Chapter 3
Non-Negative Matrix Factorization

Part 1: Introduction & computation
Motivating NMF
Reminder

\[ A = U \Sigma V^T \]

The components of the SVD are not very interpretable.
Non-negative factors

Forcing the factors to be non-negative can, and often will, improve the interpretability of the factorization.
The definition
Definition of NMF

Given a non-negative matrix $A \in \mathbb{R}^{n \times m}_+$ and an integer $k$, find non-negative matrices $W \in \mathbb{R}^{n \times k}_+$ and $H \in \mathbb{R}^{k \times m}_+$ such that

\[ \frac{1}{2} \|A - WH\|_F^2 \]

is minimized.
Non-negative rank

- The **non-negative rank** of matrix $\mathbf{A}$, $\text{rank}_+(\mathbf{A})$, is the size of the smallest exact non-negative factorization $\mathbf{A} = \mathbf{WH}$
- $\text{rank}(\mathbf{A}) \leq \text{rank}_+(\mathbf{A}) \leq \min\{n, m\}$
Some comments

• NMF is **not** unique

• If $X$ is nonnegative and with nonnegative inverse, then $WXX^{-1}H$ is equivalent valid decomposition

• Computing NMF (and non-negative rank) is NP-hard

• This was open until 2008
Example of non-uniqueness

\[
\begin{array}{c|c|c|c|c}
1 & 1 & 1 & 1 & 1 \\
\hline
0 & 1 & 0 & 1 & 0 \\
\hline
0 & 1 & 0 & 1 & 0 \\
\end{array}
= 
\begin{array}{c|c|c|c|c}
1 & 1 & 1 & 1 & 1 \\
\hline
0 & 0 & 0 & 0 & 0 \\
\hline
0 & 0 & 0 & 0 & 0 \\
\end{array}
+ 
\begin{array}{c|c|c|c|c}
0 & 0 & 0 & 0 & 0 \\
\hline
0 & 1 & 0 & 1 & 0 \\
\hline
0 & 1 & 0 & 1 & 0 \\
\end{array}
\]

\[
\begin{array}{c|c|c|c|c}
1 & 0.5 & 1 & 0.5 & 1 \\
\hline
0 & 0 & 0 & 0 & 0 \\
\hline
0 & 0 & 0 & 0 & 0 \\
\end{array}
= 
\begin{array}{c|c|c|c|c}
1 & 0.5 & 1 & 0.5 & 1 \\
\hline
0 & 0 & 0 & 0 & 0 \\
\hline
0 & 0 & 0 & 0 & 0 \\
\end{array}
+ 
\begin{array}{c|c|c|c|c}
0 & 0.5 & 0 & 0.5 & 0 \\
\hline
0 & 1 & 0 & 1 & 0 \\
\hline
0 & 1 & 0 & 1 & 0 \\
\end{array}
\]

\[
\begin{array}{c|c|c|c|c}
1 & 0 & 1 & 0 & 1 \\
\hline
0 & 0 & 0 & 0 & 0 \\
\hline
0 & 0 & 0 & 0 & 0 \\
\end{array}
= 
\begin{array}{c|c|c|c|c}
1 & 0 & 1 & 0 & 1 \\
\hline
0 & 0 & 0 & 0 & 0 \\
\hline
0 & 0 & 0 & 0 & 0 \\
\end{array}
+ 
\begin{array}{c|c|c|c|c}
0 & 1 & 0 & 1 & 0 \\
\hline
0 & 1 & 0 & 1 & 0 \\
\hline
0 & 1 & 0 & 1 & 0 \\
\end{array}
\]
NMF has no order

• The factors in NMF have no inherent order
  • The first component is no more important than the second is no more important...

