Advanced Machine Learning

Lecture 5: Non-Negative Matrix Factorization

Nora Ouzir : nora.ouzir@centralesupelec.fr
Lucca Guardiola : lucca.guardiola@centralesupelec.fr

Oct. - Nov. 2020

CentraleSupélec

Content

Non-negative Matrix Factorization

- Dimension reduction technique
- Extensions to deep-NMFs
- Find a better representation of data to be used for regression, classification, ...

What are the underlying models, how to learn these models, algorithms?

Today's Lecture

1. Introduction
   1. Main principles of NMF
   2. Key applications

2. Common Losses for NMF

3. Multiplicative Algorithms
   1. Quadratic and KL distances
   2. Weighted NMFs
   3. Regularized NMFs

4. Other NMF Algorithms

Today's course

Today's course

1. Introduction
    1. Main principles of NMF
    2. Key applications
2. Common Losses for NMF
3. Multiplicative Algorithms
    1. Quadratic and KL distances
    2. Weighted NMFs
    3. Regularized NMFs
4. Other NMF Algorithms

Matrix Factorization: Principles

r ℓ_2 norm

Given a set of data entries $\mathbf{x}_j \in \mathbb{R}^p, 1 \leq j \leq n$ and a target dimension $r < \min(p; n)$, we search for r basis elements w_k, $1 \leq k \leq r$ and weights $h_j \in \mathbb{R}^r$ such that

$$\mathbf{x}_j \approx \sum_{k=1}^{r} w_k h_j(k)$$

$r = 3$

$X_{:j} = w_1 h_j(1) + w_2 h_j(2) + w_3 h_j(3)$

Equivalently

$$X \approx WH$$

$X^{p \times n} = W^{p \times r} \times H^{r \times n}$

$p \times n$

Notations

- $X \in \mathbb{R}^{p \times n}$ s.t $X(:,j) = x_j$ for $1 \leq j \leq n$,
- $W \in \mathbb{R}^{p \times r}$ s.t $W(:,k) = w_k$ for $1 \leq k \leq r$,
- $H \in \mathbb{R}^{r \times n}$ s.t $H(:,j) = h_j$ for $1 \leq j \leq n$,

Matrix Factorization: Key Aspects

Goal: Low rank approximation/ dimension reduction

$$\boxed{X \approx WH}$$

$(X - WH)$

$p \times r$ $r \times n$

Key Aspects

1. How to evaluate the quality of the approximation ?
   - Examples: Frobenius norm, KL-divergence, logistic, Itakura-Saito.

   $\rightarrow$ Loss function

Matrix Factorization: Key Aspects

Goal: Low rank approximation/ dimension reduction

$$X \approx WH$$

Key Aspects

1. How to evaluate the quality of the approximation ?
   - Examples: Frobenius norm, KL-divergence, logistic, Itakura-Saito.

   $\rightarrow$ Loss function
2. Assumptions on the structure of W and H ?
   - Independence, sparsity, normalization, ...

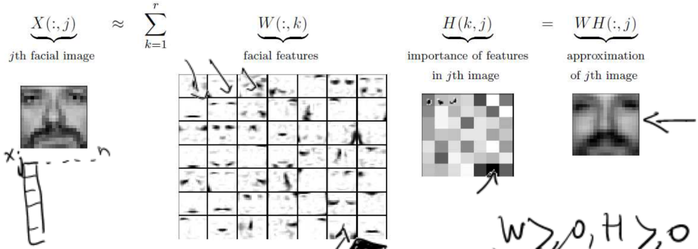
   $\rightarrow$ Non-negativity

NMF: find (W, H) s.t $X \approx WH$, $W \geq 0$, $H \geq 0$

Today's course

1. Introduction
    1. Main principles of NMF
    2. Key applications
2. Common Losses for NMF
3. Multiplicative Algorithms
    1. Quadratic and KL distances
    2. Weighted NMFs
    3. Regularized NMFs
4. Other NMF Algorithms

Example: Facial Feature Extraction



Decomposition of the CBCL face database [Lee and Seung, 1999]

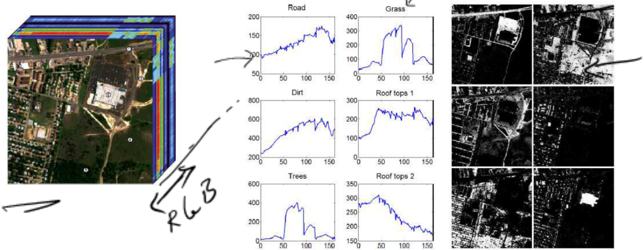
- Some of the features look like parts of facial features/organs: noses, eyes, etc.
- Decomposition of a face as having a certain weight of a some nose type, a certain amount of some eye type, etc.

