Lecture 6: Nonnegative Matrix Factorization

School of Mathematical Sciences, Xiamen University

1. Setting

- For data sets where all attributes are nonnegative numbers, it is very convenient to approximate the data by a linear combination of a few nonnegative feature vectors with nonnegative coefficients.
- Lee D.D. and Seung H.S. (1999) Learning the parts of objects by nonnegative matrix factorization, Nature 401:788–791
- Denote by $\mathbb{R}_+$ the set of nonnegative real numbers. Given a nonnegative data matrix $\mathbf{X} \in \mathbb{R}_+^{n \times p}$, the nonnegative matrix factorization (NMF) problem can be formulated as the search for an approximation

  $$\mathbf{X} \approx \mathbf{W}\mathbf{H},$$

  where $\mathbf{W} \in \mathbb{R}_+^{n \times k}$ and $\mathbf{H} \in \mathbb{R}_+^{k \times p}$ are two nonnegative matrices of rank $k \leq \min\{n, p\}$.
- Obviously, the solution is not unique. Example: Let $\mathbf{L} \in \mathbb{R}_+^{k \times k}$ be diagonal and invertible. Consider $\mathbf{W}\mathbf{L}\mathbf{L}^{-1}\mathbf{H}$, which is a new NMF.

- Let $\mathbf{X} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\top$ be SVD of $\mathbf{X}$. Then $\mathbf{U}_k\boldsymbol{\Sigma}_k\mathbf{V}_k^\top$ is the best rank k (assume $k < \operatorname{rank}(\mathbf{X})$) approximation of $\mathbf{X}$ in the Frobenius norm (rank k PCA), that is,

$$\|\mathbf{X} - \mathbf{U}_k\boldsymbol{\Sigma}_k\mathbf{V}_k^\top\|_{\mathrm{F}} \leq \|\mathbf{X} - \mathbf{W}\mathbf{H}\|_{\mathrm{F}}$$

  for every $\mathbf{W} \in \mathbb{R}^{n\times k}$, $\mathbf{H} \in \mathbb{R}^{k\times p}$.

- The nonnegative matrix factorization becomes the method of choice when the nonnegativity of the feature vectors is more important than the accuracy of the approximation.

2. The alternating nonnegative least squares algorithm

- NMF problem 1:

  Given a matrix $\mathbf{X} \in \mathbb{R}_+^{n\times p}$, find matrices $\mathbf{W} \in \mathbb{R}_+^{n\times k}$ and $\mathbf{H} \in \mathbb{R}_+^{k\times p}$ that minimize the cost function

$$f(\mathbf{W}, \mathbf{H}) = \frac{1}{2}\|\mathbf{X} - \mathbf{W}\mathbf{H}\|_{\mathrm{F}}^2.$$

Alternating nonnegative least squares (ANLS) algorithm

1. **Given**. $\mathbf{X} \in \mathbb{R}_+^{n\times p}$, $k < \min\{n, p\}$, $\tau > 0$, and `maxit`
2. **Initialize**. Generate $\mathbf{W}^0 \in \mathbb{R}_+^{n\times k}$ and scale its columns to have unit ∞-norm. Set $t = 0$ and $\delta = \infty$.
3. **Iteration**. While $\delta > \tau$ and $t <$ `maxit`, update

$$\mathbf{H}^{t+1} = \operatorname{argmin} \|\mathbf{X} - \mathbf{W}^t\mathbf{H}\|_{\mathrm{F}} \text{ s.t., } \mathbf{H} \in \mathbb{R}_+^{k\times p}$$

$$\mathbf{W}^{t+1} = \operatorname{argmin} \|\mathbf{X} - \mathbf{W}\mathbf{H}^{t+1}\|_{\mathrm{F}} \text{ s.t., } \mathbf{W} \in \mathbb{R}_+^{n\times k}$$

Scale the columns of $\mathbf{W}^{t+1}$: Define

$$\lambda_j = \max_{1\le i\le n} \mathbf{W}_{ij}^{t+1} \text{ and } \mathbf{L} = \operatorname{diag}\{\lambda_1, \cdots, \lambda_k\}$$

and set $\mathbf{W}^{t+1} = \mathbf{W}^{t+1}\mathbf{L}^{-1}$. If $t > 0$, compute

$$\delta = \frac{\|\mathbf{W}^{t+1} - \mathbf{W}^t\|_{\mathrm{F}}}{\|\mathbf{W}^t\|_{\mathrm{F}}} + \frac{\|\mathbf{H}^{t+1} - \mathbf{H}^t\|_{\mathrm{F}}}{\|\mathbf{H}^t\|_{\mathrm{F}}}.$$

Set $t = t + 1$.

