On the Equivalence of Nonnegative Matrix Factorization and Spectral Clustering

Chris Ding, Xiaofeng He, Horst D. Simon

Published on SDM 05’

Hongchang Gao

Outline

- NMF
- NMF⇔Kmeans
- NMF⇔Spectral Clustering
- NMF⇔Bipartite graph Kmeans

Outline

- NMF
- NMF⇔Kmeans
- NMF⇔Spectral Clustering
- NMF⇔Bipartite graph Kmeans

NMF

- Paatero and Tapper (1994)
  - Positive matrix factorization: a non-negative factor model with optimal utilization of error estimates of data values
  - Environmetrices
- Lee and Seung (1999, 2000)
  - Learning the parts of objects by non-negative matrix factorization, Nature
  - Algorithms for non-negative matrix factorization, NIPS

NMF

- Matrix Factorization is widely used in machine learning, such as SVD

$$X = \underset{mixed}{U} \quad \underset{nonneg}{\Sigma} \quad \underset{mixed}{V^T}$$

  - interpretation of basis vectors is difficult due to mixed signs

NMF

- Nonnegative Matrix Factorization

$$X = \underset{nonneg}{F} \quad \underset{nonneg}{G^T}$$

  - where $X \in R^{d \times n}, F \in R^{d \times k}, G \in R^{n \times k}$
  - columns of F are the underlying basis vectors
  - rows of G give the weights associated with each basis vector

Outline

- NMF
- NMF⇔Kmeans
- NMF⇔Spectral
- NMF⇔Bipartite graph Kmenas

Kmeans

- Kmeans clustering is one of most widely used clustering method.

Given n points in m-dim: $X = (x_1, x_2, \cdots, x_n)^T$

K-means objective $\min J_K = \sum_{k=1}^{K} \sum_{i \in C_k} \| x_i - c_k \|^2$

Kmeans

- Reformulate Kmeans Clustering

$$J_K = \sum_i \|x_i\|^2 - \sum_{k=1}^{K} \frac{1}{n_k} \sum_{i,j \in C_k} x_i^T x_j$$

- Cluster membership indicators:

$$h_k = (0 \cdots 0, \overbrace{1 \cdots 1}^{n_k}, 0 \cdots 0)^T / n_k^{1/2}$$

$$J_K = \sum_i x_i^2 - \sum_{k=1}^{K} h_k^T X^T X h_k$$

Kmeans

- Objective function

$$\max_{H^T H = I, H \geq 0} \mathrm{Tr}(H^T X^T X H)$$

- Replace $W = X^T X$, which is the standard inner-product linear Kernel matrix

$$\max_{H^T H = I, H \geq 0} \mathrm{Tr}(H^T W H)$$

Kernel Kmeans

- Map x to higher dimension space:

$$x_i \rightarrow \phi(x_i)$$

- Kernel Kmeans objective:

$$\min J_K^{\phi} = \sum_{k=1}^{K} \sum_{i \in C_k} \| \phi(x_i) - \phi(c_k) \|^2$$

$$J_K^{\phi} = \sum_i | \phi(x_i) |^2 - \sum_{k=1}^{K} \frac{1}{n_k} \sum_{i,j \in C_k} \phi(x_i)^T \phi(x_j)$$

$$\max J_K^{\phi} = \sum_{k=1}^{K} \frac{1}{n_k} \sum_{i,j \in C_k} \left\langle \phi(x_i), \phi(x_j) \right\rangle = \mathrm{Tr}(H^T W H)$$

NMF⇔Kmeans

- Orthogonal symmetric NMF is equivalent to Kernel Kmeans clustering

$$\text{Symmetric NMF} \quad \min_{H^T H = I, H \geq 0} \| W - HH^T \|^2$$

$$\text{Is Equivalence to} \quad \max_{H^T H = I, H \geq 0} \mathrm{Tr}(H^T W H)$$

Kernel Kmeans=>Symmetric NMF

- Factorization is equivalent to Kernel K-means clustering with the strict orthogonality relaxed

$$H = \underset{H^T H = I,\, H \geq 0}{\arg\max}\ Tr(H^T W H)$$

$$= \underset{H^T H = I,\, H \geq 0}{\arg\min}\ -2Tr(H^T W H)$$

$$= \underset{H^T H = I,\, H \geq 0}{\arg\min}\ \| W \|^2 - 2Tr(H^T W H) + \| H^T H \|^2$$

$$= \underset{H^T H = I,\, H \geq 0}{\arg\min}\ \| W - HH^T \|^2$$

Relaxing the orthogonality $H^T H = I$ completes the proof

Symmetric NMF=> Kernel Kmeans

- $W = HH^T$ factorization retains H orthogonality approxiamately.
  - Proof. $\min\|W - HH^T\|^2$ is equivalent to

$$\max_{H \geq 0} Tr(H^T W H)$$

$$\min_{H \geq 0} \| H^T H \|^2$$

  - The first one recover the objective

Symmetric NMF=> Kernel Kmeans

- The second one

$$||H^T H||^2 = \sum_{\ell k} (H^T H)^2_{\ell k} = \sum_{\ell \neq k} (\mathbf{h}_\ell^T \mathbf{h}_k)^2 + \sum_k (\mathbf{h}_k^T \mathbf{h}_k)^2$$