→ Interpretable representation

Example: Spectral Unmixing

$r < p$

$$\underbrace{X(:,j)}_{\substack{\text{spectral signature}\\ \text{of } j\text{th pixel}}} \approx \sum_{k=1}^{r} \underbrace{W(:,k)}_{\substack{\text{spectral signature}\\ \text{of } k\text{th endmember}}} \underbrace{H(k,j)}_{\substack{\text{abundance of } k\text{th endmember}\\ \text{in } j\text{th pixel}}} .$$



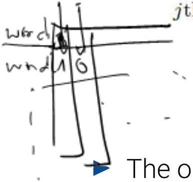
Decomposition of the Urban hyperspectral image [Ma *et al.*, 2014]

- NMF is able to compute the spectral signatures of the endmembers and simultaneously the abundance of each endmember in each pixel.

→ Here non-negativity is directly relevant for the application

Example: Topic Modelling in Text Mining

Goal: Decompose a term-document matrix, where each column represents a document, and each element in the document represents the weight of a certain word (e.g., term frequency - inverse document frequency).



$$\underbrace{X(:,j)}_{j\text{th document}} \approx \sum_{k=1}^{r} \underbrace{W(:,k)}_{k\text{th topic}} \underbrace{H(k,j)}_{\substack{\text{importance of } k\text{th topic} \\ \text{in } j\text{th document}}}$$

Topic decomposition model [Blei, 2012]

- The ordering of the words in the documents is not taken into account (=bag-of-words).
- The NMF decomposition of the term-document matrix yields components that could be considered as "topics"
- Decomposes each document into a weighted sum of topics.

Today's course

White board

$$X \approx WH$$

- $(X, WH) \to \mathcal{D}(X, WH) = \sum_{i=1}^{n} \sum_{j=1}^{p} d(x_{ij}, [WH]_{ij})$
  - → Separable
  - → $\begin{cases} d(x,x) = 0 \\ d(x,y) \geq 0 \end{cases}$

Example $d(x,y) = (x-y)^2$ QUADRATIC

$$\underbrace{\begin{matrix} W^T W = Id \\ H^T H = diag(\lambda) \end{matrix}}_{\text{SVD}} \qquad \underbrace{\begin{matrix} H^T H = Id \\ H_{kj} \in \{0,1\} \end{matrix}}_{\text{k-MEANS}} \qquad \underbrace{\begin{matrix} W \geq 0 \\ H \geq 0 \end{matrix}}_{\text{LS-NMF}}$$

- Gaussian → non-realistic
  → robust

White board

QUADRATIC.

KL-divergence

$d(x,y) = x\log\frac{x}{y} + x - y$

$/ W \geq 0, H \geq 0$ KL NMF.

Itakura-Saito

$d(x,y) = \frac{x}{y} - \log\frac{x}{y} - 1$

Music Analysis

$d(\lambda x, \lambda y) = d(x,y)$

Need more advanced solvers

White board

Today's course

1. Introduction
    1. Main principles of NMF
    2. Key applications
2. Common Losses for NMF
3. Multiplicative Algorithms
    1. Quadratic and KL distances
    2. Weighted NMFs
    3. Regularized NMFs
4. Other NMF Algorithms

NMF Algorithms

Recall: we want to find (W, H) s.t $X \approx WH$, $W \cancel{\leq} \geq 0$, $H \cancel{\leq} \geq 0$

- Loss function + constraints (+ regularization)

