Non-Negative Matrix Factorization

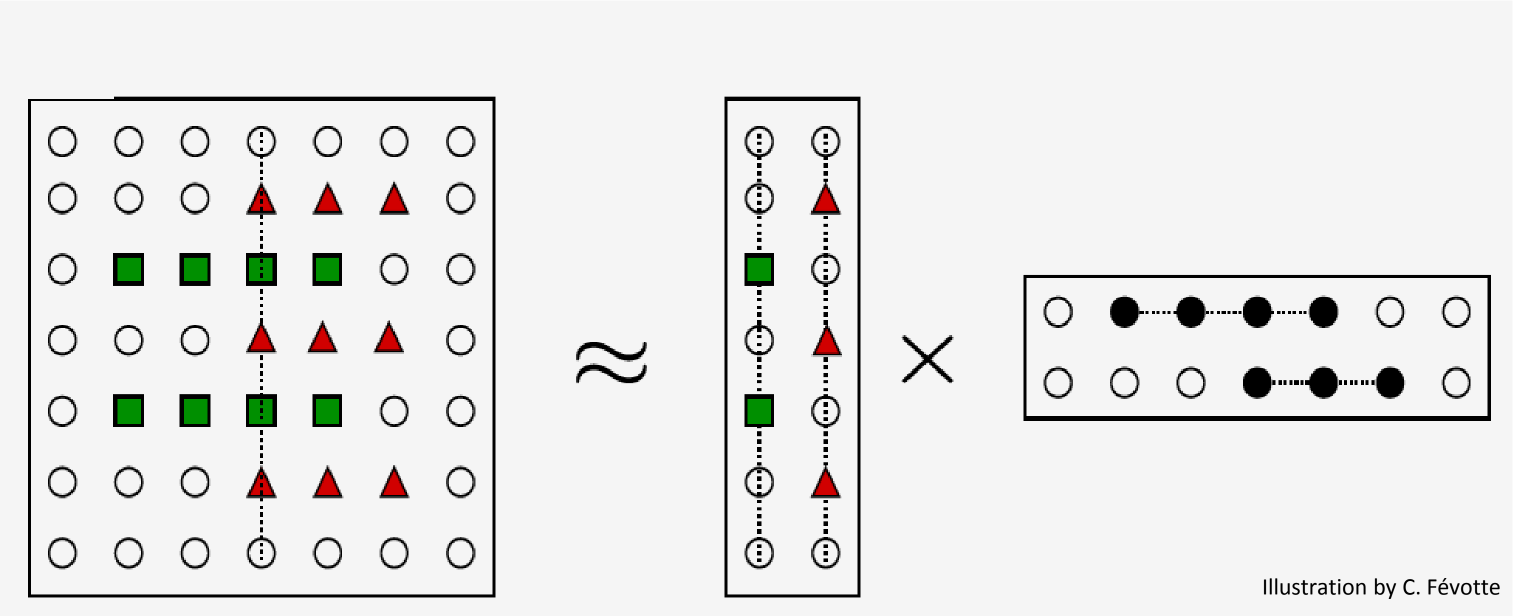
Hongchang Gao

Outline

- Matrix Factorization
- NMF
- Optimization Algorithm
- Variants of NMF
- The relationship with K-Means
- Application

Matrix Factorization

- Matrix Factorization is to find out two (or more) matrices such that we can use the product of these matrices to approximate the original matrix.



Data matrix Basis Coefficient

Matrix Factorization

- Examples:
  - PCA
    $$\min_{U,V} ||X - UV||_F^2$$
    $$s.t. U^T U = I$$
    - X: data points in high dimensional space
    - U: the basis of the low dimensional space
    - V: new representation in the low dimensional space
  - Recommendation System
    $$\min_{U,V} ||X - UV||_F^2$$
    - X: user ratings
    - U: user latent factor
    - V: item latent factor

Outline

- Matrix Factorization
- NMF
- Optimization Algorithm
- Variants of NMF
- The relationship with K-Means
- Application

