Matrix Factorization in vectorized format

Which you shouldn't do!

Andersen Ang

Mathématique et recherche opérationnelle
UMONS, Belgium

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

First draft : July 14, 2019
Last update : July 29, 2020

Problem setup

- Matrix Factorization (MF) problem
  Given $\mathbf{M} \in \mathbb{R}^{m \times n}$ and a positive integer r, find $\mathbf{W} \in \mathbb{R}^{m \times r}$ and $\mathbf{H} \in \mathbb{R}^{r \times n}$ by solving the optimization problem

$$(\mathcal{P}) \ : \ \underset{\mathbf{W},\mathbf{H}}{\text{minimize}} f(\mathbf{W}, \mathbf{H})$$

  where

$$f(\mathbf{W}, \mathbf{H}) = \frac{1}{2} \|\mathbf{M} - \mathbf{W}\mathbf{H}\|_F^2.$$

- We will illustrate how to solve this problem in matrix-wise and vector-wise format.

Matrix-form algorithm

Consider we solve $(\mathcal{P})$ by alternating gradient update

---

Algorithm 1: Alternating gradient update for solving $(\mathcal{P})$

Result: $\mathbf{W}, \mathbf{H}$ that approximately solve the problem

Initialize $\mathbf{W}_0, \mathbf{H}_0$;

for $k = 1, 2, \ldots$ **do**

$\quad \mathbf{W}_{k+1} = \mathbf{W}_k - \alpha_k^{\mathbf{W}} \nabla_{\mathbf{W}} f(\mathbf{W}_k, \mathbf{H}_k);$

$\quad \mathbf{H}_{k+1} = \mathbf{H}_k - \alpha_k^{\mathbf{H}} \nabla_{\mathbf{H}} f(\mathbf{W}_{k+1}, \mathbf{H}_k);$

end

---

where $\alpha_k^{\mathbf{W}}$, $\alpha_k^{\mathbf{H}}$ are stepsizes.

A simple rule for stepsize selection is to use the inverse Lipschitz constant

$$\begin{aligned} \alpha_k^{\mathbf{W}} &= \frac{1}{\|\mathbf{H}\mathbf{H}^\top\|_2}, \\ \alpha_k^{\mathbf{H}} &= \frac{1}{\|\mathbf{W}^\top\mathbf{W}\|_2}. \end{aligned}$$

The vec operator and the Kronecker product

- To update the variable in vectorized format, we need two tools: the vec operator and the Kronecker product.

- Give $\mathbf{A} = [\mathbf{a}_1 \ \mathbf{a}_2 \ \ldots \ \mathbf{a}_n]$ where $\mathbf{a}_i$ is the i^{th} column of $\mathbf{A}$, the operator vec: $\mathbb{R}^{m \times n} \to \mathbb{R}^{mn}$ turns $\mathbf{A}$ into a long column vector as

$$\text{vec}\mathbf{A} = \begin{bmatrix} \mathbf{a}_1 \\ \mathbf{a}_2 \\ \vdots \\ \mathbf{a}_n \end{bmatrix}.$$

- For Kronecker product, given two matrices $\mathbf{P} \in \mathbb{R}^{m \times n}$ and $\mathbf{Q} \in \mathbb{R}^{p \times q}$

$$\mathbf{P} \boxtimes \mathbf{Q} = \begin{bmatrix} P_{11}\mathbf{Q} & P_{12}\mathbf{Q} & \ldots & P_{1n}\mathbf{Q} \\ P_{21}\mathbf{Q} & P_{22}\mathbf{Q} & \ldots & P_{2n}\mathbf{Q} \\ \vdots & \vdots & \vdots & \ddots \\ P_{m1}\mathbf{Q} & P_{m2}\mathbf{Q} & \ldots & P_{mn}\mathbf{Q} \end{bmatrix} \in \mathbb{R}^{mp \times nq}.$$

  In fact, the expression of the Kronecker product is a two-dimensional compression of the four-dimensional tensor product.

