Algorithms for Non-negative Matrix Factorization (NMF)

D. D. Lee, and H. S. Seung

Adel Elmahdy
University of Minnesota, ECE
9 Mar. 2017
Contents

• What is NMF?

• Applications of NMF.

• Proposed NMF Algorithm.

• Discussion on the algorithm convergence.
Non-Negative Matrix Factorization (NMF)

• Given a matrix $V, n \times m$, with non-negative entires, we seek non-negative factors $W, n \times r$, and $H$, and $r \times m$, such that

$$V \approx WH$$

• Usually, $r$ is chosen to be smaller than $m$ or $n$, i.e., Compressed version of the original data

• Good approximation is achieved when the latent structure in the data $V$ is discovered.
Why Non-Negative?

• Often the data to be analyzed is nonnegative, and the low-rank data are further required to be comprised of nonnegative values in order to avoid contradicting physical realities.

• By not allowing negative entries, NMF enables a non-subtractive combination of parts to form a whole.

• Features may be parts of faces in image data, topics or clusters in textual data, or specific absorption characteristics in hyper-spectral data
Decomposition Model of Simple NMF

D. Kitamura, et al., 2014
Roadmap for the proposed NMF Algorithm

1. Defining the cost function(s).

2. Constructing the optimization problem(s).

3. Introducing the auxiliary function(s) associated with the Residual Error function(s).

4. Deriving the “Multiplicative Update Rule(s)”.

5. Discussing the algorithm convergence.
1. Defining the Cost Function(s)

- Two cost functions are proposed
  - The square of Euclidean Distance
    \[ \| A - B \|^2 = \sum_{i,j} (A_{ij} - B_{ij})^2 \]
  - The Generalized Kullback-Leibler (KL) Divergence
    \[ D (A\|B) = \sum_{i,j} \left( A_{ij} \log \frac{A_{ij}}{B_{ij}} - A_{ij} + B_{ij} \right)^2 \]
2. Constructing the Optimization Problem(s)

• Problem 1:

\[
\min_{W, H} \| V - WH \|
\]

s.t. \( W_{ia} \geq 0, H_{a\mu} \geq 0, \forall i, a, \mu \)

• Problem 2:

\[
\min_{W, H} D (V\|WH)
\]

s.t. \( W_{ia} \geq 0, H_{a\mu} \geq 0, \forall i, a, \mu \)

• **Not** jointly convex in \((W, H)\) !!
3. Auxiliary Functions

- $G(h, h')$ is an auxiliary function for $F(h)$ if $G(h, h') \geq F(h)$, $G(h, h) = F(h)$
3.a. [Auxiliary Function] [Euclidean Distance]

• The auxiliary function for

\[ F(h) = \frac{1}{2} \sum_{i} (v_i - \sum_a W_{ia} h_a)^2 \]

is

\[ G(h, h^t) = F(h^t) + (h - h^t) \nabla F(h^t) + \frac{1}{2} (h - h^t)^T K(h^t)(h - h^t) \]

where

\[ K_{ab}(h^t) = \delta_{ab}(W^T W h^t)_a / h_a^t \]
3.a. [Auxiliary Function] [Euclidean Distance] (Cont’d)

• Proof Sketch

  • \( G(h, h) = F(h) \) is straightforward. We can show that

    \[
    F(h) = F(h^t) + (h - h^t)\nabla F(h^t) + \frac{1}{2}(h - h^t)^T(W^TW)(h - h^t)
    \]

  • In order to verify that \( G(h, h^t) - F(h) \geq 0 \), we need to prove the positive semi-definiteness of \( K(h^t) - W^TW \)

    \[
    G(h, h^t) = F(h^t) + (h - h^t)\nabla F(h^t) + \frac{1}{2}(h - h^t)^TK(h^t)(h - h^t)
    \]

  • It can be shown that \( \nu^T(K(h^t) - W^TW)\nu^T \geq 0 \)
3.b. [Auxiliary Function] [KL Divergence]

- The auxiliary function for

\[
F(h) = \sum_i v_i \log \left( \frac{v_i}{\sum_a W_{ia} h_a} \right) - v_i + \sum_a W_{ia} h_a
\]

is

\[
G(h, h^t) = \sum_i (v_i \log v_i - v_i) + \sum_{ia} W_{ia} h_a
\]

\[
- \sum_i v_i \frac{W_{ia} h_a^t}{\sum_b W_{ib} h_b^t} \left( \log(w_{ia} h_a) - \log\left(\frac{W_{ia} h_a^t}{\sum_b W_{ib} h_b^t}\right) \right)
\]
3.b. [Auxiliary Function] [KL Divergence] (Cont’d)

• **Proof Sketch**

  • \( G(h, h) = F(h) \) is straightforward.