• NMF is not **hierarchical**
  • The factors of rank-($k+1$) decomposition can be completely different to those of rank-$k$ decomposition
Example

\[
\begin{array}{cccccc}
1 & 1 & 1 & 1 & 1 & 1 \\
0 & 1 & 0 & 1 & 0 & \\
0 & 1 & 0 & 1 & 0 & \\
\end{array}
\approx
\begin{array}{c}
0.8 \\
0.5 \\
0.5 \\
\end{array}
\approx
\begin{array}{cccccc}
0.7 & 1.5 & 0.7 & 1.5 & 0.7 \\
\end{array}
\}
\text{Rank-1}

= \begin{array}{cc}
1 & 0 \\
0 & 1 \\
0 & 1 \\
\end{array}
= \begin{array}{ccccccc}
1 & 1 & 1 & 1 & 1 & 1 & 1 \\
0 & 1 & 0 & 1 & 0 & \\
0 & 1 & 0 & 1 & 0 & \\
\end{array}
\}
\text{Rank-2}
Interpreting NMF
Parts-of-whole

• NMF works over anti-negative semiring
  • There is no subtraction
  • Each rank-1 component $w_i h_i$ explains a part of the whole
    • This can yield to sparse factors
NMF example: faces

Row of original

Row of reconstruction

=  \times 

PCA/SVD
NMF example: faces

Row of original

\[ \text{Row of original} = \text{NMF} \times \]
NMF example: digits

NMF factors correspond to patterns and background

A

H
Some NMF applications

- Text mining (more later)
- Bioinformatics
- Microarray analysis
- Mineral exploration
- Neuroscience
- Image understanding
- Air pollution research
- Weather forecasting
- ...

Figure 4.8 Illustration for (a) benchmark used in large-scale experiments with 10 nonnegative sources; (b) Typical 1000 mixtures; (c) Ten estimated components by using FAST HALS NMF from the observations matrix $Y$ of dimension $1000 \times 1000$. (d) Performance expressed via the PSNR using the Beta HALS NMF algorithm for $\beta = 0.5, 1, 1.5, 2$ and 3.
Computing NMF
General idea

• NMF is not convex, but it is **biconvex**

  • If $W$ is fixed, $\frac{1}{2} \|A - WH\|^2_F$ is convex

• Start from random $W$ and **repeat**

  • Fix $W$ and update $H$

  • Fix $H$ and update $W$

• **until** the error doesn’t decrease anymore
Notes on the general idea

• How to create a good random starting point?
  • Is the algorithm robust to initial solutions?

• How to update $W$ and $H$?

• When (and how quickly) has the process converged?
  • Fixed number of iterations? Minimum change in error?
Alternating least squares

- Without the non-negativity constraint, this is the standard least-squares:
  - \( w_i \leftarrow \arg \min_w \|wH - a_i\|_F \)
  - we can update \( W \leftarrow AH^+ \) and \( H \leftarrow W^+ A \)
  - \( X^+ \) is the pseudo-inverse of \( X \) which is LS-optimal
- The method is called **alternating least-squares** (ALS)
- This can introduce negative values
Enforcing non-negativity in ALS

• Least-squares optimal update of \( W \) (or \( H \)) with non-negativity constraints is convex optimization problem
  • In theory in P, in practice slow, but subject to much research
• Simple approach: truncate all negative values to 0
  • Update \( W \leftarrow [AH^+]_+ \)
The NMF-ALS algorithm

1. $W \leftarrow \text{random}(n, k)$

2. repeat

   2.1. $H \leftarrow [W^+A]_+$

   2.2. $W \leftarrow [AH^+]_+$

3. until convergence
When has there been enough convergence?