$\rightarrow$ Non-convex optimization problem wrt W and H

Main types of algorithms

- Multiplicative methods
- Alternating LS
- Gradient descent

Today's course

Multiplicative Algorithms for NMF

Solve

$$\min_{W,H} \mathcal{D}(X, WH) \text{ s.t } W \geq 0, \; H \geq 0$$

Challenges

NMF is NP-hard and ill-posed (solution is not unique):

- Most algorithms are only guaranteed to converge to a stationary point
- Sensitive to initialization
- In practice: priors on W and H/regularization

[Lee and Seung, 1999]

→ Popular class of methods relying on multiplicative updates

→ Key assumption $X \geq 0$

Multiplicative Algorithms: [Lee and Seung]

Assuming $X \geq 0$, the resulting updates are multiplicative and take the following forms

- Frobenius norm $\|X - WH\|_F^2$

elementwise.

$$W \leftarrow W \circ \frac{XH^T}{WHH^T}$$

$$H \leftarrow H \circ \frac{W^T X}{W^T WH}$$

- KL-divergence: $\mathcal{KL}(X, WH)$

$$W_{ik} \leftarrow W_{ik} \frac{\sum_{l=1}^{n} (H_{kl} X_{il} / [WH]_{il})}{\sum_{l=1}^{n} H_{kl}}$$

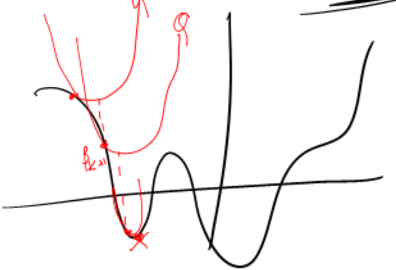
$$H_{kj} \leftarrow H_{kj} \frac{\sum_{i=1}^{p} (W_{ik} X_{ij} / [WH]_{ij})}{\sum_{i=1}^{p} W_{ik}}$$

White Board

$$\underbrace{F(\beta)}\leq\underbrace{Q(\beta,\beta')}$$

$(F(\beta') = Q(\beta',\beta'))$

$(\beta)_k$



LEE & SEUNG : using Quadratic Loss.

White Board

FROBENIUS: $\|WH - X\|_F^2 = \sum_{j=1} \|Wh_j - x_j\|^2$

① Build Q

② Separable w.r.t lines (i) x_{ij}

$\beta' = \bar{H}$ $\quad \|Wh_j - x_j\|^2 \leq \sum_{i=1}^{p} \frac{1}{[W\bar{h}_j]_i} \sum_{k=1}^{r} w_{ik} \bar{H}_{kj} \left(x_{ij} - \frac{H_{kj}}{\bar{H}_{kj}} [W\bar{h}_j]_i \right)^2$

JENSEN INEQUALITY

$f\left(\sum \lambda_i x\right) \leq \sum \lambda f(x)$

Can show that $=$ when $H = \bar{H}$

$$F \leq \underbrace{\sum_{j=1}^{n} \sum_{i=1}^{p} \sum_{k=1}^{r} \frac{w_{ik} \bar{H}_{kj}}{[W\bar{h}_j]_i} \left(x_{ij} - \frac{H_{kj}}{\bar{H}_{kj}} [W\bar{h}_j]_i \right)^2}_{Q(H, \bar{H})}$$

White Board

$$\underbrace{\frac{\partial Q}{\partial H_{kj}}} = 2\sum_i^p \frac{w_{ik}\cancel{\bar{H}_{kj}}}{\cancel{[W\bar{h}_j]_i}} \times \sum \frac{\cancel{[W\bar{h}_j]_i}}{\cancel{\bar{H}_{kj}}} \times \left(x_{ij} - \frac{H_{kj}}{\bar{H}_{kj}}(W\bar{h}_j)_i\right)$$

$$\frac{\partial Q}{\partial H_{kj}} = 0 \Rightarrow \sum_{i=1}^p w_{ik}\left(x_{ij} - \frac{H_{kj}}{\bar{H}_{kj}}[W\bar{h}_j]_i\right) = 0$$

$$\Longrightarrow \underbrace{\sum^p w_{ik} X_{ij}}_{[W^T X]_{kj}} = \frac{H_{kj}}{\bar{H}_{kj}} \underbrace{\sum^p w_{ik}[W\bar{h}_j]_i}_{[W^T W \bar{H}]_{kj}}$$

$X \geq 0$

$$H_{kj} = \bar{H}_{kj} \times \frac{[W^T X]_{kj}}{[W^T W \bar{H}]_{kj}}$$

$\rightarrow \underline{Q(H, \bar{H})}$

Jensen ineq.

