Algorithms for Non-negative Matrix Factorization

- By D. D. Lee and H. S. Seung,2001

Contents

- NMF
- Cost Functions
- Multiplicative update rules
- Interpretation as gradient descent
- Proving the monotonic convergence

NMF

- Nonnegative Matrix: $V \geq 0 \Leftrightarrow \forall \mathrm{i,j}\ V_{ij} \geq 0$
- Non-negative matrix factorization(NMF):Using two non-negative matrices to approximate another non-negative matrix.

$$\begin{cases} \boldsymbol{V} \approx \boldsymbol{WH} (\boldsymbol{V} \in R^{n\times m}, \boldsymbol{W} \in \mathrm{R}^{n\times r}, \boldsymbol{H} \in \mathrm{R}^{r\times m}) \\ \boldsymbol{W} \geq 0, \boldsymbol{H} \geq 0 \end{cases}$$

- Significance: Relatively few basis vectors are used to represent many data vectors.

Cost Function

- Quantify the quality of approximation.

- Square of the Euclidean distance between *A* and *B*

$$\left\| \boldsymbol{A} - \boldsymbol{B} \right\|^2 = \sum_{ij} \left(\boldsymbol{A}_{ij} - \boldsymbol{B}_{ij} \right)^2 \geq 0$$

- Generalized Kullback-Leibler divergence of A from B

$$D\left(\boldsymbol{A} \| \boldsymbol{B} \right) = \sum_{ij} \left(\boldsymbol{A}_{ij} \log \frac{\boldsymbol{A}_{ij}}{\boldsymbol{B}_{ij}} - \boldsymbol{A}_{ij} + \boldsymbol{B}_{ij} \right) \geq 0$$

Optimization Problems

- Problem 1: $$\begin{cases} \min\limits_{\boldsymbol{W},\boldsymbol{H}} \left\| \boldsymbol{V} - \boldsymbol{W}\boldsymbol{H} \right\| \\ \boldsymbol{W},\boldsymbol{H} \geq 0 \end{cases}$$

- Problem 2: $$\begin{cases} \min\limits_{\boldsymbol{W},\boldsymbol{H}} D\left(\boldsymbol{V} \,\middle\|\, \boldsymbol{W}\boldsymbol{H} \right) \\ \boldsymbol{W},\boldsymbol{H} \geq 0 \end{cases}$$

How to tackle the problems above?

- They are convex in W only or H only, but not convex in both variables together.
- Goal: Finding local minima

Gradient decent?

- Convergence can be slow
- Sensitive to the step size

Conjugate gradient?

- More complicated to complement

Multiplicative update rules

Euclidean distance	Divergence
$\boldsymbol{H}_{a\mu} \leftarrow \boldsymbol{H}_{a\mu} \frac{(\boldsymbol{W}^T\boldsymbol{V})_{a\mu}}{(\boldsymbol{W}^T\boldsymbol{W}\boldsymbol{H})_{a\mu}}$	$\boldsymbol{H}_{a\mu} \leftarrow \boldsymbol{H}_{a\mu} \frac{\sum_i \boldsymbol{W}_{ia}\boldsymbol{V}_{i\mu}/(\boldsymbol{W}\boldsymbol{H})_{i\mu}}{\sum_k W_{ka}}$
$\boldsymbol{W}_{ia} \leftarrow \boldsymbol{W}_{ia} \frac{(\boldsymbol{V}\boldsymbol{H}^T)_{ia}}{(\boldsymbol{W}\boldsymbol{H}\boldsymbol{H}^T)_{ia}}$	$\boldsymbol{W}_{ia} \leftarrow \boldsymbol{W}_{ia} \frac{\sum_\mu \boldsymbol{H}_{a\mu}\boldsymbol{V}_{i\mu}/(\boldsymbol{W}\boldsymbol{H})_{i\mu}}{\sum_v \boldsymbol{H}_{av}}$