- When the error doesn’t change too much
  \[ \|A - W^{(k)}H^{(k)}\|_F - \|A - W^{(k+1)}H^{(k+1)}\|_F \leq \epsilon \]
- After some number of maximum iterations has been achieved
- Usually, whichever of these two happens first
Gradient descent

• We can compute the gradient of the error function (with one factor matrix fixed)
  • \( f(H) = \frac{1}{2} \| A - WH \|_F^2 = \frac{1}{2} \sum_i \| a_i - Wh_i \|_F^2 \)
  • \( \nabla_{H_{ij}} f(H) = (W^T A)_{ij} - (W^T WH)_{ij} \)
• We can move slightly towards the negative gradient
  • How much is the step size and deciding it is a big problem
The NMF gradient descent algorithm

1. $W \leftarrow \text{random}(n, k)$
2. $H \leftarrow \text{random}(k, m)$
3. repeat
   3.1. $H \leftarrow H - \epsilon_H \frac{\partial f}{\partial H}$
   3.2. $W \leftarrow W - \epsilon_W \frac{\partial f}{\partial W}$
4. until convergence
Oblique Projected Landweber (OPL) for NMF

• OPL provides one way to select the step size

• With $H \leftarrow H - \varepsilon H \frac{\partial f}{\partial H}$ updates, the convergence radius is $2/\lambda_{\text{max}}(W^T W)$, where $\lambda_{\text{max}}$ is the largest eigenvalue

  • $\lambda_{\text{max}} \leq \max(\text{rowSums}(W^T W))$

• We can set the learning rates to $1/\text{rowSums}(W^T W)$ for a good convergence
The OPL algorithm for updating $H$

1. $\eta \leftarrow \text{diag}(1 / \text{rowSums}(W^T W))$

2. repeat

2.1. $G \leftarrow W^T W H - W^T A$

2.2. $H \leftarrow [H - \eta G]_+$

3. until a stopping criterion is met

   (small) number of iterations

   OR

   $H$ doesn’t change much
Interior Point Gradient (IPG) for NMF

• In OPL, we might (temporarily) have negative values in $W$ or $H$

• In **Interior Point Gradient** (IPG) algorithm, we set the step sizes so that we never update to negative
The IPG algorithm for updating $H$

1. **repeat until** a stopping criterion is met

1.1. $G \leftarrow W^T(WH - A)$

1.2. $D \leftarrow H / (W^TWH)$

1.3. $P \leftarrow -D \ast G$

1.4. $\eta^* \leftarrow -\langle \text{vec}(P), \text{vec}(G) \rangle / \langle \text{vec}(WP), \text{vec}(WP) \rangle$

1.5. $\eta' \leftarrow \max\{\eta : H + \eta P \geq 0\}$

1.6. $\eta \leftarrow \min\{\tau \eta', \eta^*\}$

1.7. $H \leftarrow H + \eta P$

---

Gradient

Scaling

Update direction

/ and $\ast$ are element-wise

Best step size

Positive step size

Update
Multiplicative updates

• The KKT conditions for $H$ in NMF are
  
  • $H \geq 0; \nabla_H ||A - WH||^2/2 \geq 0$
  
  • $H * \nabla_H ||A - WH||^2/2 = 0$

• Substituting $\nabla_H ||A - WH||^2/2 = W^TWH - W^TA$
  one gets $H * (W^TWH) = H * (W^TA)$

• This gives us an update rule for $H$
The NMF multiplicative updates algorithm

1. $W \leftarrow \text{random}(n, k)$

2. $H \leftarrow \text{random}(k, m)$

3. repeat

   3.1. $h_{ij} \leftarrow h_{ij} \frac{(W^T A)_{ij}}{(W^T WH)_{ij} + \varepsilon}$

   3.2. $w_{ij} \leftarrow w_{ij} \frac{(AH^T)_{ij}}{(WHH^T)_{ij} + \varepsilon}$

4. until convergence
Notes on multiplicative updates

• Proposed by Lee & Seung (Nature, 1999)

• Equivalent to gradient descent with dynamic step size

• Zeros in initial solutions will never turn into non-zeros; non-zeros will never turn into zeros

• Problems if the correct solution contains zeros
Summary

• NMF can provide factorizations that are more interpretable than those given by SVD
• Harder to compute than SVD, but many different approaches
  • Or are they so different...
• In two weeks: Applications & alternations of NMF... *Stay tuned!*