Non-negative Matrix Factorization

Gauri Bhagawantrao Jagatap

Term project
IE 631 Non-linear Programming
Iowa State University

April 25, 2017

Outline

Motivation

- Non-negative matrix factorization (NMF) is defined as a decomposition $M \approx WH$ which lies in a low rank subspace, where $M \in \mathbb{R}_+^{d \times n}, W \in \mathbb{R}_+^{d \times r}, H \in \mathbb{R}_+^{r \times n}$ and $r << d, n$.
- Dimensionality reduction, similar to principal component analysis (PCA).
- Matrix factors W and H are non-negative, making them interpretable, in applications such as image segmentation and text mining.

Mathematical Model

The premise of non-negative matrix factorization of positive matrix $M \in \mathbb{R}_+^{d \times n}$ is the minimization

$$\begin{aligned} \min_{W \geq 0, H \geq 0} \quad & \|M - WH\|_F^2 = \sum_{i,j} (M - WH)_{ij}^2 \\ \text{such that} \quad & M(:, i) \approx \sum_{k=1}^{r} W(:, k) H(k, i) \quad \text{for all} \quad i \in \{1, 2 \ldots n\} \\ & M \approx WH \end{aligned}$$

where $W \in \mathbb{R}_+^{d \times r}$ and $H \in \mathbb{R}_+^{r \times n}$.

NMF Illustration

2429 (=n) such examples (25 shown),
each of dimension 19×19 ($= 361 = d$):

NMF Illustration

Vectorized images, are stacked into matrix M.
Any column (or face) can be reconstructed as $M(:, i) \approx WH(:, i)$:

Figure: Original

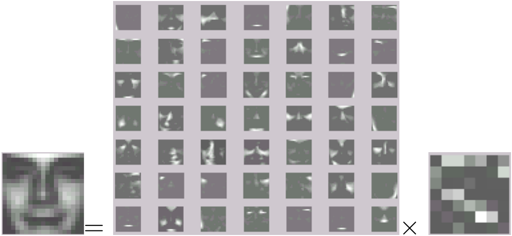

Figure: Reconstruction

Rank considered is $r = 49$.

Standard NMF

Framework

- Employs the block-coordinate descent (BCD) method, which alternatingly minimizes two non-negative least squares (NLS) problems:

$$\min_{H \geq 0} \quad \|M - WH\|_F^2 \quad \text{for fixed } W \tag{1}$$

$$\min_{W^T \geq 0} \quad \|M^T - H^T W^T\|_F^2 \quad \text{for fixed } H^T \tag{2}$$

  until the stopping condition is met, which is determined by KKT conditions.
- Since the technique to solve the two sub-problems is symmetric in H and W^T, one can focus on solving just the NLS in (1).

Standard NMF

Framework

- In [1], Guan, et. al solve the NLS sub-problems in (1) and (2) using Nesterov's optimal gradient method [2].
- In each minimization, the matrix factor (W or H) is updated by using the projected gradient method and a step size which is determined by the Lipschitz constant.

NMF using Nesterov's Optimal Gradient Method

NeNMF

OGM is optimal gradient method.

Input: $M \in \mathbf{R}_+^{d \times n}$, $1 \leq r \leq \min\{d, n\}$
Output: $W \in \mathbf{R}_+^{d \times r}$, $H \in \mathbf{R}_+^{r \times n}$
Initialize: $W^1 \geq 0,\ H^1 \geq 0,\ t = 1$
Repeat:

$$H^{t+1} = OGM(W^t, H^t),$$
$$W^{t+1} = OGM\left(H^{t+1}, W^t\right),$$
$$t \leftarrow t + 1.$$

until: KKT conditions are met for both minimizations.

NMF using Nesterov's Optimal Gradient Method

Optimal Gradient Method

Solving the minimization (OGM):

$$H^{t+1} = \arg\min_{H \geq 0} F(W^t, H) = \frac{1}{2}||M - W^t H||_F^2$$

Input: W^t, H^t

Output: H^{t+1}

Initialize: $Y_0 = H^t,\ \alpha_0 = 1,\ L = ||{W^t}^T W^t||_2,\ k = 0$

Repeat:

$$\begin{aligned}
H_k &= P\left(Y_k - \frac{1}{L}\nabla_H F\left(W^t, Y_k\right)\right), \\
\alpha_{k+1} &= \frac{1 + \sqrt{4\alpha_k^2 + 1}}{2}, \\
Y_{k+1} &= H_k + \frac{\alpha_k - 1}{\alpha_{k+1}}(H_k - H_{k-1}). \\
k &\leftarrow k + 1
\end{aligned}$$

Until: KKT conditions are met.

NMF using Nesterov's Optimal Gradient Method

NeNMF

The crux of this algorithm is in implementing the optimal gradient step:

$$
\begin{aligned}
H_k &= \arg\min_{H \geq 0} \phi(Y_k, H) \\
&= \arg\min_{H \geq 0} F\left(W^t, Y_k\right) + \left\langle \nabla_H F(W^t, Y_k), H - Y_k \right\rangle \\
&\qquad\qquad\qquad\qquad + \frac{L}{2} ||H - Y_k||_F^2 \\
&= P\left(Y_k - \frac{1}{L} \nabla_H F\left(W^t, Y_k\right)\right)^+
\end{aligned}
$$

where the Lipschitz constant is $L = ||{W^t}^T W^t||_2$, $\phi(Y_k, H)$ is the proximal function of $F(W^t, H)$ on Y_k, and Y_k stores the search point:

$$
Y_{k+1} = H_k + \frac{\alpha_k - 1}{\alpha_{k+1}}(H_k - H_{k-1}) \quad \text{where} \quad \alpha_{k+1} = \frac{1 + \sqrt{4\alpha_k^2 + 1}}{2}
$$