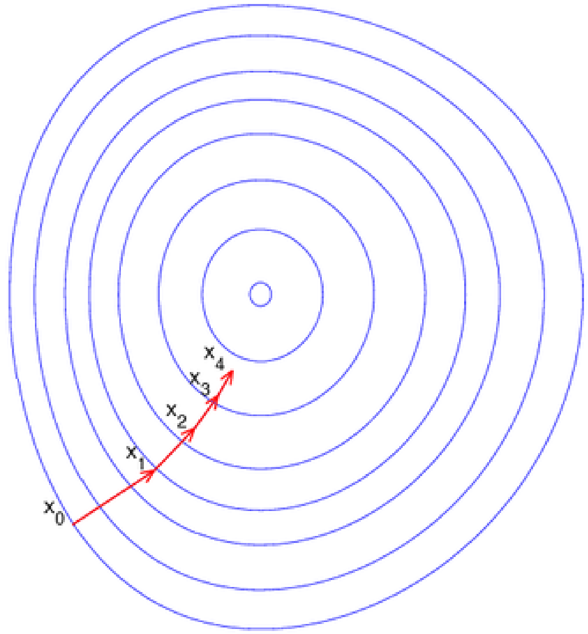
The Euclidean distance and divergence are not increasing under the update rules.

Gradient decent

- Gradient decent for $\mathrm{f}(\boldsymbol{x})$

$$\boldsymbol{x}_{n+1}=\boldsymbol{x}_n-\lim_{\eta\to 0^+}\eta\frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{x}},n\geq 0\Rightarrow f(\boldsymbol{x}_0)\geq f(\boldsymbol{x}_1)\geq f(\boldsymbol{x}_2)\geq....$$

- The sequence $\{x_n\}$ converges to the desired local minimum.



Multiplicative vs additive update rules

- Euclidean distance

$$H_{a\mu} \leftarrow H_{a\mu} + \eta_{a\mu}\left[\left(W^T V\right)_{a\mu} - \left(W^T W H\right)_{a\mu}\right]$$

$$\eta_{a\mu} = \frac{H_{a\mu}}{\left(W^T W H\right)_{a\mu}}$$

$$H_{a\mu} \leftarrow H_{a\mu} \frac{\left(W^T V\right)_{a\mu}}{\left(W^T W H\right)_{a\mu}}$$

- Divergence

$$H_{a\mu} \leftarrow H_{a\mu} + \eta_{a\mu}\left[\sum_i W_{ia} \frac{V_{i\mu}}{(WH)_{i\mu}} - \sum_i W_{ia}\right]$$

$$\eta_{a\mu} = \frac{H_{a\mu}}{\sum_i W_{ia}}$$

$$H_{a\mu} \leftarrow H_{a\mu} \frac{\sum_i W_{ia} V_{i\mu} / (WH)_{i\mu}}{\sum_k W_{ka}}$$

- Is it convergent, even if η is not necessarily small enough?

Proofs of converge

- Define an auxiliary function $G(h, h^t)$ for $F(h)$

$$G(\boldsymbol{h},\boldsymbol{h}^t) \geq F(\boldsymbol{h}), G(\boldsymbol{h},\boldsymbol{h}) = F(\boldsymbol{h})$$

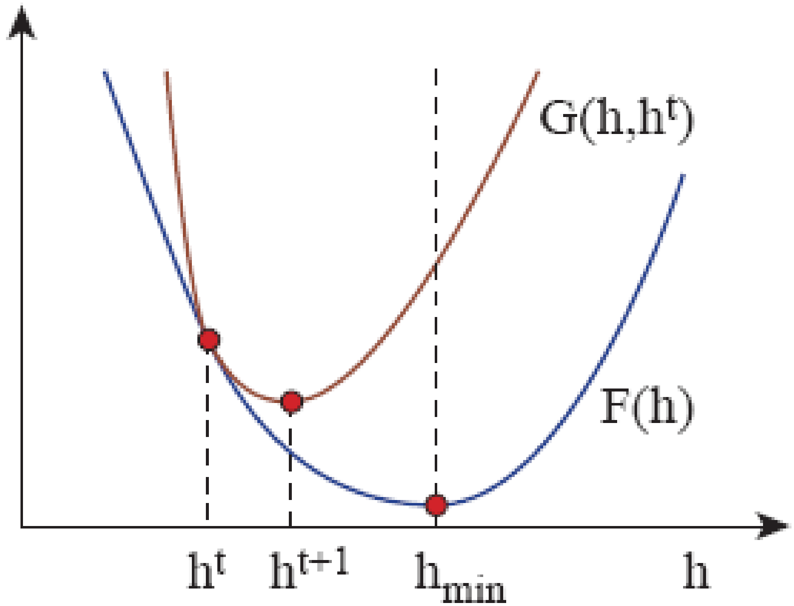
- Find a local minimum of G by iterating the update

