Non-negative Matrix Factorization and Multiplicative Update

Andersen Ang

Mathématique et recherche opérationnelle
UMONS, Belgium

manshun.ang@umons.ac.be Homepage: angms.science

First draft : Jun 6, 2017
Last update : February 25, 2019

Overview

Introduction to NMF

Non-negative Matrix Factroization (NMF) is a decomposition on non-negative matrix.

A matrix $\mathbf{X} \in \mathbb{R}^{m \times n}$ is called *non-negative* if all the elements in $\mathbf{X}$ are non-negative.

Equivalent notations of non-negativity :

1. $[X]_{ij} \geq 0, \forall i, j$
2. $\mathbf{X} \in \mathbb{R}_{+}^{m \times n}$
3. $\mathbf{X} \geq 0$

The NMF problems

The exact NMF Problem : given $\mathbf{X} \in \mathbb{R}_+^{m \times n}$, find $\mathbf{W} \in \mathbb{R}_+^{m \times r}$ and $H \in \mathbb{R}_+^{r \times n}$ such that $\mathbf{X} = \mathbf{WH}$.

- A very difficult problem : NP-hard !

The NMF problems

The exact NMF Problem : given $\mathbf{X} \in \mathbb{R}_+^{m \times n}$, find $\mathbf{W} \in \mathbb{R}_+^{m \times r}$ and $H \in \mathbb{R}_+^{r \times n}$ such that $\mathbf{X} = \mathbf{WH}$.

- A very difficult problem : NP-hard !

The approximate NMF problem: given $\mathbf{X} \in \mathbb{R}_+^{m \times n}$, find $\mathbf{W} \in \mathbb{R}_+^{m \times r}$ and $H \in \mathbb{R}_+^{r \times n}$ such that $\mathbf{X} \approx \mathbf{WH}$.

i.e., given $\mathbf{X} \in \mathbb{R}_+^{m \times n}$, find $\mathbf{W}$ and $\mathbf{H}$ that minimizes

$$\|\mathbf{X} - \mathbf{WH}\|_\xi$$

where $\|.\|_\xi$ is some norm that measures the discrepancy between $\mathbf{X}$ and $\mathbf{WH}$.

There are many types of $\|.\|_\xi$, here we consider Frobenius norm, that is

$$\|\mathbf{X} - \mathbf{WH}\|_F$$

Minimizing this we arrive at the non-convex optimization problem

$$[\mathbf{W}\ \mathbf{H}] = \underset{\mathbf{W} \geq 0, \mathbf{H} \geq 0}{\arg\min}\ f(\mathbf{W}, \mathbf{H}) = \frac{1}{2}\|\mathbf{X} - \mathbf{WH}\|_F^2$$

Solving the NMF minimization problem

We solve the problem using alternating Gradient Descent (GD).

We first review GD. GD is a first-order iterative optimization algorithm.

For minimizing a single variable function $f(x)$, GD starts with an inital x_0 and iterates the following update:

$$x^{k+1} = x^k - t^k \nabla_x f(x^k)$$

where

- $k \in \mathbb{N}$ is the step counter
- x^k is the current variable
- x^{k+1} is the variable of the next iteration
- $t^k \geq 0$ is the step size
- $\nabla_x f(x)$ is the gradient of the objective function f wit respect to x

Solving the NMF minimization problem - 2

Now instead of one variable, NMF objective function $f(\mathbf{W}, \mathbf{H})$ has two variables: $\mathbf{W}$ and $\mathbf{H}$. So we can either:

1. update $\mathbf{W}$ and $\mathbf{H}$ at the same time
2. update $\mathbf{W}$ and $\mathbf{H}$ separately

No matter which strategy is used, the goal is to produce a sequence $\left\{\mathbf{W}^k, \mathbf{H}^k\right\}_{k \in \mathbb{N}}$ such that the cost function is monotonically decreasing

$$f\left(\mathbf{W}^{k+1}, \mathbf{H}^{k+1}\right) \leq f\left(\mathbf{W}^k, \mathbf{H}^k\right)$$

Commonly the second method is used. That is, we first update $\mathbf{W}$ for a fix $\mathbf{H}$, then update $\mathbf{H}$ for a fix $\mathbf{W}$. Such method belongs to the category of *alternating minimization* or *(block) coordinate descent*.

Solving the NMF by Alternating Gradient Descent

Goal: solve the optimization problem:

$$[\mathbf{W}\ \mathbf{H}] = \underset{\mathbf{W}\geq 0, \mathbf{H}\geq 0}{\arg\min}\ f\left(\mathbf{W}, \mathbf{H}\right) = \frac{1}{2}\|\mathbf{X} - \mathbf{W}\mathbf{H}\|_F^2$$

Method: alternating GD

$$\mathbf{W}^{k+1} = \mathbf{W}^k - t_{\mathbf{W}}^k \nabla_{\mathbf{W}} f(\mathbf{W}^k, \mathbf{H}^k)$$

$$\mathbf{H}^{k+1} = \mathbf{H}^k - t_{\mathbf{H}}^k \nabla_{\mathbf{H}} f(\mathbf{W}^{k+1}, \mathbf{H}^k)$$