Non-Negative Matrix Factorization

- “Learning the parts of objects by non-negative matrix factorization”

—Nature 1999

- “Algorithms for non-negative matrix factorization”

—NIPS 2001

- Definition

$$\min \| X - FG^T \|_F^2$$
$$s.t.\ \ F \geq 0, G \geq 0$$

Interpretation with NMF

$$X \approx FG^T$$

- Columns of F are the underlying basis vectors

$$F = [f_1, f_2, \ldots., f_k]$$

- Rows of G give the weights associated with each basis vector.

$$G = \begin{bmatrix} g^1 \\ g^2 \\ \vdots \\ g^k \end{bmatrix} \qquad x_i = f_1 g_1^i + f_2 g_2^i + \ldots. + f_k g_k^i$$

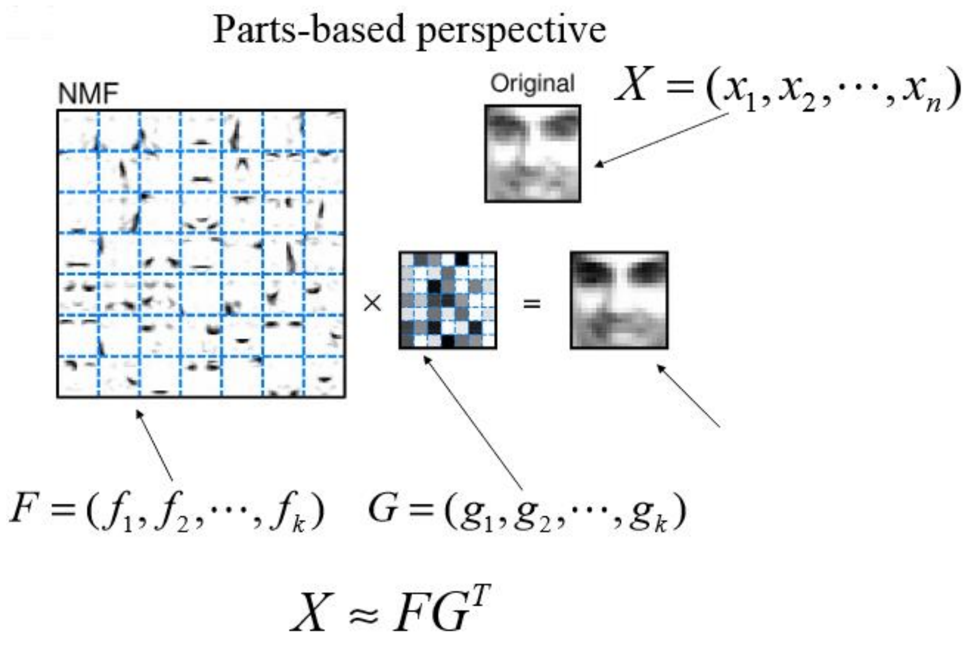
only additive combinations!!!

Interpretation with NMF

- Parts-Based Representation

  –by Nature 1999

  - The basis images contain several versions of mouths, noses and other facial parts, which are in different locations or forms
  - A whole face is generated by combining these different parts



Outline

- Matrix Factorization
- NMF
- Optimization Algorithm
- Variants of NMF
- The relationship with K-Means
- Application

How to solve NMF?

- Objective function:

$$\min \| X - FG^T \|_F^2$$
$$s.t.\ \ F \geq 0, G \geq 0$$

  - Non-convex for F and G simultaneously
  - Convex for F or G separately

Convex Function

f is convex if $\mathbf{dom}\, f$ is a convex set and Jensen's inequality holds:

$$f(\theta x + (1-\theta)y) \leq \theta f(x) + (1-\theta)f(y) \quad \forall x, y \in \mathbf{dom}\, f,\ \theta \in [0,1]$$

first-order condition

for (continuously) differentiable f, Jensen's inequality can be replaced with

$$f(y) \geq f(x) + \nabla f(x)^T (y - x) \quad \forall x, y \in \mathbf{dom}\, f$$