$$\boldsymbol{h}^{t+1} = \arg\min_{h} G(\boldsymbol{h},\boldsymbol{h}^t)$$

- The sequence converges to a local minimum of $F(h)$

$$F(\boldsymbol{h}^{t+1}) \leq G(\boldsymbol{h}^{t+1},\boldsymbol{h}^t) \leq G(\boldsymbol{h}^t,\boldsymbol{h}^t) \leq F(\boldsymbol{h}^t)$$

Auxiliary Function



$$\boldsymbol{h}^{t+1} = \arg\min_{h} G(\boldsymbol{h}, \boldsymbol{h}^{t})$$

Updates for H of Euclidean distance

$$F(\boldsymbol{h})=\frac{1}{2}\sum_i\left(\boldsymbol{v}_i-\sum_a \boldsymbol{W}_{ia}\boldsymbol{h}_a\right)^2$$

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

$$K_{ab}(\boldsymbol{h}^t)=\delta_{ab}(\boldsymbol{W}^T\boldsymbol{W}\boldsymbol{h}^t)_a/\boldsymbol{h}_a^t$$

Proving steps:

Step one

- Show $G(h,h^t)$ is an auxiliary function for $F(h)$

Step two

- Obtain the update rule by setting the gradient of $G(h,h^t)$ to zero

Step three

- Check the equivalence between the update rule and $\boldsymbol{H}_{a\mu}\leftarrow\boldsymbol{H}_{a\mu}\frac{\left(\boldsymbol{W}^T\boldsymbol{V}\right)_{a\mu}}{\left(\boldsymbol{W}^T\boldsymbol{W}\boldsymbol{H}\right)_{a\mu}}$

Auxiliary function $G(\boldsymbol{h},\boldsymbol{h}^t)$ for $F(\boldsymbol{h})$

- $F(\boldsymbol{h}) = \frac{1}{2}\sum_i(\boldsymbol{v}_i - \sum_a \boldsymbol{W}_{ia}\boldsymbol{h}_a)^2$
- $F(\boldsymbol{h}) = F(\boldsymbol{h}^t)+(\boldsymbol{h}-\boldsymbol{h}^t)^T\nabla F(\boldsymbol{h}^t)+\frac{1}{2}(\boldsymbol{h}-\boldsymbol{h}^t)^T(\boldsymbol{W}^T\boldsymbol{W})(\boldsymbol{h}-\boldsymbol{h}^t)$
- $G(\boldsymbol{h},\boldsymbol{h}^t) = F(\boldsymbol{h}^t)+(\boldsymbol{h}-\boldsymbol{h}^t)^T\nabla F(\boldsymbol{h}^t)+\frac{1}{2}(\boldsymbol{h}-\boldsymbol{h}^t)^T K(\boldsymbol{h}^t)(\boldsymbol{h}-\boldsymbol{h}^t)$

$$G(\boldsymbol{h},\boldsymbol{h}^t) \geq F(\boldsymbol{h}) \Leftarrow (\boldsymbol{h}-\boldsymbol{h}^t)^T[K(\boldsymbol{h}^t)-\boldsymbol{W}^T\boldsymbol{W}](\boldsymbol{h}-\boldsymbol{h}^t) \geq 0$$

$$\boldsymbol{M}_{ab}(\boldsymbol{h}^t) = \boldsymbol{h}_a^t[\boldsymbol{K}(\boldsymbol{h}^t)-\boldsymbol{W}^T\boldsymbol{W}]_{ab}\boldsymbol{h}_b^t$$