To apply alternating GD, we need to know

- $\nabla_{\mathbf{W}} f$ and $\nabla_{\mathbf{H}} f$
- $t_{\mathbf{W}}^k$, $t_{\mathbf{H}}^k$

Deriving the gradient of the NMF objective function

Useful formulae from Linear Algebra 101 and *The Matrix Cookbook*

$$\|\mathbf{X}\|_F^2 = \operatorname{tr}\left(\mathbf{X}^\top \mathbf{X}\right) \tag{1}$$

$$\operatorname{tr}(\mathbf{X}) = \operatorname{tr}\left(\mathbf{X}^\top\right) \tag{2}$$

$$\nabla_{\mathbf{X}} \operatorname{tr}(\mathbf{AXB}) = \mathbf{A}^\top \mathbf{B}^\top \tag{3}$$

$$\nabla_{\mathbf{X}} \operatorname{tr}\left(\mathbf{X}^\top \mathbf{AX}\right) = (\mathbf{A} + \mathbf{A}^\top)\mathbf{X} \tag{4}$$

$$\nabla_{\mathbf{X}} \operatorname{tr}\left(\mathbf{B}^\top \mathbf{X}^\top \mathbf{XB}\right) = 2\mathbf{XBB}^\top \tag{5}$$

Hence

$$\begin{aligned}
\|\mathbf{X} - \mathbf{WH}\|_F^2 &\overset{(1)}{=} \operatorname{tr}\left\{\left(\mathbf{X} - \mathbf{WH}\right)^\top \left(\mathbf{X} - \mathbf{WH}\right)\right\} \\
&= \operatorname{tr}\left\{\mathbf{X}^\top\mathbf{X} - \mathbf{X}^\top\mathbf{WH} - (\mathbf{WH})^\top\mathbf{X} + (\mathbf{WH})^\top\mathbf{WH}\right\} \\
&\overset{(2)}{=} \operatorname{tr}\left(\mathbf{X}^\top\mathbf{X} - 2(\mathbf{WH})^\top\mathbf{X} + \mathbf{H}^\top\mathbf{W}^\top\mathbf{WH}\right) \\
&\overset{(1)}{=} \|\mathbf{X}\|_F^2 - 2\operatorname{tr}(\mathbf{WH})^\top\mathbf{X} + \operatorname{tr}\mathbf{H}^\top\mathbf{W}^\top\mathbf{WH}
\end{aligned}$$

Deriving the gradient of the NMF objective function

So $f(\mathbf{W}, \mathbf{H}) = \frac{1}{2}\|\mathbf{X} - \mathbf{WH}\|_F^2 = \|\mathbf{X}\|_F^2 - \operatorname{tr}\mathbf{X}^\top\mathbf{WH} + \frac{1}{2}\operatorname{tr}\mathbf{H}^\top\mathbf{W}^\top\mathbf{WH}$, and

$$\begin{aligned} \nabla_{\mathbf{W}} f &\overset{(3,5)}{=} (\mathbf{WH} - \mathbf{X})\mathbf{H}^\top \\ \nabla_{\mathbf{H}} f &\overset{(3,4)}{=} \mathbf{W}^\top(\mathbf{WH} - \mathbf{X}) \end{aligned}$$

So we get

$$\begin{aligned} \mathbf{W}^{k+1} &= \mathbf{W}^k - t_{\mathbf{W}}^k(\mathbf{WH} - \mathbf{X})\mathbf{H}^\top \\ \mathbf{H}^{k+1} &= \mathbf{H}^k - t_{\mathbf{H}}^k\mathbf{W}^\top(\mathbf{WH} - \mathbf{X}) \end{aligned}$$

Now the remaining question is on choosing the step sizes t

On the step sizes and the update formulae

There are multiple ways to select step size, but now consider

$$[t_{\mathbf{W}}]_{ij} = \frac{[\mathbf{W}]_{ij}}{[\mathbf{W}\mathbf{H}\mathbf{H}^\top]_{ij}} \qquad [t_{\mathbf{H}}]_{ij} = \frac{[\mathbf{H}]_{ij}}{[\mathbf{W}^\top\mathbf{W}\mathbf{H}]_{ij}}$$

where $[\mathbf{A}]_{ij}$ denotes the element of matrix $\mathbf{A}$ in the ith row jth column.

These step sizes are **element-wise**: instead of a common step size for the whole matrix, each element in the matrix has different step size.

A question naturally follows: why select the step sizes in such way?
Answer : because these step sizes guarantee non-negativity !

