Non-negative Matrix Factorization

Runze Mao

1 K-means

Given the non-negative data matrix $X \in R_+^{p\times n}$, the objective is to divide the n column vectors into K clusters. The problem is to solve the objective function:

$$\min \sum_k^K \sum_{i\in C_k} \|x_i - m_k\|_2^2 \tag{1.1}$$

where m_k is the centroid vector of the k-th cluster, that is to say, $m_k = \frac{1}{n_k}\sum_{i\in C_k} x_i$, and n_k is the size of the k-th cluster.

Definition 1.1 Let $H \in R_+^{n\times K}$ be $[h_1, \ldots, h_K]$ where:

$$h_{ik} = \begin{cases} 1/\sqrt{n_k} & , x_i \text{ belongs to } C_k \\ 0 & , \text{otherwise} \end{cases}$$

It is obvious that $H^T H = I$. H is called the *indicator matrix* of X for its ability to indicate whether the k-th cluster includes each column vector of X.

Theorem 1.1 The K-means problem defined by (1.1) is equivalent to the problem:

$$\max_{H^T H = I, H\geq 0} tr(H^T W H) \tag{1.2}$$

where $W = X^T X$.

Proof Assume $J_1 = \sum_k^K \sum_{i\in C_k} \|x_i - m_k\|_2^2$, and $J_2 = tr(H^T X H)$.

$$\begin{aligned}
J_1 =& \sum_{k=1}^K \sum_{i\in C_k} \left(\|x_i\|_2^2 - 2x_i^T m_k + \|m_k\|_2^2\right) \\
=& \sum_{k=1}^K \|x_i\|_2^2 - 2\sum_{k=1}^K \sum_{i\in C_k} x_i^T \sum_{j\in C_k} x_j/n_k \\
&+ \sum_{k=1}^K n_k \big(\sum_{j\in C_k} x_j/n_k\big)^2 \\
=& \sum_{k=1}^K \|x_i\|_2^2 - \sum_{k=1}^K \frac{2}{n_k} \sum_{i,j\in C_k} x_i^T x_j \\
&+ \sum_{k=1}^K \frac{1}{n_k} \sum_{i,j\in C_k} x_i^T x_j \\
=& \sum_{k=1}^K \|x_i\|_2^2 - \sum_{k=1}^K \frac{1}{n_k} \sum_{i,j\in C_k} x_i^T x_j
\end{aligned}$$

Notice that

$$J_2 = \sum_{k=1}^K h_k^T W h_k$$

where h_k^T selects the rows of W and h_k selects columns. And after extracting the common factor $1/\sqrt{n_k}$:

$$\begin{aligned}
J_2 =& \sum_{k=1}^K \frac{1}{\sqrt{n_k}} \frac{1}{\sqrt{n_k}} \sum_{i,j\in C_k} w_{ij} \\
=& \sum_{k=1}^K \frac{1}{n_k} \sum_{i,j\in C_k} x_i^T x_j
\end{aligned}$$

Therefore, minimizing J_1 is equivalent to maximizing J_2, which is actually the weighted within-cluster similarities.

Moreover, substituting $\kappa(x_i, x_j)$ for $x_i^T x_j$ leads to the *kernal* K-means.

Theorem 1.2 The problem (1.2) can be solved by *symmetric* NMF:

$$\min_{H^T H = I, H\geq 0} \|W - HH^T\|_{\mathcal{F}}^2$$

Proof To maximize J_2 is to minimize $J_3 = \|W\|_{\mathcal{F}}^2 - 2tr(H^T W H) + \|H^T H\|_{\mathcal{F}}^2$ because W and $H^T H$ are both constants. Furthermore,

$$\begin{aligned}
\|W - HH^T\|_{\mathcal{F}}^2 =& \sum_{i,j} \left(w_{ij} - (HH^T)_{ij}\right)^2 \\
=& \sum_{i,j} w_{ij}^2 - 2\sum_{i,j} w_{ij}(HH^T)_{ij} + \sum_{i,j}(HH^T)_{ij}^2 \\
=& \|W\|_{\mathcal{F}}^2 - 2tr(WHH^T) + \|HH^T\|_{\mathcal{F}}^2 \\
=& \|W\|_{\mathcal{F}}^2 - 2tr(H^T W H) + \|H^T H\|_{\mathcal{F}}^2 \\
=& J_3
\end{aligned}$$

Therefore, maximizing J_2 is equivalent to minimizing $\|W - HH^T\|_{\mathcal{F}}^2$, with the non-negativity and orthogonality constraints. Interestingly, even if the strict orthogonality is relaxed, we can still keep $H^T H \approx I$ [1].