  - Minimize the first term, we get

$$\mathbf{h}_\ell^T \mathbf{h}_k \approx 0$$

  - Minimize the second term

$$\min \ ||\mathbf{h}_1||^4 + \cdots + ||\mathbf{h}_K||^4$$

    - We should make sure H cannot be all zero

$$\sum_{ij} w_{ij} \approx \sum_{ij} (HH^T)_{ij} = \sum_{kij} h_{ik} h_{jk} = \sum_k |\mathbf{h}_k|^2$$

Outline

- NMF
- NMF⇔Kmeans
- NMF⇔Spectral
- NMF⇔Bipartite graph Kmeans

Spectral Clustering

- Spectral clustering objective functions

$$J = \sum_{1 \le p < q \le K} \frac{s(C_p, C_q)}{\rho(C_p)} + \frac{s(C_p, C_q)}{\rho(C_q)} = \sum_{k=1}^{K} \frac{s(C_k, \bar{C}_k)}{\rho(C_k)}$$

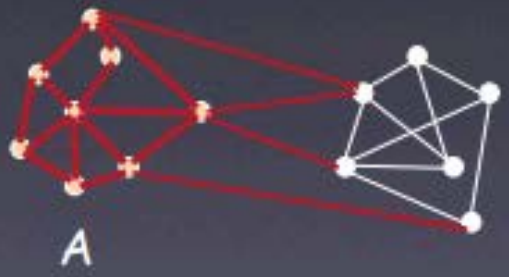
$$\rho(C_k) = \begin{cases} |C_k| & \text{for Ratio Cut} \\ \sum_{i \in C_k} d_i & \text{for Normalized Cut} \\ s(C_k, C_k) & \text{for MinMax Cut} \end{cases}$$

Graph Terminology

Similarity matrix $S = [\, s_{ij} \,]$

Degree of node: $d_i = \sum_j S_{ij}$

Volume of set:



Graph Cuts

Spectral Clustering

- Reformulate the objective based on Ncut

$$J = \frac{1}{K}\sum_{l=1}^{K}\frac{cut(V_l, V - V_l)}{vol(V_l)} = \frac{1}{K}\sum_{l=1}^{K}\frac{h_l^T (D - W) h_l}{h_l^T D h_l}$$

- Replace

$$z_l = \frac{D^{1/2} h_l}{\| D^{1/2} h_l \|}$$

- Then,

$$J = \sum_{l=1}^{K} z_l^T (I - \tilde{W}) z_l = \sum_{l=1}^{K} z_l^T z_l - Tr(Z^T \tilde{W} Z)$$

NMF⇔Spectral Clustering

- The objective of spectral clustering

$$\max_{Z^T Z = I, Z \geq 0} Tr(Z^T \tilde{W} Z)$$

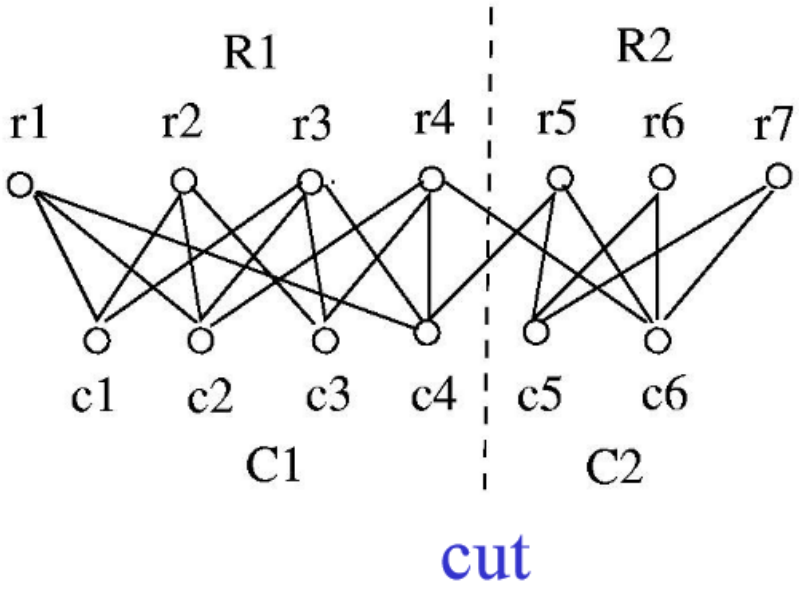
- This is identical to the Kernel Kmeans clustering
- Spectral Clustering⇔Kernel Kmeans ⇔ NMF