NeNMF

Stopping criterion for Optimal Gradient Method

KKT conditions:

$$\nabla_H^P F\left(W^t, H_k\right)_{ij} = 0$$

where

$$\nabla_H^P F\left(W^t, H_k\right)_{ij} = \begin{cases} \nabla_H F\left(W^t, H_k\right)_{ij}, & \left(H_k\right)_{ij} > 0 \\ \min\left\{0, \nabla_H F(W^t, H_k)_{ij}\right\}, & \left(H_k\right)_{ij} = 0. \end{cases}$$

NeNMF

Stopping criterion for NeNMF

$$\nabla_H^P F(W^t, H^t) = \mathbf{0},$$
$$\nabla_W^P F(W^t, H^t) = \mathbf{0}.$$

Separable NMF

Framework

- The separable NMF framework requires an additional assumption that the there exists an index set $\mathcal{K}$, with cardinality less than rank $r < \min(d, n)$ of M.

$$M = M(:, \mathcal{K})H$$

- A subset of r columns of M can approximately generate a convex cone containing all columns of M.
- The goal of separable NMF is to identify the subset of columns with index in $\mathcal{K}$.
- The separability assumption has been used in several applications such as text mining and hyper-spectral imaging.

Separable NMF

Framework

Gillis and Vasavis in [5], demonstrated fast and robust recursive algorithms for separable NMF, using a successive projection algorithm (SPG) to find this subset $\mathcal{K}$. Once $W = M(:, \mathcal{K})$ has been found, the second part of the exercise:

$$\min_{H \geq 0} \quad \|M - WH\|_F^2 \quad \text{for fixed } H$$

is the same as that in the standard NMF framework.

- Lower computational complexity! Instead of solving 2 minimization problems repeatedly, solve both of them *exactly*.
- Only works under the *separability* assumption.

Separable NMF

FastSepNMF

For $f(x) = \|x\|_2^2$,

Input: Let $R = M$, $J = \{\}$, $j = 1$.

Output: $W = M(:, J)$

Repeat:

$$
\begin{aligned}
j^* &= \arg\max_j f(R_{:j}) \\
u_j &= R_{:j^*} \\
R &\leftarrow \left(I - \frac{u_j u_j^T}{||u_j||_2^2} \right) R \\
J &= J \cup \{j^*\} \\
j &= j + 1
\end{aligned}
$$

Until: $R \neq 0$ and $j \leq r$

- Subset of the original TDT2 corpus dataset.
- The largest 30 (=r) categories were retained.
- $9,394$ (=n) documents in total.
- Total number of words in all documents are $36,771$ (=d).

```
Top 5 topics using Nesterov's OGM for NMF...

topic 1:
spkr voice people news president
topic 2:
president clinton lewinsky house white
topic 3:
nuclear india pakistan tests indias
topic 4:
iraq un weapons united iraqi
topic 5:
percent economic government market crisis
```

```
Top 5 topics using Separable NMF...

topic 1:
spkr voice people peterjennings news
topic 2:
iraq un weapons united iraqi
topic 3:
president clinton house lewinsky white
topic 4:
percent market stock economic bank
topic 5:
nuclear india pakistan tests weapons
```

```
Top 7 topics using Nesterov's OGM for NMF...
topic 1:
iraq un weapons united iraqi
topic 2:
spkr voice people president news
topic 3:
world people time olympic team
topic 4:
president clinton lewinsky house white
topic 5:
percent economic government market crisis
topic 6:
tobacco industry companies bill smoking
topic 7:
nuclear india pakistan tests indias
```

```
Top 7 topics using Separable NMF...
topic 1:
spkr voice people peterjennings news
topic 2:
iraq un weapons united iraqi
topic 3:
president clinton house white lewinsky
topic 4:
percent market stock economic bank
topic 5:
nuclear india pakistan tests weapons
topic 6:
tobacco industry smoking companies bill
topic 7:
ms lewinsky tripp lawyers jones
```

Comparison

So which method is better?

- Depends on the application!
  - NeNMF has a computational complexity of $\mathcal{O}(dnr + dr^2 + nr^2) + T\mathcal{O}(dr^2 + nr^2)$, where total number of runs of NeNMF $T < r$.
  - SepFastNMF has a computational complexity of $\mathcal{O}(dnr)$.
  - However, M need not always be separable. NeNMF is a more general framework!

Summary

So why NMF?

- Useful applications in text mining and image segmentation.
- Can be used in non-stationary speech denoising.
- Useful for interpreting common themes/topics in data.
- Helps decompose data into meaningful components.
- Data compression.
- Clustering in gene expression data.

For Further Reading I

- N. Guan, D. Tao, Z. Luo, and B. Yuan, "Nenmf: an optimal gradient method for nonnegative matrix factorization," *IEEE Transactions on Signal Processing*, vol. 60, no. 6, pp. 2882–2898, 2012.
- Y. Nesterov, "A method of solving a convex programming problem with convergence rate o (1/k2),"
- C.-J. Lin, "Projected gradient methods for nonnegative matrix factorization," *Neural computation*, vol. 19, no. 10, pp. 2756–2779, 2007.
- H. Kim and H. Park, "Nonnegative matrix factorization based on alternating nonnegativity constrained least squares and active set method," *SIAM journal on matrix analysis and applications*, vol. 30, no. 2, pp. 713–730, 2008.

For Further Reading II

N. Gillis and S. A. Vavasis, "Fast and robust recursive algorithmsfor separable nonnegative matrix factorization," *IEEE transactions on pattern analysis and machine intelligence*, vol. 36, no. 4, pp. 698–714, 2014.