$$\begin{aligned}
\boldsymbol{v}^T\boldsymbol{M}\boldsymbol{v} &= \sum_{ab}\boldsymbol{v}_a\boldsymbol{M}_{ab}\boldsymbol{v}_b = \sum_{ab}\boldsymbol{h}_a^t(\boldsymbol{W}^T\boldsymbol{W})_{ab}\boldsymbol{h}_b^t\boldsymbol{v}_a^2 - \boldsymbol{v}_a\boldsymbol{h}_a^t(\boldsymbol{W}^T\boldsymbol{W})_{ab}\boldsymbol{h}_b^t\boldsymbol{v}_b \\
&= \sum_{ab}(\boldsymbol{W}^T\boldsymbol{W})_{ab}\boldsymbol{h}_a^t\boldsymbol{h}_b^t[\frac{1}{2}\boldsymbol{v}_a^2+\frac{1}{2}\boldsymbol{v}_b^2-\boldsymbol{v}_a\boldsymbol{v}_b] \\
&= \frac{1}{2}\sum_{ab}(\boldsymbol{W}^T\boldsymbol{W})_{ab}\boldsymbol{h}_a^t\boldsymbol{h}_b^t(\boldsymbol{v}_a+\boldsymbol{v}_b)^2 \geq 0
\end{aligned}$$

Minimum of $G(\boldsymbol{h}, \boldsymbol{h}^t)$ and update rules

$$\frac{\partial G(\boldsymbol{h},\boldsymbol{h}^t)}{\partial \boldsymbol{h}} = \nabla F(\boldsymbol{h}^t) + (\boldsymbol{h}-\boldsymbol{h}^t)K(\boldsymbol{h}^t) = 0 \Rightarrow \boldsymbol{h}^{t+1} = \boldsymbol{h}^t - K(\boldsymbol{h}^t)^{-1}\nabla F(\boldsymbol{h}^t)$$

$$\boldsymbol{h}_a^{t+1} = \boldsymbol{h}_a^t \frac{(\boldsymbol{W}^T\boldsymbol{v})_a}{(\boldsymbol{W}^T\boldsymbol{W}\boldsymbol{h}^t)_a}$$

$$\boldsymbol{H}_{a\mu} \leftarrow \boldsymbol{H}_{a\mu} \frac{(\boldsymbol{W}^T\boldsymbol{v})_{a\mu}}{(\boldsymbol{W}^T\boldsymbol{W}\boldsymbol{H})_{a\mu}}$$

By reversing the role of H and W,*F* can be shown to be non-increasing under the update rules of for W.

Updates for H of divergence

$$F(\boldsymbol{h}) = \sum_i \boldsymbol{v}_i \log\left(\frac{\boldsymbol{v}_i}{\sum_a \boldsymbol{W}_{ia}\boldsymbol{h}_a}\right) - \boldsymbol{v}_i + \sum_a \boldsymbol{W}_{ia}\boldsymbol{h}_a$$

$$G(\boldsymbol{h},\boldsymbol{h}^t) = \sum_i (\boldsymbol{v}_i \log \boldsymbol{v}_i - \boldsymbol{v}_i) + \sum_{ia} \boldsymbol{W}_{ia}\boldsymbol{h}_a - \sum_{ia} \boldsymbol{v}_i \frac{\boldsymbol{W}_{ia}\boldsymbol{h}_a^t}{\sum_b \boldsymbol{W}_{ib}\boldsymbol{h}_b^t}\left(\log \boldsymbol{W}_{ia}\boldsymbol{h}_a - \log \frac{\boldsymbol{W}_{ia}\boldsymbol{h}_a^t}{\sum_b \boldsymbol{W}_{ib}\boldsymbol{h}_b^t}\right)$$

Step one

- Show $G(h, h^t)$ is an auxiliary function for $F(h)$

Step two

- Obtain the update rule by setting the gradient of $G(h, h^t)$ to zero

Step three

- Check the equivalence between the update rule and $\boldsymbol{H}_{a\mu} \leftarrow \boldsymbol{H}_{a\mu} \frac{\sum_i \boldsymbol{W}_{ia}\boldsymbol{V}_{i\mu} / (\boldsymbol{WH})_{i\mu}}{\sum_k \boldsymbol{W}_{ka}}$