$\rightarrow \underline{\frac{\partial Q}{\partial H_{kj}}} = 0 \rightarrow$ MU

Multiplicative Algorithms: Surrogate functions

The multiplicative schemes rely on separable surrogate functions that majorize the loss w.r.t W and H. For every $(X, W, H, \bar{H}) \geq \mathbf{0}$, and $1 \leq j \leq n$:

- Frobenius norm:

$$\|Wh_j - x_j\|_2^2 \leq \sum_{i=1}^{p} \frac{1}{[W\bar{h}_j]_i} \sum_{k=1}^{r} W_{ik}\bar{H}_{kj}\left(X_{ij} - \frac{H_{kj}}{\bar{H}_{kj}}[W\bar{h}_j]_i\right)^2$$

- KL-divergence:

$$\mathcal{KL}(x_j, Wh_j) \leq \sum_{i=1}^{p}\Big[X_{ij}\log X_{ij} - X_{ij} + [Wh_j]_i - \frac{X_{ij}}{[W\bar{h}_j]_i}\sum_{k=1}^{r} W_{ik}\bar{H}_{kj}\log\Big(\frac{H_{kj}}{\bar{H}_{kj}}[W\bar{h}_j]_i\Big)\Big]$$

Multiplicative Algorithms: Implementation Key points

MU algorithms

- Build a quadratic majorant function
- Jensen inequality $\rightarrow$ separability
- Guarantee positive updates by construction

Implementation: initialization

- Apply factorization without non-negativity constraint (*e.g.*, SVD) $\rightarrow$ truncate/threshold
- Avoid too simple initializations (*e.g.*, identity, random)
- Clustering methods, r columns of X ...

Today's course

Weighted NMF

Application example: Matrix completion to predict unobserved data (*e.g.*, user-rating) $\rightarrow$ Binary weights indicate the position of the available entries in X.

- Weighted Frobenius norm $\|\Sigma \circ (X - WH)\|_F^2$

$$W \leftarrow W \circ \frac{(\Sigma \circ X)H^T}{(\Sigma \circ WH)H^T}$$

$$H \leftarrow H \circ \frac{W^T(\Sigma \circ X)}{W^T(\Sigma \circ WH)}$$

- Weighted KL-divergence: $\mathcal{KL}(X, \mathrm{Diag}(p)WH\mathrm{Diag}(q))$

$p > 0$
$q > 0$

$$W_{ik} \leftarrow W_{ik} \frac{\sum_{l=1}^{n}(H_{kl}X_{il}/p_i[WH]_{il})}{\sum_{l=1}^{n} q_l H_{kl}}$$

$$H_{kj} \leftarrow H_{kj} \frac{\sum_{i=1}^{p}(W_{ik}X_{ij}/q_j[WH]_{ij})}{\sum_{i=1}^{p} p_i W_{ik}}$$

White Board

Today's course

1. Introduction
   1. Main principles of NMF
   2. Key applications

2. Common Losses for NMF

3. Multiplicative Algorithms
   1. Quadratic and KL distances
   2. Weighted NMFs
   3. Regularized NMFs

4. Other NMF Algorithms

Regularized NMFs

Regularized Frobenius norm: (sparsity, stability, scale ambiguity)

$$\frac{1}{2}\|X - WH\|_F^2 + \frac{\mu}{2}\|H\|_F^2 + \lambda\|H\|_1 + \frac{\nu}{2}\|W\|_F^2$$

Resulting MU

$\sum h$

$\ell_1 \to$ sparsity

$$W \leftarrow W \circ \frac{XH^T}{W(HH^T + \nu I_r)}$$

$$H \leftarrow H \circ \frac{W^T X - \lambda 1_{r\times n}}{(W^T W + \mu I_r)H}$$

→ The regularization terms are already separable!