second-order condition

for twice differentiable f, Jensen's inequality can be replaced with

$$\nabla^2 f(x) \succeq 0 \quad \forall x \in \mathbf{dom}\, f$$

Multiplicative Update Method

- The most common used method
- Proposed by Lee and Seung (2001)
- The update rule:
  - Fix F, solve for G
  - Fix G, solve for F

$$F_{ik} \leftarrow F_{ik} \frac{(XG)_{ik}}{(FG^T G)_{ik}} \qquad G_{jk} \leftarrow G_{jk} \frac{(X^T F)_{jk}}{(GF^T F)_{jk}}$$

Multiplicative Update Method

- Arise from gradient descent method

$$F_{ik} \leftarrow F_{ik} + \varepsilon_{ik}[(XG)_{ik} - (FG^T G)_{ik}]$$

  - Where ε_{ik} is a small positive number.

- Set it as

$$\varepsilon_{ik} = \frac{F_{ik}}{(FG^T G)_{ik}}$$

- Then

$$F_{ik} = F_{ik} \frac{(XG)_{ik}}{(FG^T G)_{ik}}$$

Multiplicative Update Method

Algorithm 2 Algorithm to solve NMF.

Initialize F and G

repeat

Update F:

$$F_{ik} \Leftarrow F_{ik} \frac{(XG)_{ik}}{(FG^TG)_{ik}}$$

Update G:

$$G_{jk} \Leftarrow G_{jk} \frac{(X^TF)_{jk}}{(GF^TF)_{jk}}$$

until Converges

Outline

- Matrix Factorization
- NMF
- Optimization Algorithm
- Variants of NMF
- The relationship with K-Means
- Application

The Variants of NMF: Semi-NMF

- Problems:
  - NMF fails to deal with the data with mixed signs
- Semi-NMF:
  - restrict *G* to be nonnegative while placing no restriction on the signs of *F*.

$$X_{\pm} \approx F_{\pm} G_{+}^{T}$$

  - Semi-NMF can be motivated by K-means clustering

$$J_{K-means} = \sum_{i=1}^{n} \sum_{k=1}^{K} g_{ik} \|\mathbf{x}_i - \mathbf{f}_k\|^2 = \|X - FG^T\|^2$$

  - Semi-NMF can be thought as a soft clustering by relaxing the element of *g* from binary to continuous nonnegative values

The Variants of NMF: Orthogonal NMF

- Problems:
  - The solution of NMF F and G are not unique.
- G-orthogonal NMF

$$\min \| X - FG^T \|_F^2$$
$$s.t.\ \ F \geq 0, G \geq 0, G^T G = I$$

- Advantages:
  - uniqueness of the solution
  - Clustering interpretations

The Variants of NMF: Tri-NMF

- Simultaneously cluster rows and columns of the input data matrix $X \in \mathbb{R}_{+}^{p \times n}$

$$
\begin{aligned}
&\min \| X - FSG^T \|_F^2 \\
&s.t.\ \ F \geq 0, G \geq 0, S \geq 0, \\
&F^T F = I, G^T G = I
\end{aligned}
$$

- *Note:*
    - $F \in \mathbb{R}_{+}^{p \times k}$ gives row clusters and $G \in \mathbb{R}_{+}^{n \times \ell}$ gives column clusters

Outline

- Matrix Factorization
- NMF
- Optimization Algorithm
- Variants of NMF
- The relationship with K-Means
- Application

The Relationship with K-Means

- K-Means clustering is one of most widely used clustering method.