  • In order to verify that \( G(h, h^t) \geq F(h) \), the convexity of the log function is used to derive the inequality

  • For all non-negative \( \alpha_a \) that sum to one,

\[
- \log \sum_a W_{ia} h_a \leq - \sum_a \alpha_a \log \frac{W_{ia} h_a}{\alpha_a}
\]
3.b. [Auxiliary Function] [KL Divergence] (Cont’d)

• Proof Sketch (Cont’d)

• Setting $\alpha_a = \frac{W_{ia} h_a^t}{\sum_b W_{ib} h_b^t}$, we obtain

$$-\log \sum_a W_{ia} h_a \leq - \sum_i v_i \frac{W_{ia} h_a^t}{\sum_b W_{ib} h_b^t} \left( \log(w_{ia} h_a) - \log(\frac{W_{ia} h_a^t}{\sum_b W_{ib} h_b^t}) \right)$$

• From this inequality, it follows that $G(h, h^t) \geq F(h)$. 

$\blacksquare$
4. Multiplicative Update Rule(s)

- The Euclidean Distance (KL Divergence) is non-increasing under the update rules:

  - Euclidean Distance
    \[ W_{ia} \leftarrow W_{ia} \frac{(VHT)_{ia}}{(WHHT)_{ia}} \]

  - KL Divergence
    \[ W_{ia} \leftarrow W_{ia} \frac{\sum_{\mu} H_{a\mu} V_{i\mu} / (WH)_{i\mu}}{\sum_{\nu} H_{a\nu}} \]
    \[ H_{a\mu} \leftarrow H_{a\mu} \frac{(WTV)_{a\mu}}{(WTWH)_{a\mu}} \]
    \[ H_{a\mu} \leftarrow H_{a\mu} \frac{\sum_i W_{ia} V_{i\mu} / (WH)_{i\mu}}{\sum_k W_{ka}} \]
Recall: Auxiliary Functions

- \( G(h, h') \) is an auxiliary function for \( F(h) \) if
  \[
  G(h, h') \geq F(h), \quad G(h, h) = F(h)
  \]

- Lemma: \( F \) is non-increasing under the update
  \[
  h^{t+1} = \arg \min_h G(h, h^t)
  \]  \hspace{1cm} (1)

Proof: \( F(h^{t+1}) \leq G(h^{t+1}, h^t) \leq G(h^t, h^t) = F(h^t) \)
5.a. [Update Rule] [Euclidean Distance]

- Getting the minimum of $G(h, h^t)$ with respect to $h$

\[ h^{t+1} = h^t - K(h^t)^{-1} \nabla F(h^t) G(h, h^t) \]

- Writing the components of this equation explicitly, we obtain

\[ h_a^{t+1} = h_a^t \frac{(W^T v)_a}{(W^T W h^t)_a} G(h, h^t) \]

- By reversing the roles of $W$ and $H$, $F$ can be shown to be non-increasing under the update rules for $W$
5.b. [Update Rule] [KL Divergence]

• Getting the minimum of $G(h, h^t)$ with respect to $h$

\[
\frac{dG(h, h^t)}{dh_a} = - \sum_i v_i \frac{W_{ia} h^t_a}{\sum_b W_{ib} h^t_b} \frac{1}{h_a} + \sum_i W_{ia} = 0
\]

• Thus, the update rule takes the form of

\[
h_a^{t+1} = \frac{h_a^t}{\sum_b W_{kb} \sum_i v_i \sum_b W_{ib} h^t_b} W_{ia}
\]

• By reversing the roles of $W$ and $H$, $F$ can be shown to be non-increasing under the update rules for $W$
5. Algorithm Convergence

- By iterating the update rule, the authors claim that the obtained sequence of estimates converges to a local minimum of the objective function.

\[ h_{\text{min}} = \arg \min_h F(h) \]

- In other words, it is claimed that the Euclidean Distance (KL Divergence) is invariant iff \( W \) and \( H \) are at a stationary point.

- Does the fact that the cost function is non-increasing under these updates imply the convergence to a local minimum?!
### Multiplicative Update Rule Algorithm

**(a)** $m = 25, r = 5, n = 125.$

<table>
<thead>
<tr>
<th>$\epsilon$</th>
<th>Time</th>
<th>#iterations</th>
<th>Objective values</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$10^{-3}$</td>
<td>$10^{-4}$</td>
<td>$10^{-5}$</td>
</tr>
<tr>
<td>mult</td>
<td>0.10</td>
<td>1.22</td>
<td>2.60</td>
</tr>
<tr>
<td>alspgrad</td>
<td><strong>0.03</strong></td>
<td><strong>0.10</strong></td>
<td><strong>0.45</strong></td>
</tr>
<tr>
<td>pgrad</td>
<td>0.05</td>
<td>0.24</td>
<td>0.68</td>
</tr>
<tr>
<td>lsqnonneg</td>
<td>6.32</td>
<td>27.76</td>
<td>57.57</td>
</tr>
</tbody>
</table>

C.-J. Lin, 2007
State-of-the-art NMF Algorithms

- There are two main approaches that have been considered in studying NMF:
  - Optimization-Based Methods
  - Geometry-Based Methods
In a nutshell...

• The authors claim that multiplicative updates converge to a local minimum of the cost function.

• However, the proof that shows that the cost function is non-increasing under these updates, which is slightly different from the convergence to a local minimum.

• Though some papers, such as [2], pointed out its possible slow convergence, this method is popular due to its simplicity.
References