Proof of $G(\boldsymbol{h}, \boldsymbol{h}^t)$ by convexity

$$F(\boldsymbol{h}) = \sum_i \boldsymbol{v}_i \log\left(\frac{\boldsymbol{v}_i}{\sum_a \boldsymbol{W}_{ia}\boldsymbol{h}_a}\right) - \boldsymbol{v}_i + \sum_a \boldsymbol{W}_{ia}\boldsymbol{h}_a$$

$$G(\boldsymbol{h}, \boldsymbol{h}^t) = \sum_i (\boldsymbol{v}_i \log \boldsymbol{v}_i - \boldsymbol{v}_i) + \sum_{ia} \boldsymbol{W}_{ia}\boldsymbol{h}_a - \sum_{ia} \boldsymbol{v}_i \frac{\boldsymbol{W}_{ia}\boldsymbol{h}_a^t}{\sum_b \boldsymbol{W}_{ib}\boldsymbol{h}_b^t}\left(\log \boldsymbol{W}_{ia}\boldsymbol{h}_a - \log \frac{\boldsymbol{W}_{ia}\boldsymbol{h}_a^t}{\sum_b \boldsymbol{W}_{ib}\boldsymbol{h}_b^t}\right)$$

- The convexity of log function

$$-\log \sum_a \boldsymbol{W}_{ia}\boldsymbol{h}_a = -\log \sum_a \alpha_a \frac{\boldsymbol{W}_{ia}\boldsymbol{h}_a}{\alpha_a} \leq -\sum_a \alpha_a \log \frac{\boldsymbol{W}_{ia}\boldsymbol{h}_a}{\alpha_a}$$

$$\alpha_a = \frac{\boldsymbol{W}_{ia}\boldsymbol{h}_a^t}{\sum_b \boldsymbol{W}_{ib}\boldsymbol{h}_b^t}$$

$$-\log \sum_a \boldsymbol{W}_{ia}\boldsymbol{h}_a \leq -\sum_a \frac{\boldsymbol{W}_{ia}\boldsymbol{h}_a^t}{\sum_b \boldsymbol{W}_{ib}\boldsymbol{h}_b^t}\left(\log \boldsymbol{W}_{ia}\boldsymbol{h}_a - \log \frac{\boldsymbol{W}_{ia}\boldsymbol{h}_a^t}{\sum_b \boldsymbol{W}_{ib}\boldsymbol{h}_b^t}\right) \Rightarrow G(\boldsymbol{h}, \boldsymbol{h}^t) \geq F(\boldsymbol{h})$$

Minimum of $G(\boldsymbol{h}, \boldsymbol{h}^t)$ and update rules

$$\frac{\partial G(\boldsymbol{h}, \boldsymbol{h}^t)}{\partial \boldsymbol{h}_a} = \sum_i \boldsymbol{W}_{ia} - \frac{1}{\boldsymbol{h}_a} \sum_i \boldsymbol{v}_i \frac{\boldsymbol{W}_{ia} \boldsymbol{h}_a^t}{\sum_b \boldsymbol{W}_{ib} \boldsymbol{h}_b^t} = 0$$

$$\boldsymbol{h}^{t+1} = \frac{\boldsymbol{h}_a^t}{\sum_i \boldsymbol{W}_{ia}} \sum_i \frac{\boldsymbol{v}_i}{\sum_b \boldsymbol{W}_{ib} \boldsymbol{h}_b^t} \boldsymbol{W}_{ia}$$

$$\boldsymbol{H}_{a\mu} \leftarrow \boldsymbol{H}_{a\mu} \frac{\sum_i \boldsymbol{W}_{ia} \boldsymbol{V}_{i\mu} / (\boldsymbol{W}\boldsymbol{H})_{i\mu}}{\sum_k W_{ka}}$$

- By reversing the role of H and W, the update rule for W can similarly be shown to be non-increasing.