Given n points in m-dim: $X = (x_1, x_2, \cdots, x_n)^T$

K-means objective $\min J_K = \sum_{k=1}^{K} \sum_{i \in C_k} \| x_i - c_k \|^2$

The Relationship with K-Means

- Reformulate K-Means Clustering

$$J_K = \sum_i \|x_i\|^2 - \sum_{k=1}^{K} \frac{1}{n_k} \sum_{i,j \in C_k} x_i^T x_j$$

- Cluster membership indicators:

$$h_k = (0 \cdots 0, \overbrace{1 \cdots 1}^{n_k}, 0 \cdots 0)^T / n_k^{1/2}$$

$$J_K = \sum_i x_i^2 - \sum_{k=1}^{K} h_k^T X^T X h_k$$

The Relationship with K-Means

- Objective function of K-Means

$$\max_{H^T H = I, H \geq 0} \operatorname{Tr}(H^T X^T X H)$$

- Replace $W = X^T X$, then

$$\max_{H^T H = I, H \geq 0} \operatorname{Tr}(H^T W H)$$

NMF ⇔ K-Means

- Orthogonal symmetric NMF is equivalent to Kernel K-Means clustering

Symmetric NMF $$\min_{H^T H=I, H\geq 0} \| W - HH^T \|^2$$

Is Equivalence to $$\max_{H^T H=I, H\geq 0} \mathrm{Tr}(H^T W H)$$

NMF ⇔ K-Means

- Factorization is equivalent to Kernel K-means clustering with the strict orthogonality relaxed

$$H = \underset{H^T H = I,\, H \geq 0}{\arg\max}\ Tr(H^T W H)$$

$$= \underset{H^T H = I,\, H \geq 0}{\arg\min}\ -2Tr(H^T W H)$$

$$= \underset{H^T H = I,\, H \geq 0}{\arg\min}\ \| W \|^2 - 2Tr(H^T W H) + \| H^T H \|^2$$