Outline

- NMF
- NMF⇔Kmeans
- NMF⇔Spectral
- NMF⇔Bipartite graph Kmenas

Bipartite graph Kmeans

- Simultaneous clustering of rows and columns



row indicators

$$\begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix} = (f_1, f_2, f_3) = F$$

column indicators

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} = (g_1, g_2, g_3) = G$$

Bipartite graph Kmeans

- Simultaneously cluster the rows and columns of data matrix

$$B = (x_1, x_{2,\ldots,} x_n)$$

- Row Clustering

$$\max_{F^T F = I, F \geq 0} Tr(F^T B B^T F)$$

- Column Clustering

$$\max_{G^T G = I, G \geq 0} Tr(G^T B^T B G)$$

Bipartite graph Kmeans

- Equivalent problem:

$$\max_{\substack{F^T F = I; \\ G^T G = I; \\ F, G \geq 0}} J_2 = \frac{1}{2}\mathrm{Tr}\begin{pmatrix} F \\ G \end{pmatrix}^T \begin{pmatrix} 0 & B \\ B^T & 0 \end{pmatrix} \begin{pmatrix} F \\ G \end{pmatrix} = \mathrm{Tr}\ F^T B G$$

- Solution

$$Bg_k = \lambda_k f_k, B^T f_k = \lambda_k g_k$$

- Then,

$$B^T B g_k = \lambda_k^2 g_k, BB^T f_k = \lambda_k^2 f_k$$

Bipartite graph Kmeans=>NMF

- The simultaneous row and column Kmeans clustering is equivalent to the following optimization problem

$$\min_{\substack{F^T F = I; \\ G^T G = I; \\ F, G \geq 0}} ||B - FG^T||^2$$

Bipartite graph Kmeans=>NMF

- Proof.

$$\max_{F,G} Tr(F^T BG)$$

$$\Rightarrow \min_{F,G} -Tr(F^T BG)$$

$$\Rightarrow \min_{F,G} \| B \|^2 - 2Tr(F^T BG) + Tr(F^T FG^T G)$$

$$\Rightarrow \min_{F,G} \| B - FG^T \|^2$$

- Therefore, NMF is equivalent to Kmeans clustering with relaxed orthogonality contraints.

NMF=>Bipartite graph Kmeans

- In the previous, we assume both F and G are orthogonal. If one of them is orthogonal, we can explicitly write $\| B - FG^T \|^2$ as a Kmeans clustering objective function.
- NMF with orthogonal G is identical to Kmeans clustering of the columns of B.

NMF=>Bipartite graph Kmeans

- Proof.
  - At first, normalize the row of G, s.t. $\sum_{r=1}^{k} g_{ir} = 1$
  - Then, for the objective function

$$J_2 = \|B - FG^T\|^2 = \sum_{i=1}^{n} \left\| \mathbf{x}_i - \sum_{k=1}^{\kappa} g_{ik}\mathbf{f}_k \right\|^2$$

  - We have

$$\left\| \sum_{k=1}^{\kappa} g_{ik}(\mathbf{x}_i - \mathbf{f}_k) \right\|^2 = \sum_{k=1}^{\kappa} g_{ik}^2 ||\mathbf{x}_i - \mathbf{f}_k||^2 = \sum_{k=1}^{\kappa} g_{ik} ||\mathbf{x}_i - \mathbf{f}_k||^2$$

NMF=>Bipartite graph Kmeans

- The orthogonality condition of G implies that in each row of G, only one element is nonzero and $g_{ik} = 0,1$
- Summing over i:

$$J_2 = \sum_{k=1}^{K} \sum_{i \in C_k} \| x_i - f_k \|^2$$

  – which is the Kmeans clustering

Reference

- Ding, Chris HQ, Xiaofeng He, and Horst D. Simon. "On the Equivalence of Nonnegative Matrix Factorization and Spectral Clustering." *SDM*. Vol. 5. 2005.
- Li, Tao, and Chris Ding. "The relationships among various nonnegative matrix factorization methods for clustering." *Data Mining, 2006. ICDM'06. Sixth International Conference on*. IEEE, 2006.
- Von Luxburg, Ulrike. "A tutorial on spectral clustering." *Statistics and computing* 17.4 (2007): 395-416.
- Shi, Jianbo, and Jitendra Malik. "Normalized cuts and image segmentation." *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 22.8 (2000): 888-905.

Thanks