→ The ambiguity due to the rescaling and rotation of (W, H) is frozen by the penalty terms.

White Board

Multiplicative NMF Algorithms: Summary

To sum up

- Dimension reduction for non-negative data
- Easily interpretable representation
- Weighted and regularized formulations

In practice MU is...

+ Simple to implement
- Slow convergence
- Sensitive to initialization (stuck with small values)
- Choice of r ? Heuristics exist, domain knowledge,...

Today's course

1. Introduction
   1. Main principles of NMF
   2. Key applications

2. Common Losses for NMF

3. Multiplicative Algorithms
   1. Quadratic and KL distances
   2. Weighted NMFs
   3. Regularized NMFs

4. Other NMF Algorithms

Other NMF Algorithms: Alternating LS

The subproblem in one variable is convex!

Alternating LS: Unconstrained solution w.r.t. W or H followed by a projection onto non-negative orthant.

- Easy to implement but oscillations can arise (no convergence guarantees) $\rightarrow$ initialization purposes.

Other NMF Algorithms: Alternating LS

The subproblem in one variable is convex!

Alternating LS: Unconstrained solution w.r.t. W or H followed by a projection onto non-negative orthant.

- Easy to implement but oscillations can arise (no convergence guarantees) $\rightarrow$ initialization purposes.

Alternating Non-negative LS: Solve the constrained problem (alternate w.r.t. W,H) exactly using an inner solver (e.g., projected gradient, Quasi-Newton, active set).

- Expensive $\rightarrow$ useful as a refinement step for a cheaper MU.

Other NMF Algorithms: Alternating LS

The subproblem in one variable is convex!

Alternating LS: Unconstrained solution w.r.t. W or H followed by a projection onto non-negative orthant.

- Easy to implement but oscillations can arise (no convergence guarantees) $\rightarrow$ initialization purposes.

Alternating Non-negative LS: Solve the constrained problem (alternate w.r.t. W,H) exactly using an inner solver (e.g., projected gradient, Quasi-Newton, active set).

- Expensive $\rightarrow$ useful as a refinement step for a cheaper MU.

Hierarchical Alternative LS: Exact coordinate descent method, updating one column of W (resp. one line of H) at a time.

- Simple to implement, performance is similar to MU.

Other NMF Algorithms: Alternate Scheme

Algorithm CD Two-Block Coordinate Descent – Framework of Most NMF Algorithms

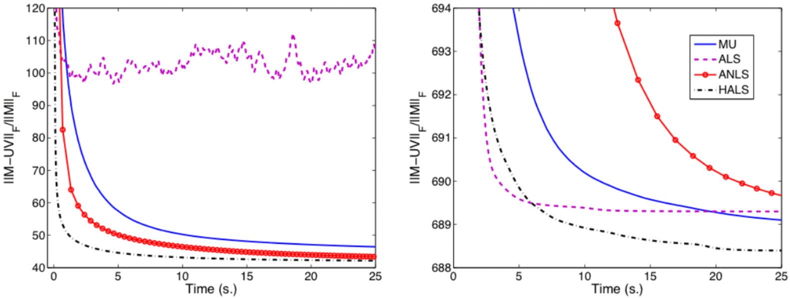
Input: Input nonnegative matrix $X \in \mathbb{R}_+^{p \times n}$ and factorization rank r.
Output: $(W, H) \geq 0$: A rank-r NMF of $X \approx WH$.

1: Generate some initial matrices $W^{(0)} \geq 0$ and $H^{(0)} \geq 0$; see Section 3.1.8.
2: **for** $t = 1, 2, \ldots^{\dagger}$ **do**
3: $\quad W^{(t)} = \text{update}\big(X, H^{(t-1)}, W^{(t-1)}\big)$.
4: $\quad {H^{(t)}}^T = \text{update}\Big(X^T, {W^{(t)}}^T, {H^{(t-1)}}^T\Big)$.
5: **end for**

†See Section 3.1.7 for stopping criteria.

Algorithm: Coordinate descent framework

Other NMF Algorithms: Comparison



Performance comparison to MU using same initialization:
Average error vs execution times