$$= \underset{H^T H = I,\, H \geq 0}{\arg\min}\ \| W - H H^T \|^2$$

Relaxing the orthogonality $H^T H = I$ completes the proof

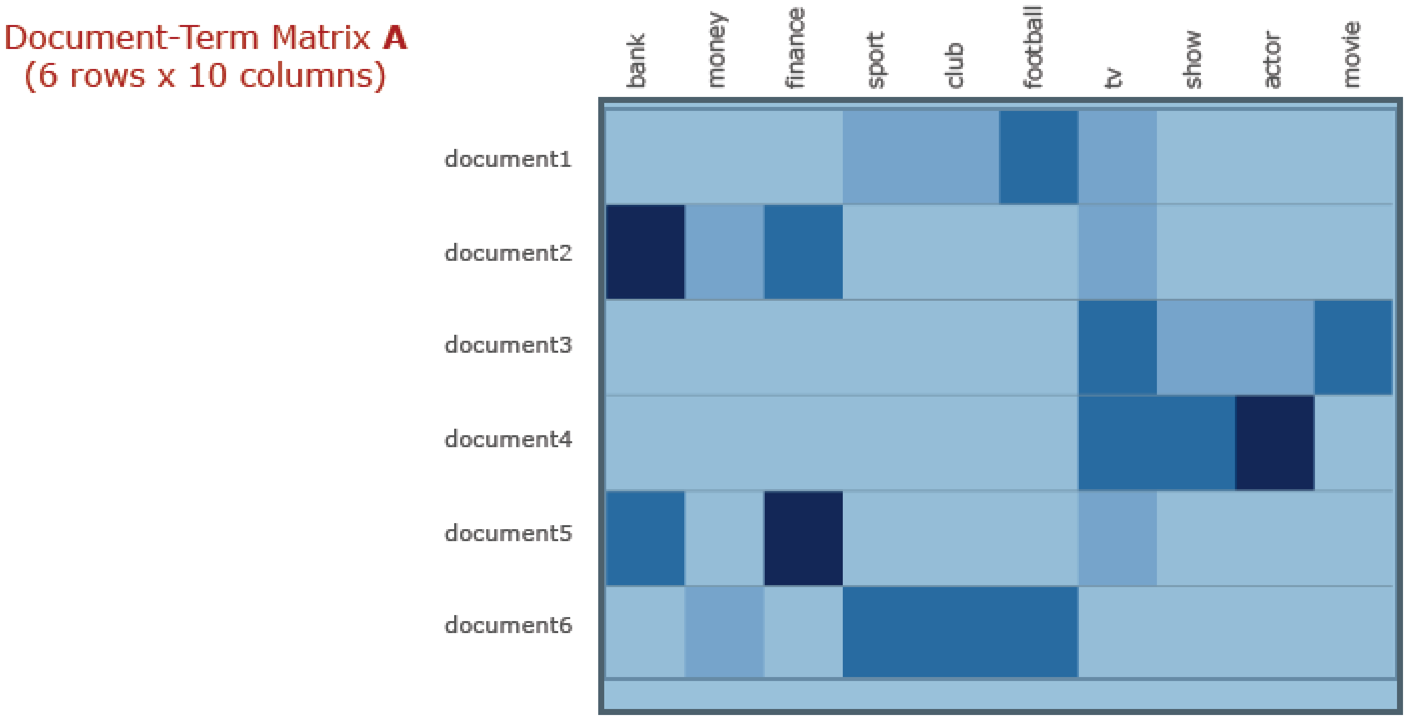
Outline

- Matrix Factorization
- NMF
- Optimization Algorithm
- Variants of NMF
- The relationship with K-Means
- Application

Application Example: Topic Models

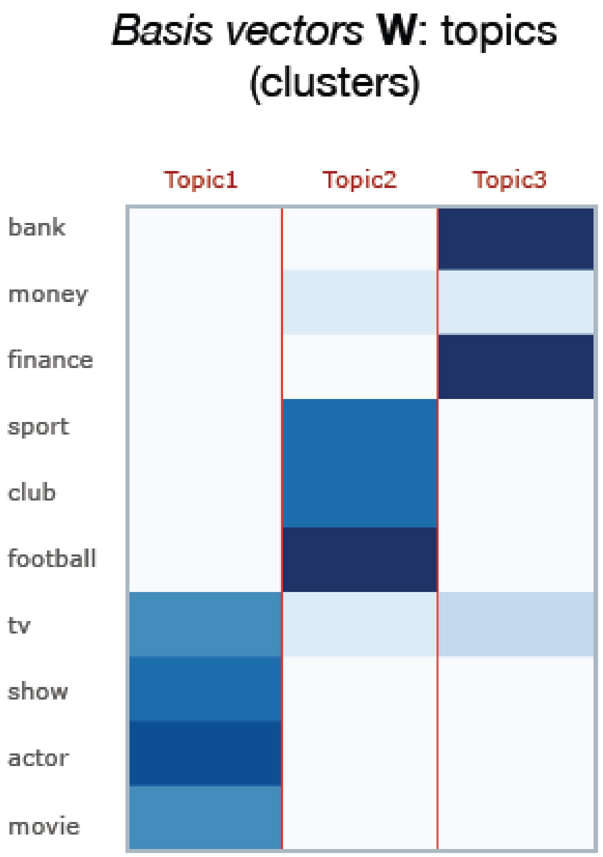
- Algorithm
  - 1. Construct vector space model for documents, resulting in a term-document matrix X.
  - 2. Apply TF-IDF term weight to X.
    - TF-IDF is a statistic that reflect how important a word is to a document
  - 3. Normalize TF-IDF vectors to unit length.
  - 4. Perform NMF on X.
- Output
  - Basis vectors: the topics (clusters) in the data.
  - Coefficient matrix: the membership weights for documents relative to each topic (cluster).