- Express the Frobenius norm from a columnwise perspective

$$\|\mathbf{X} - \mathbf{W}\mathbf{H}\|_{\mathrm{F}}^2 = \sum\nolimits_{j=1}^{p} \|\mathbf{X}_{:,j} - \mathbf{W}\mathbf{H}_{:,j}\|_2^2$$

  Updating $\mathbf{H}$ need to solve p constrained least squares propblmes:

$$\text{minimize } \|\mathbf{W}^t\mathbf{H}_{:,j} - \mathbf{X}_{:,j}\|_2 \quad \text{s.t.} \quad \mathbf{H}_{:,j} \in \mathbb{R}_+^k$$

- Express the Frobenius norm from a rowwise perspective

$$\|\mathbf{X} - \mathbf{W}\mathbf{H}\|_{\mathrm{F}}^2 = \sum\nolimits_{i=1}^{n} \|\mathbf{X}_{i,:} - \mathbf{W}_{i,:}\mathbf{H}\|_2^2$$

  Updating $\mathbf{W}$ need to solve n constrained least squares propblmes:

$$\text{minimize } \|(\mathbf{H}^{t+1})^\top(\mathbf{W}_{i,:})^\top - (\mathbf{X}_{i,:})^\top\|_2 \quad \text{s.t.} \quad (\mathbf{W}_{i,:})^\top \in \mathbb{R}_+^k$$

- MATLAB: `lsqnonneg`

3. Multiplicative updating formula

- To guarantee the nonnegativity, we use a change of variables:

$$\mathbf{W}_{ij} = \mathrm{e}^{\xi_{ij}} \quad \text{and} \quad \mathbf{H}_{ij} = \mathrm{e}^{\zeta_{ij}}.$$

We have

$$\begin{aligned}\frac{\partial f(\mathbf{W},\mathbf{H})}{\partial \xi_{\mu\nu}} &= \frac{1}{2}\frac{\partial}{\partial \xi_{\mu\nu}}\left(\sum_{i,j}(\mathbf{X}_{ij} - (\mathbf{WH})_{ij})^2\right) \\ &= -\sum_{i,j}(\mathbf{X}_{ij} - (\mathbf{WH})_{ij})\frac{\partial (\mathbf{WH})_{ij}}{\partial \xi_{\mu\nu}},\end{aligned}$$

where

$$\frac{\partial (\mathbf{WH})_{ij}}{\partial \xi_{\mu\nu}} = \frac{\partial}{\partial \xi_{\mu\nu}}\left(\sum_{\ell}\mathbf{W}_{i\ell}\mathbf{H}_{\ell j}\right) = \sum_{\ell}\frac{\partial \mathbf{W}_{i\ell}}{\partial \xi_{\mu\nu}}\mathbf{H}_{\ell j}.$$

Observe that

$$\frac{\partial \mathbf{W}_{i\ell}}{\partial \xi_{\mu\nu}} = 0 \quad \text{if} \quad i \neq \mu \quad \text{or} \quad \ell \neq \nu,$$

and for $i = \mu$ and $\ell = \nu$,

$$\frac{\partial \mathbf{W}_{\mu\nu}}{\partial \xi_{\mu\nu}} = \mathrm{e}^{\xi_{\mu\nu}} = \mathbf{W}_{\mu\nu}.$$

Using the Kronecker symbols δ_{ij} yields

$$\begin{aligned} \frac{\partial (\mathbf{W}\mathbf{H})_{ij}}{\partial \xi_{\mu\nu}} &= \sum_{\ell} \frac{\partial \mathbf{W}_{i\ell}}{\partial \xi_{\mu\nu}} \mathbf{H}_{\ell j} = \sum_{\ell} \mathbf{W}_{i\ell} \delta_{\mu i} \delta_{\nu\ell} \mathbf{H}_{\ell j} \\ &= \delta_{\mu i} \sum_{\ell} \mathbf{W}_{i\ell} \delta_{\nu\ell} \mathbf{H}_{\ell j} \\ &= \delta_{\mu i} \mathbf{W}_{i\nu} \mathbf{H}_{\nu j}. \end{aligned}$$

