.691 & 0. & 1. & 0. & 0.104 & 0.769 \\ 0.45 & 0.722 & 1. & 0. & 1. & 0.995 & 0.64 \\ 0.54 & 0.79 & 0.995 & 0.104 & 0.995 & 1. & 0.716 \\ 0.974 & 0.994 & 0.64 & 0.769 & 0.64 & 0.716 & 1. \end{bmatrix}$$

Application 1 : Collaborative Filtering

- Suppose that we want our recommender system to decide whether or not to recommend Black Mirror to me.

Application 1 : Collaborative Filtering

- Suppose that we want our recommender system to decide whether or not to recommend Black Mirror to me.
- The first three nearest neighbors to Black Mirror are Sacred Games, GoT and TBBT. Assign ranks(k_m) 1,2 and 3 respectively.

Application 1 : Collaborative Filtering

- Suppose that we want our recommender system to decide whether or not to recommend Black Mirror to me.
- The first three nearest neighbors to Black Mirror are Sacred Games, GoT and TBBT. Assign ranks(k_m) 1,2 and 3 respectively.
- Let the weights be $0.5(1 - \frac{k_m}{\sum_{m=1}^{3}(k_m)})$. Thus, the weights are $\frac{5}{12}, \frac{4}{12},$ and $\frac{3}{12}$ respectively.

Application 1 : Collaborative Filtering

- Suppose that we want our recommender system to decide whether or not to recommend Black Mirror to me.
- The first three nearest neighbors to Black Mirror are Sacred Games, GoT and TBBT. Assign ranks(k_m) 1,2 and 3 respectively.
- Let the weights be $0.5(1 - \frac{k_m}{\sum_{m=1}^{3}(k_m)})$. Thus, the weights are $\frac{5}{12}, \frac{4}{12}, \text{ and } \frac{3}{12}$ respectively.
- The estimated rating is then 1.175. Thus, Black Mirror wouldn't be recommended to me. Right on point!

Is this the best approach?

- Definitely not! There are some problems with this naive version.

Is this the best approach?

- Definitely not! There are some problems with this naive version.
- For instance, we didn't normalize each row of X in which case the average rating would be 0. Had we done that, how do we represent an unavailable rating?

Is this the best approach?

- Definitely not! There are some problems with this naive version.
- For instance, we didn't normalize each row of X in which case the average rating would be 0. Had we done that, how do we represent an unavailable rating?
- If we leave that cell empty, then *NMF* algorithm would crash!

Is this the best approach?

- Definitely not! There are some problems with this naive version.
- For instance, we didn't normalize each row of X in which case the average rating would be 0. Had we done that, how do we represent an unavailable rating?
- If we leave that cell empty, then *NMF* algorithm would crash!
- Non-uniqueness of this approximation might also cause some troubles. However, in most cases, the local optima obtained has proved to be effective in applications.
- Nevertheless, *NMF* was formerly used by big names like Netflix, Amazon etc... The present technology involves a low rank matrix completion problem.

Application 2 : Document Classification

- Document classification is another crucial problem in *artificial intelligence*.
- Question: Given a large unstructured collection of documents, can we identify the hidden patterns in them and there-by group them?

Application 2 : Document Classification

- Document classification is another crucial problem in *artificial intelligence*.
- Question: Given a large unstructured collection of documents, can we identify the hidden patterns in them and there-by group them?
- Essentially, this is a clustering/unsupervised classification problem in *machine learning* terminology.

Application 2 : Document Classification

- Document classification is another crucial problem in *artificial intelligence*.
- Question: Given a large unstructured collection of documents, can we identify the hidden patterns in them and there-by group them?
- Essentially, this is a clustering/unsupervised classification problem in *machine learning* terminology.
- Applications involve spam filtering, sentiment analysis, tagging content/genre classification etc...

Application 2 : Document Classification

- Suppose we have a dictionary of words for our collection of documents, and the frequencies of each words in each of the document.

Application 2 : Document Classification

- Suppose we have a dictionary of words for our collection of documents, and the frequencies of each words in each of the document.
- Our matrix X, in this case, is a word by document matrix i.e rows represent the words and the columns represent documents. Therefore, x_{ij} is the frequency of the i^{th} word of the dictionary in j^{th} document.

Application 2 : Document Classification

- Suppose we have a dictionary of words for our collection of documents, and the frequencies of each words in each of the document.
- Our matrix X, in this case, is a word by document matrix i.e rows represent the words and the columns represent documents. Therefore, x_{ij} is the frequency of the i^{th} word of the dictionary in j^{th} document.
- W will be a word by category, and H will be a category by document matrix.

Application 2 : Document Classification

- Suppose we have a dictionary of words for our collection of documents, and the frequencies of each words in each of the document.
- Our matrix X, in this case, is a word by document matrix i.e rows represent the words and the columns represent documents. Therefore, x_{ij} is the frequency of the i^{th} word of the dictionary in j^{th} document.
- W will be a word by category, and H will be a category by document matrix.
- We know that the entries of H indicate the extent to which a document belongs to a category.

Application 2 : Document Classification

- Suppose we have a dictionary of words for our collection of documents, and the frequencies of each words in each of the document.
- Our matrix X, in this case, is a word by document matrix i.e rows represent the words and the columns represent documents. Therefore, x_{ij} is the frequency of the i^{th} word of the dictionary in j^{th} document.
- W will be a word by category, and H will be a category by document matrix.
- We know that the entries of H indicate the extent to which a document belongs to a category.
- Therefore, a document j will be classified as the category-k if h_kj is the largest entry in column h_j.

Application 2 : Document Classification

- As an example we shall look at the topic extraction example of a dataset in *sklearn.datatsets* of documents from 20 different news groups. The code can be found at the website of *scikit*$_l$*earn*.
- This data has 2000 documents with 1000 features. We try to group them into 10 clusters based on the topics they deal with.

Application 3 : Image Processing

- Image processing typically entails dimensionality reduction (decomposition into components), object classification, facial recognition etc...
- The data matrix X is typically a pixel by image matrix. Each column has the pixes values of the corresponding image.
- Question: Can we find a lower dimensional representation for this image data?

Application 3 : Image Processing

- Image processing typically entails dimensionality reduction (decomposition into components), object classification, facial recognition etc...
- The data matrix X is typically a pixel by image matrix. Each column has the pixes values of the corresponding image.
- Question: Can we find a lower dimensional representation for this image data?
- Yes! We know that *NMF* is capable of doing this!

Application 4 : Bioinformatics

- The field of cancer prognosis has also benefitted from *NMF*. This is just one mention of the various applications of *NMF* in bioinformatics.

Application 4 : Bioinformatics

- The field of cancer prognosis has also benefitted from NMF. This is just one mention of the various applications of NMF in bioinformatics.
- X would be an expression level by gene matrix. Then H would be a metagene (the representative genes of various classes or sub-classes of cancer) by gene matrix.

Application 4 : Bioinformatics

- The field of cancer prognosis has also benefitted from *NMF*. This is just one mention of the various applications of *NMF* in bioinformatics.
- X would be an expression level by gene matrix. Then H would be a metagene (the representative genes of various classes or sub-classes of cancer) by gene matrix.
- In a way similar to document classification, we can identify the correspondence of a particular gene with a particular class/subclass of cancer.