Useful lemma and the vector expression

- **Lemma** For $\mathbf{A} \in \mathbb{R}^{m \times n}, \mathbf{X} \in \mathbb{R}^{n \times p}$ and $\mathbf{B} \in \mathbb{R}^{p \times q}$,

$$\text{vec}(\mathbf{AXB}) = (\mathbf{B}^\top \boxtimes \mathbf{A})\text{vec}(\mathbf{X}).$$

- From this lemma, with the fact that the vec operator is distributive, then

$$\begin{aligned} \|\mathbf{M} - \mathbf{WH}\|_F &= \|\text{vec}(\mathbf{M} - \mathbf{WH})\|_2 \\ &= \|\text{vec}\mathbf{M} - \text{vec}(\mathbf{WH})\|_2 \\ &= \|\text{vec}\mathbf{M} - (\mathbf{H}^\top \boxtimes \mathbf{I}_m)\text{vec}(\mathbf{W})\|_2 \quad \mathbf{W} \text{ as subject} \\ &= \|\text{vec}\mathbf{M} - (\mathbf{I}_n \boxtimes \mathbf{W})\text{vec}(\mathbf{H})\|_2 \quad \mathbf{H} \text{ as subject} \end{aligned}$$

- The last two expressions are in the form $\|\mathbf{Ax} - \mathbf{b}\|_2$:
  For $\mathbf{W}$, we have $\mathbf{A} = \mathbf{H}^\top \boxtimes \mathbf{I}_m$, $\mathbf{b} = \text{vec}\mathbf{M}$ and $\mathbf{x} = \text{vec}\mathbf{W}$.
  For $\mathbf{H}$, we have $\mathbf{A} = \mathbf{I}_n \boxtimes \mathbf{W}$, $\mathbf{b} = \text{vec}\mathbf{M}$ and $\mathbf{x} = \text{vec}\mathbf{H}$.

The same algorithm in vector-form

The same algorithm in the vector-form can be expressed as follows.

Algorithm 2: Alternating gradient update for solving $(\mathcal{P})$, vector-form

Result: $\mathbf{W}, \mathbf{H}$ that approximately solve the problem

Set $\mathbf{b} = \text{vec}(\mathbf{M})$;

Initialize $\mathbf{W}_0, \mathbf{H}_0$;

for $i = 1, 2, \ldots$ **do**

- $\mathbf{A} = \mathbf{H}^\top \boxtimes \mathbf{I}_m$;
- $\mathbf{x} = \text{vec}\mathbf{W}$;
- $\mathbf{x} = \mathbf{x} - \alpha(\mathbf{A}^\top \mathbf{A}\mathbf{x} - \mathbf{A}^\top \mathbf{b})$, $\alpha = (\|\mathbf{A}^\top \mathbf{A}\|_2)^{-1}$;
- $\mathbf{W} = \text{vec}^{-1}(\mathbf{x})$;

- $\mathbf{A} = \mathbf{I}_n \boxtimes \mathbf{W}$;
- $\mathbf{x} = \text{vec}\mathbf{H}$;
- $\mathbf{x} = \mathbf{x} - \alpha(\mathbf{A}^\top \mathbf{A}\mathbf{x} - \mathbf{A}^\top \mathbf{b})$, $\alpha = (\|\mathbf{A}^\top \mathbf{A}\|_2)^{-1}$;
- $\mathbf{H} = \text{vec}^{-1}(\mathbf{x})$;

end

where vec^{-1} is the inverse operation of vec.

Improving the computation of the vector-form algorithm

- To improve the computation of the vector-form algorithm, we can make use of the lemma $\text{vec}(\mathbf{AXB}) = (\mathbf{B}^\top \boxtimes \mathbf{A})\text{vec}(\mathbf{X})$, which implies

  $$(\mathbf{C} \boxtimes \mathbf{D})\mathbf{y} \;=\; \text{vec}(\mathbf{C}\text{vec}^{-1}(\mathbf{y})\mathbf{D}^\top).$$

  So for the $\mathbf{A}^\top \mathbf{b}$ term on $\mathbf{W}$:

  $$\begin{aligned}\mathbf{A}^\top \mathbf{b} &= (\mathbf{H}^\top \boxtimes \mathbf{I}_m)^\top \text{vec}\mathbf{M} \\ &= (\mathbf{I}_m \boxtimes \mathbf{H})\text{vec}\mathbf{M} \\ &= \text{vec}(\mathbf{I}_m \text{vec}^{-1}(\text{vec}\mathbf{M})\mathbf{H}^\top) \\ &= \text{vec}(\mathbf{I}_m \mathbf{M}\mathbf{H}^\top) \\ &= \text{vec}(\mathbf{M}\mathbf{H}^\top)\end{aligned}$$

- That is, to compute $\mathbf{A}^\top \mathbf{b}$ for the update of $\text{vec}\mathbf{W}$, in fact we do not need to compute it as $(\mathbf{H}^\top \boxtimes \mathbf{I}_m)^\top \text{vec}\mathbf{M}$, which is expensive. We can compute $\mathbf{b}$ simpliy as the vectorization of the product $\mathbf{M}\mathbf{H}^\top$. This also agrees with the matrix-form algorithm: When we update the matrix variable $\mathbf{W}$, we do have to compute the term $\mathbf{M}\mathbf{H}^\top$.