Then we have

$$
\begin{aligned}
\frac{\partial f(\mathbf{W},\mathbf{H})}{\partial \xi_{\mu\nu}} &= -\sum_{i,j}(\mathbf{X}_{ij} - (\mathbf{W}\mathbf{H})_{ij})\delta_{\mu i}\mathbf{W}_{i\nu}\mathbf{H}_{\nu j} \\
&= -\sum_{j}(\mathbf{X}_{\mu j} - (\mathbf{W}\mathbf{H})_{\mu j})\mathbf{W}_{\mu\nu}\mathbf{H}_{\nu j} \\
&= -\mathbf{W}_{\mu\nu}(\mathbf{X}\mathbf{H}^\top)_{\mu\nu} + \mathbf{W}_{\mu\nu}(\mathbf{W}\mathbf{H}\mathbf{H}^\top)_{\mu\nu}.
\end{aligned}
$$

To find a critical point, we set

$$
\frac{\partial f(\mathbf{W},\mathbf{H})}{\partial \xi_{\mu\nu}} = -\mathbf{W}_{\mu\nu}(\mathbf{X}\mathbf{H}^\top)_{\mu\nu} + \mathbf{W}_{\mu\nu}(\mathbf{W}\mathbf{H}\mathbf{H}^\top)_{\mu\nu} = 0.
$$

Similarly, we have

$$
\frac{\partial f(\mathbf{W},\mathbf{H})}{\partial \zeta_{\mu\nu}} = -\mathbf{H}_{\mu\nu}(\mathbf{W}^\top\mathbf{X})_{\mu\nu} + \mathbf{H}_{\mu\nu}(\mathbf{W}^\top\mathbf{W}\mathbf{H})_{\mu\nu} = 0.
$$

- Let $\mathbf{W}^c$ and $\mathbf{H}^c$ denote the current values of $\mathbf{W}$ and $\mathbf{H}$, and write the current approximation of the data matrix as

$$\mathbf{X}^c = \mathbf{W}^c \mathbf{H}^c.$$

  Let $\mathbf{W}^+$ and $\mathbf{H}^+$ denote the updated matrices. Set

$$-\mathbf{W}^c_{\mu\nu}(\mathbf{X}(\mathbf{H}^c)^\top)_{\mu\nu} + \mathbf{W}^+_{\mu\nu}(\mathbf{X}^c(\mathbf{H}^c)^\top)_{\mu\nu} = 0$$

  and solve for the next iterate, yielding

$$\mathbf{W}^+_{\mu\nu} = \frac{(\mathbf{X}(\mathbf{H}^c)^\top)_{\mu\nu}}{(\mathbf{X}^c(\mathbf{H}^c)^\top)_{\mu\nu}} \mathbf{W}^c_{\mu\nu}.$$

  Similarly, we have

$$\mathbf{H}^+_{\mu\nu} = \frac{((\mathbf{W}^c)^\top \mathbf{X})_{\mu\nu}}{((\mathbf{W}^c)^\top \mathbf{X}^c)_{\mu\nu}} \mathbf{H}^c_{\mu\nu}.$$

NMF multiplicative updating algorithm I

1. **Given**. $\mathbf{X} \in \mathbb{R}_{+}^{n\times p}$, $k < \min\{n, p\}$, $\tau > 0$, and `maxit`
2. **Initialize**. Generate $\mathbf{W}^0 \in \mathbb{R}_{++}^{n\times k}$ and scale its columns to have unit ∞-norm. Generate $\mathbf{H}^0 \in \mathbb{R}_{++}^{k\times p}$. Set $t = 0$ and $\delta = \infty$.
3. **Iteration**. While $\delta > \tau$ and $t <$ `maxit`,

   Compute $\mathbf{X}^c = \mathbf{W}^t\mathbf{H}^t$ and update $\mathbf{H}$,

   $$\mathbf{H}^{t+1} = ((\mathbf{W}^t)^\top \mathbf{X})./((\mathbf{W}^t)^\top \mathbf{X}^c).*\mathbf{H}^t.$$