Update formulae derivation

Put the gradient and step size we have

$$
\begin{aligned}
[\mathbf{W}]_{ij} &= [\mathbf{W}]_{ij} - [t_{\mathbf{W}}]_{ij} \left[(\mathbf{WH} - \mathbf{X})\mathbf{H}^\top\right]_{ij} \\
&= [\mathbf{W}]_{ij} - \frac{[\mathbf{W}]_{ij}}{[\mathbf{WHH}^\top]_{ij}} \left[(\mathbf{WH} - \mathbf{X})\mathbf{H}^\top\right]_{ij} \\
&= [\mathbf{W}]_{ij} - \frac{\left[\mathbf{W}(\mathbf{WH} - \mathbf{X})\mathbf{H}^\top\right]_{ij}}{[\mathbf{WHH}^\top]_{ij}} \\
&= [\mathbf{W}]_{ij} \left(1 - \frac{\left[(\mathbf{WH} - \mathbf{X})\mathbf{H}^\top\right]_{ij}}{[\mathbf{WHH}^\top]_{ij}}\right) \\
&= [\mathbf{W}]_{ij} \left(\frac{[\mathbf{WHH}^\top]_{ij}}{[\mathbf{WHH}^\top]_{ij}} - \frac{\left[(\mathbf{WH} - \mathbf{X})\mathbf{H}^\top\right]_{ij}}{[\mathbf{WHH}^\top]_{ij}}\right) \\
&= [\mathbf{W}]_{ij} \frac{\left[\mathbf{XH}^\top\right]_{ij}}{[\mathbf{WHH}^\top]_{ij}}
\end{aligned}
$$

All elements in $\mathbf{X}$, $\mathbf{W}$, $\mathbf{H}$ non-negative $\implies$ result also non-negative.

Update formulae derivation - 2

$$
\begin{aligned}
[\mathbf{H}]_{ij} &= [\mathbf{H}]_{ij} - [t_{\mathbf{H}}]_{ij} \left[\mathbf{W}^\top(\mathbf{WH}-\mathbf{X})\right]_{ij} \\
&= [\mathbf{H}]_{ij} - \frac{[\mathbf{H}]_{ij}}{[\mathbf{W}^\top\mathbf{WH}]_{ij}} \mathbf{W}^\top(\mathbf{WH}-\mathbf{X}) \\
&= [\mathbf{H}]_{ij} - \frac{[\mathbf{H}]_{ij}}{[\mathbf{W}^\top\mathbf{WH}]_{ij}} \left[\mathbf{W}^\top(\mathbf{WH}-\mathbf{X})\right]_{ij} \\
&= [\mathbf{H}]_{ij} \left(1 - \frac{\left[\mathbf{W}^\top(\mathbf{WH}-\mathbf{X})\right]_{ij}}{[\mathbf{W}^\top\mathbf{WH}]_{ij}}\right) \\
&= [\mathbf{H}]_{ij} \left(\frac{\left[\mathbf{W}^\top\mathbf{WH}\right]_{ij}}{[\mathbf{W}^\top\mathbf{WH}]_{ij}} - \frac{\left[\mathbf{W}^\top(\mathbf{WH}-\mathbf{X})\right]_{ij}}{[\mathbf{W}^\top\mathbf{WH}]_{ij}}\right) \\
&= [\mathbf{H}]_{ij} \frac{\left[\mathbf{W}^\top\mathbf{X}\right]_{ij}}{[\mathbf{W}^\top\mathbf{WH}]_{ij}}
\end{aligned}
$$

All elements in $\mathbf{X}$, $\mathbf{W}$, $\mathbf{H}$ non-negative $\Longrightarrow$ result also non-negative.

Update formulae

Using $\circ$ to denote element-wise multiplication and noting that the fraction is element-wise, we have

$$\mathbf{W}^{k+1} = \mathbf{W}^k \circ \frac{\mathbf{X}\mathbf{H}^\top}{\mathbf{W}\mathbf{H}\mathbf{H}^\top} \quad \mathbf{H}^{k+1} = \mathbf{H}^k \circ \frac{{\mathbf{W}^k}^\top \mathbf{X}}{(\mathbf{W}^k)^\top \mathbf{W}^k \mathbf{H}^k}$$

This update is called *Multiplicative Update* (MU).

Properties :

- Theory of MU : the objective function $\|\mathbf{X} - \mathbf{W}^k\mathbf{H}^k\|_F^2$ is non-increasing with respect to MU. See the proof here.
- However, the iterates of MU is no guaranteed to converge to a 1st-order stationary point of f.
- To keep non-negativity, the step size of MU is "shrinked", hence MU has slow convergence.

Last page - summary

- Introduction of NMF
- Formulation of NMF as a minimization problem

$$[\mathbf{W}\ \mathbf{H}] = \underset{\mathbf{W}\geq 0, \mathbf{H}\geq 0}{\arg\min}\ f\left(\mathbf{W}, \mathbf{H}\right) = \frac{1}{2}\|\mathbf{X} - \mathbf{W}\mathbf{H}\|_F^2$$

- Derivation of gradient descent update

$$\mathbf{W}^{k+1} = \mathbf{W}^k \circ \frac{\mathbf{X}(\mathbf{H}^k)^\top}{\mathbf{W}^k\mathbf{H}^k(\mathbf{H}^k)^\top} \qquad \mathbf{H}^{k+1} = \mathbf{H}^k \circ \frac{{\mathbf{W}^k}^\top\mathbf{X}}{(\mathbf{W}^k)^\top\mathbf{W}^k\mathbf{H}^k}$$

- Non-negativity is retained with the specially selected step size

End of document