The improved algorithm in vector-form

Algorithm 3: Alternating gradient update for solving $(\mathcal{P})$, vector-form

Result: $\mathbf{W}, \mathbf{H}$ that approximately solve the problem

Initialize $\mathbf{W}_0, \mathbf{H}_0$;

for $i = 1, 2, \ldots$ **do**

$\mathbf{A} = \mathbf{H}^\top \boxtimes \mathbf{I}_m$;

$\mathbf{x} = \text{vec}\mathbf{W}$;

$\mathbf{x} = \mathbf{x} - \alpha\Big(\mathbf{A}^\top \mathbf{A}\mathbf{x} - \text{vec}(\mathbf{M}\mathbf{H}^\top)\Big)$, $\alpha = (\|\mathbf{A}^\top \mathbf{A}\|_2)^{-1}$;

$\mathbf{W} = \text{vec}^{-1}(\mathbf{x})$;

$\mathbf{A} = \mathbf{I}_n \boxtimes \mathbf{W}$;

$\mathbf{x} = \text{vec}\mathbf{H}$;

$\mathbf{x} = \mathbf{x} - \alpha\Big(\mathbf{A}^\top \mathbf{A}\mathbf{x} - \text{vec}(\mathbf{W}^\top \mathbf{M})\Big)$, $\alpha = (\|\mathbf{A}^\top \mathbf{A}\|_2)^{-1}$;

$\mathbf{H} = \text{vec}^{-1}(\mathbf{x})$;

end

Drawbacks of the vectorization approach on MF

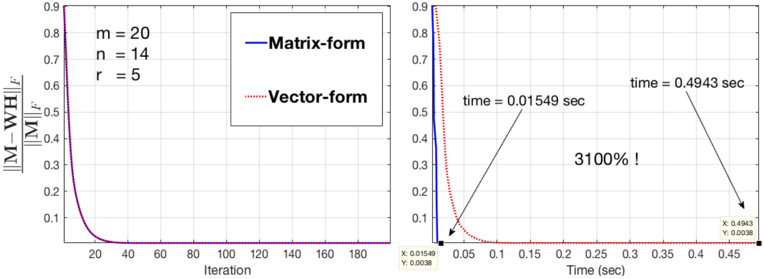
- Dimension explosion.
  - In the matrix case, we have $\mathbf{M}, \mathbf{W}, \mathbf{H}$ as m-by-n, m-by-r and r-by-n.
  - In the vector case, vec$\mathbf{X}$ is mn-by-1, vec$\mathbf{W}$ is mr-by-1 and vec$\mathbf{H}$ is nr-by-1, and finally $(\mathbf{H}^\top \boxtimes \mathbf{I}_m)$ is rm-by-nm and $(\mathbf{I}_n \boxtimes \mathbf{W})$ is nm-by-nr. These make the computation of $\mathbf{A}$, $\mathbf{A}^\top\mathbf{A}$, $\mathbf{A}^\top\mathbf{b}$ and α much more expansive than the matrix version.

- Additional structure of matrix – the rank structure is loss by the vectorization, and hence some regularizer such as the nuclear norm $\|\cdot\|_*$ cannot be accommodate by the algorithm:
  - In the vector-form algorithm, nuclear norm has no equivalent expression in vectorized format.
  - In the matrix-form algorithm, nuclear norm can be handled by proximal gradient, which resulted in the singular value thresholding operator.

  The vector-form can only handle regularizers that are separable : such as F-norm (sum of l_2 norm of all elements), column-wise l_1 norm.

Dimension explosion

- The biggest drawback of the vectorization approach is dimension explosion, which leads to high computational load.
- For example, for $m = 20$, $n = 14$, $r = 5$



MATLAB code

- In terms of iteration, both algorithms perform the same, which is as expected since the two algorithms are exactly the same iteration-wise.
- In terms of computational time, the runtime of the vector-form algorithm is 3100% of that of the matrix-form algorithm.

Last page - summary

- A useful lemma

$$\text{vec}(\mathbf{A}\mathbf{X}\mathbf{B}) = (\mathbf{B}^\top \boxtimes \mathbf{A})\text{vec}(\mathbf{X}),$$

  Or equivalently

$$(\mathbf{C} \boxtimes \mathbf{D})\mathbf{y} \quad = \quad \text{vec}(\mathbf{C}\text{vec}^{-1}(\mathbf{y})\mathbf{D}^\top).$$

- MF problem in vectorized format
- MF algorithm in vectorized format
- Drawback of the vectorized format : dimension expansion $\implies$ high computational load $\implies$ slow algorithm

- END of Document -