   Recompute $\mathbf{X}^c = \mathbf{W}^t\mathbf{H}^{t+1}$ and update $\mathbf{W}$,

   $$\mathbf{W}^{t+1} = (\mathbf{X}(\mathbf{H}^{t+1})^\top)./(\mathbf{X}^c(\mathbf{H}^{t+1})^\top).*\mathbf{W}^t.$$

   Scale the columns of $\mathbf{W}^{t+1}$: Define

   $$\lambda_j = \max_{1\le i\le n} \mathbf{W}_{ij}^{t+1} \text{ and } \mathbf{L} = \mathrm{diag}\{\lambda_1, \cdots, \lambda_k\}$$

   and set $\mathbf{W}^{t+1} = \mathbf{W}^{t+1}\mathbf{L}^{-1}$. Compute

   $$\delta = \frac{\|\mathbf{W}^{t+1} - \mathbf{W}^t\|_{\mathrm{F}}}{\|\mathbf{W}^t\|_{\mathrm{F}}} + \frac{\|\mathbf{H}^{t+1} - \mathbf{H}^t\|_{\mathrm{F}}}{\|\mathbf{H}^t\|_{\mathrm{F}}}.$$

   Set $t = t + 1$.

4. Alternative cost functions

- It is natural to quantify how close two nonnegative matrices are by resorting to tools developed in the context of information theory, statistical physics, and probability theory, for example, *entropy divergence*, defined as

$$D(\mathbf{A}||\mathbf{B}) = \sum_{i=1}^{n} \sum_{j=1}^{p} \left(\mathbf{A}_{ij} \log \frac{\mathbf{A}_{ij}}{\mathbf{B}_{ij}} - \mathbf{A}_{ij} + \mathbf{B}_{ij} \right).$$

- Theorem: For any two matrices $\mathbf{A}, \mathbf{B} \in \mathbb{R}_{++}^{n \times p}$, the entropy divergence $D(\mathbf{A}||\mathbf{B})$ is nonnegative, and is equal to zero if and only if $\mathbf{A} = \mathbf{B}$.
- The entropy divergence $D(\mathbf{A}||\mathbf{B})$ is a dissimilarity measure rather than a proper distance, since in general,

$$D(\mathbf{A}||\mathbf{B}) \neq D(\mathbf{B}||\mathbf{A}).$$

- NMF problem 2:

  Given a matrix $\mathbf{X} \in \mathbb{R}_{+}^{n\times p}$, find matrices $\mathbf{W} \in \mathbb{R}_{++}^{n\times k}$ and $\mathbf{H} \in \mathbb{R}_{++}^{k\times p}$ that minimize the cost function $g(\mathbf{W},\mathbf{H}) = D(\mathbf{X}||\mathbf{WH})$.

- We have

$$\begin{aligned}
\frac{\partial g(\mathbf{W},\mathbf{H})}{\partial \xi_{\mu\nu}} &= \sum_{i,j} \Bigg(\mathbf{X}_{ij} \frac{\partial}{\partial \xi_{\mu\nu}} (\log \mathbf{X}_{ij} - \log(\mathbf{WH})_{ij}) \\
&\qquad - \frac{\partial}{\partial \xi_{\mu\nu}} (\mathbf{X}_{ij} - (\mathbf{WH})_{ij}) \Bigg) \\
&= \sum_{i,j} \left(-\frac{\mathbf{X}_{ij}}{(\mathbf{WH})_{ij}} + 1 \right) \frac{\partial (\mathbf{WH})_{ij}}{\partial \xi_{\mu\nu}} \\
&= \sum_{i,j} \left(-\frac{\mathbf{X}_{ij}}{(\mathbf{WH})_{ij}} + 1 \right) \mathbf{W}_{i\nu} \mathbf{H}_{\nu j} \delta_{\mu i} \\
&= \sum_{j} \left(-\frac{\mathbf{X}_{\mu j}}{(\mathbf{WH})_{\mu j}} + 1 \right) \mathbf{W}_{\mu\nu} \mathbf{H}_{\nu j}.
\end{aligned}$$