Theorem 1.3 The K-means problem is also equivalent to the general NMF:

$$
\begin{aligned}
\min \|X - FG^T\|_{\mathcal{F}}^2 \\
\text{subject to } G^T G = I, \qquad (1.3)\\
F \geq 0, G \geq 0
\end{aligned}
$$

where $F \in R_+^{p\times K}, G \in R_+^{n\times K}$ and G is in fact the indicator matrix of X.

Proof We first show that the orthogonality together with non-negativity implies that in each row of G, at most one element is non-zero.

Assume g_i, g_j are two different column vectors of G, *i.e.*, $i \neq j$. Because of the orthogonality, $g_i^T g_j = 0$. Because of the non-negativity, $\forall l : g_{li} g_{lj} \geq 0$, which means $g_i^T g_j = \sum_{l=1}^n g_{li} g_{lj} = 0$ *iff.* $\forall l : g_{li} g_{lj} = 0$. So, for any pair of different g_i and g_j, at most one element is non-zero in one row. Thus, we can get the conclusion that in each row of G, at most one element is non-zero. Next, we show that

$$
\begin{aligned}
J_4 =& \|X - FG^T\|_{\mathcal{F}}^2 \\
=& \|X\|_{\mathcal{F}}^2 - 2tr(XGF^T) + tr(FG^T GF^T) \\
=& \|X\|_{\mathcal{F}}^2 - 2tr(F^T XG) + tr(F^T F)
\end{aligned}
$$

Therefore, $\partial \frac{J_4}{\partial F} = -2XG + 2F$, which indicates that at the optimal points, $F = XG$. Thus, $J_4 = \|X\|_{\mathcal{F}}^2 - tr(G^T X^T XG)$. So, minimizing J_4 is equivalent to maximizing $tr(G^T X^T XG)$ where $G^T G = I$ and $G \geq 0$. According to Theorem 1.1, this is identical to K-means clustering, and F consists of the centroids of the clusters [2].

Theorem 1.4 Adding the orthogonality of F to (1.3) results in the *co-clustering* problem:

$$
\begin{aligned}
\min \|X - FG^T\|_{\mathcal{F}}^2 \qquad (1.4)\\
\text{subject to } G \geq 0, G^T G = I, \\
F \geq 0, F^T F = I
\end{aligned}
$$

F and G are the indicator matrices for the rows and columns of X, respectively.

Proof According to Theorem 1.1, the co-clustering problem is to simultaneously solve:

$$
\begin{cases}
\max tr(G^T X^T XG), s.t.\ G \geq 0, G^T G = I \\
\max tr(F^T XX^T F), s.t.\ F \geq 0, F^T F = I
\end{cases} \qquad (1.5)
$$

which is equivalent to

$$
\begin{aligned}
&\max J_5 = \frac{1}{2} tr(G^T X^T XG + F^T XX^T F) \\
&s.t.\ G \geq 0, G^T G = I, \\
&\quad F \geq 0, F^T F = I
\end{aligned}
$$

We can simplify J_5 into

$$
J_5 = \frac{1}{2} tr(\begin{bmatrix} F \\ G \end{bmatrix}^T \begin{bmatrix} 0 & X \\ X^T & 0 \end{bmatrix}^2 \begin{bmatrix} F \\ G \end{bmatrix})
$$

Because the matrices are all non-negative, minimizing J_5 is equivalent to

$$
\max J_6 = \frac{1}{2} tr(\begin{bmatrix} F \\ G \end{bmatrix}^T \begin{bmatrix} 0 & X \\ X^T & 0 \end{bmatrix} \begin{bmatrix} F \\ G \end{bmatrix})
$$

with all the constraints preserved *(This is just a conjecture which I cannot prove yet. Another rigorous proof is given in [1])*. Notice that

$$
J_6 = tr(F^T XG)
$$

and that

$$
\|X - FG^T\|_{\mathcal{F}}^2 = \|X\|_{\mathcal{F}}^2 - 2tr(F^T XG) + \|FG^T GF^T\|_{\mathcal{F}}^2
$$

Since the first and third terms are both constants, (1.4) is equivalent to maximizing J_6, thus equivalent to (1.5).

References

[1] C. Ding, X. He, H. D. Simon. On the Equivalence of Nonnegative Matrix Factorization and Spectral Clustering. Proc. SIAM Data Mining Conf., 2005.

[2] C. Ding, T. Li, W. Peng, H. Park. Orthogonal nonnegative matrix tri-factorizations for clustering. In SIGKDD, 2006:126–135.