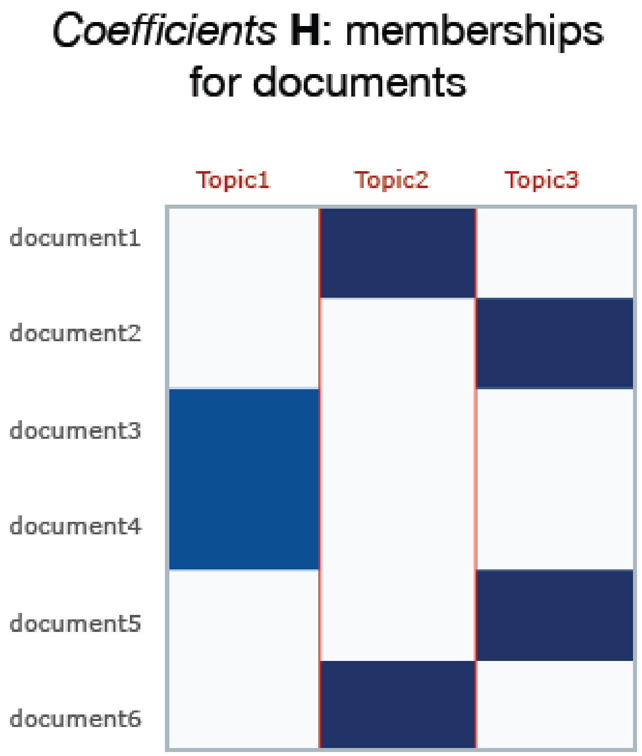
Application Example: Topic Models



- Apply TF-IDF and unit length normalization to rows of **A**.
- Run Euclidean NMF on normalized **A** (k=3, random initialization).

Application Example: Topic Models





Application Example: Face Image

- Given face image data set

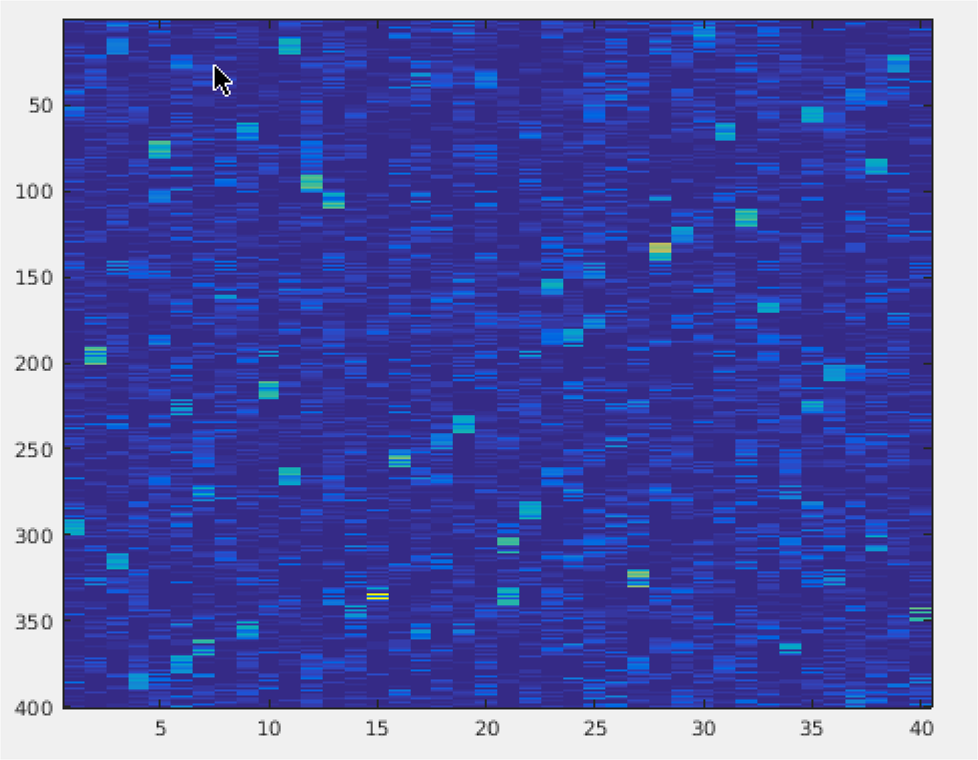
$$[\, f_1,\, f_2, \cdots,\, f_n]$$

Application Example: Face Image

- Output

Basis vector F (reshaped)

Coefficient matrix G



Tips

- Initialization:
  - Random initialize F and G
    - lead to instability
  - K-Means

- How to get the final clustering result?
  - Find the index corresponding to the maximal value in each row of G
  - Perform K-Means on G

Thank you!