Similarly, we have

$$\frac{\partial g(\mathbf{W},\mathbf{H})}{\partial \zeta_{\mu\nu}} = \sum_i \left(-\frac{\mathbf{X}_{i\nu}}{(\mathbf{WH})_{i\nu}} + 1 \right) \mathbf{W}_{i\mu}\mathbf{H}_{\mu\nu}.$$

- Set the updating formulas

$$\mathbf{W}^+_{\mu\nu} = \left(\frac{1}{\sum_j \mathbf{H}^c_{\nu j}} \sum_j \frac{\mathbf{X}_{\mu j}}{(\mathbf{W}^c\mathbf{H}^c)_{\mu j}} \mathbf{H}^c_{\nu j} \right) \mathbf{W}^c_{\mu\nu},$$

$$\mathbf{H}^+_{\mu\nu} = \left(\frac{1}{\sum_i \mathbf{W}^c_{i\mu}} \sum_i \frac{\mathbf{X}_{i\nu}}{(\mathbf{W}^c\mathbf{H}^c)_{i\nu}} \mathbf{W}^c_{i\mu} \right) \mathbf{H}^c_{\mu\nu}.$$

- The columns of **W** can be scaled to have a unit 1-norm.

NMF multiplicative updating algorithm II

1. **Given**. $\mathbf{X} \in \mathbb{R}_{+}^{n\times p}$, $k < \min\{n, p\}$, $\tau > 0$, and `maxit`
2. **Initialize**. Generate $\mathbf{W}^0 \in \mathbb{R}_{++}^{n\times k}$ and scale its columns to have unit 1-norm. Generate $\mathbf{H}^0 \in \mathbb{R}_{++}^{k\times p}$. Set $t = 0$ and $\delta = \infty$.
3. **Iteration**. While $\delta > \tau$ and $t <$ `maxit`,

   Compute $\mathbf{X}^c = \mathbf{W}^t\mathbf{H}^t$ and update $\mathbf{H}$,

   $$\mathbf{H}_{\mu\nu}^{t+1} = (\textstyle\sum_i (\mathbf{X}_{i\nu}/\mathbf{X}_{i\nu}^c)\mathbf{W}_{i\mu}^t)\mathbf{H}_{\mu\nu}^t.$$

   Recompute $\mathbf{X}^c = \mathbf{W}^t\mathbf{H}^{t+1}$ and update $\mathbf{W}$,

   $$\mathbf{W}_{\mu\nu}^{t+1} = (1/\textstyle\sum_j \mathbf{H}_{\nu j}^{t+1})(\sum_j (\mathbf{X}_{\mu j}/\mathbf{X}_{\mu j}^c)\mathbf{H}_{\nu j}^{t+1})\mathbf{W}_{\mu\nu}^t.$$

   Scale the columns of $\mathbf{W}^{t+1}$: Define

   $$\lambda_j = \textstyle\sum_i \mathbf{W}_{ij}^{t+1} \text{ and } \mathbf{L} = \operatorname{diag}\{\lambda_1, \cdots, \lambda_k\}$$

   and set $\mathbf{W}^{t+1} = \mathbf{W}^{t+1}\mathbf{L}^{-1}$. Compute

   $$\delta = \frac{\|\mathbf{W}^{t+1} - \mathbf{W}^t\|_{\mathrm{F}}}{\|\mathbf{W}^t\|_{\mathrm{F}}} + \frac{\|\mathbf{H}^{t+1} - \mathbf{H}^t\|_{\mathrm{F}}}{\|\mathbf{H}^t\|_{\mathrm{F}}}.$$

   Set $t = t + 1$.

5. Computed example: images as sums of their parts

- Data set: the well-known MNIST data set. The test sample contains $p = 10000$ images, and hence $\mathbf{X} \in \mathbb{R}^{784\times 10000}$.
- Test the NMF multiplicative updating algorithm I for

$$k = 9, \quad k = 81, \quad \tau = 0.01.$$

- Plot the history of the relative change, i.e., δ
- Plot the $k = 9$ and $k = 81$ columns of the feature vector matrix $\mathbf{W}$, visualized as 28×28 images.
- Observations: when k is small, NMF produces a summary of the data, compressing in few feature vectors the contents of the data; when k is relatively large, NMF decomposes the data into elementary features, or building blocks, allowing us to look for local similarities among the data vectors.