Spectral Clustering and Normalized Cut

Kai Yao[1,a]

University of the Balearic Islands, Carretera de Valldemossa, km 7.5, 07122 Palma, Illes Balears, Spain

k.yao@uib.es

Graph Cut

An example of graph cut is shown in Fig. 1, which is equivalent to the clustering result. Define $w(A, B)$, where

$$A \subset V, B \subset V, A \cap B = \emptyset, W(A, B) = \sum_{i \in A} \sum_{i \in B} w_{ij} \tag{1}$$

Given K categories in the dataset, the cut of this graph is

$$Cut(V) = Cut(v_1, v_2, \cdots, v_k) = \sum_{k=1}^{K} W(A_k, \overline{A_k}) = \sum_{k=1}^{K} [W(A_k, V) - W(A_k, A_k)], \text{ where } \overline{A_k} = V - A_k \tag{2}$$

In order to balance the weights inside of each category, a normalization is required. Define the degree of a node, which is represented as $d_i = \sum_{i=1}^{n} w_{ij}$. Then, the degree of a set is $\Delta_k = \text{degree}(A_k) = \sum_{i \in A_k} d_i$. Therefore, the normalized cut is

$$NCut = \sum_{k=1}^{K} \frac{w(A_k, \overline{A_k})}{\sum_{i \in A_k} d_i} \tag{3}$$

In a weighted undirected graph, the degree of a node is the sum of the weights of all the edges connected to the node. Therefore,

$$\text{degree}(A_k) = \sum_{i \in A_k} d_i, \text{ where } d_i = \sum_{j=1}^{N} w_{ij} \tag{4}$$

Spectral Clustering

There are two kinds of ideas in clustering methods, which can be categorized as:

- Compactness, such as K-means and GMM, which are used to deal with convex datasets.
- Connectivity, such as spectral clustering

Spectral clustering is based on weighted undirected graphical models, which is represented by a graph $G = \{E, V\}$, where $V = \{v_1, v_2, \cdots, v_n\}, E = \{w_{ij}\}$. Here, w_{ij} is the similarity between two nodes, and W is the similarity (affinity) matrix. As usual, w_{ij} is calculated by the RBF kernel, as shown below.

$$\begin{cases} w_{ij} = k(x_i, x_j) = \exp(-\frac{||x_i - x_j||_2^2}{2\sigma^2}), (i, j) \in E \\ w_{ij} = 0, (i, j) \notin E \end{cases} \tag{5}$$

Let,

$$\begin{cases} y_i \in \{0, 1\}^k \\ \sum_{j=1}^{k} y_{ij} = 1 \end{cases} \tag{6}$$

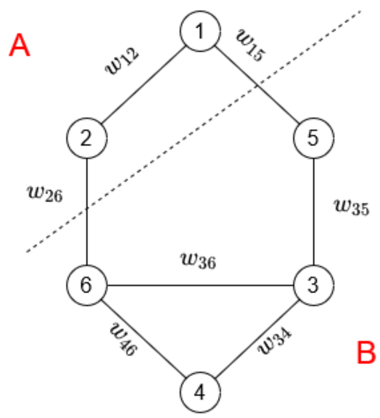

Fig. 1: Graph cut

So, $y_i \in \mathbb{R}^{K\times 1}$ is an one-hot vector. Let $Y = \{y_1, y_2, \cdots, y_n\}^T, Y \in \mathbb{R}^{N\times K}$. Therefore the spectral clustering model is

$$
\begin{aligned}
\{A_k\}_{k=1}^{K} &= \arg\min_{k} NCut \\
\hat{Y} &= \arg\max_{Y} \sum_{k=1}^{K} \frac{W(A_k, \overline{A_K})}{\sum_{i\in A_k} d_i}, d_i = \sum_{j=1}^{N} w_{ij}
\end{aligned}
\tag{7}
$$

As described in Eq. 3, the result of $NCut$ is a scalar, therefore we can use the trace of matrix to represent Eq. 3 as below.

$$
\begin{aligned}
NCut = \sum_{k=1}^{K} \frac{W(A_k, \overline{A}_k)}{\sum_{i\in A_k} d_i} &= tr \begin{Bmatrix} \frac{W(A_1,\overline{A_1})}{\sum_{i\in A_1} d_1} & \cdots & \cdots & 0 \\ 0 & \frac{W(A_2,\overline{A_2})}{\sum_{i\in A_2} d_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & \cdots & \frac{W(A_k,\overline{A_k})}{\sum_{i\in A_k} d_k} \end{Bmatrix}_{K\times K} \\
&= tr \left[\underbrace{\begin{Bmatrix} W(A_1, \overline{A_1}) & \cdots & \cdots & 0 \\ 0 & W(A_2, \overline{A_2}) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & \cdots & W(A_k, \overline{A_k}) \end{Bmatrix}_{K\times K}}_{O} \underbrace{\begin{Bmatrix} \sum_{i\in A_1} d_1 & \cdots & \cdots & 0 \\ 0 & \sum_{i\in A_2} d_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & \cdots & \sum_{i\in A_k} d_k \end{Bmatrix}_{K\times K}^{-1}}_{P} \right] =
\end{aligned}
\tag{8}
$$

In spectral clustering, the affinity matrix $W \in \mathbb{R}^{N\times N}$ and the labels $Y \in \mathbb{R}^{N\times K}$ are known.

Firstly, look at the matrix of $Y^TY \in \mathbb{R}^{K\times K}$.

$$
\begin{aligned}
Y^TY = \sum_{i=1}^{K} y_i^T y_i = \{y_1, y_2, \cdots, y_n\} \begin{Bmatrix} y_1^T \\ y_2^T \\ \vdots \\ y_n^T \end{Bmatrix} \\
= \begin{Bmatrix} \sum\limits_{i\in A_1} 1 & 0 & \cdots & 0 \\ 0 & \sum\limits_{i\in A_2} 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sum\limits_{i\in A_k} 1 \end{Bmatrix}_{K\times K} \\
\text{Define } D = \begin{Bmatrix} d_1 & 0 & \cdots & 0 \\ 0 & d_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & d_n \end{Bmatrix}_{N\times N} \\
\therefore \sum_{i=1}^{K} y_i^T d_i y_i = \begin{Bmatrix} \sum\limits_{i\in A_1} d_1 & 0 & \cdots & 0 \\ 0 & \sum\limits_{i\in A_2} d_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sum\limits_{i\in A_k} d_k \end{Bmatrix}_{K\times K} = Y^TDY = P
\end{aligned}
\tag{9}
$$

Secondly, compute the matrix O. As known, $W(A_i, \overline{A_i}) = W(A_i, V) - W(A_i, A_i)$, therefore

$$
\begin{aligned}
O = \underbrace{\begin{Bmatrix} W(A_1, V) & \cdots & \cdots & 0 \\ 0 & W(A_2, V) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & \cdots & W(A_k, V) \end{Bmatrix}_{K\times K}}_{\sum\limits_{i\in A_k} d_i} - \underbrace{\begin{Bmatrix} W(A_1, A_1) & \cdots & \cdots & 0 \\ 0 & W(A_2, A_2) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & \cdots & W(A_k, A_k) \end{Bmatrix}_{K\times K}}_{b} \\
= Y^TDY - \begin{Bmatrix} W(A_1, A_1) & \cdots & \cdots & 0 \\ 0 & W(A_2, A_2) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & \cdots & W(A_k, A_k) \end{Bmatrix}_{K\times K}
\end{aligned}
\tag{10}
$$

Look at the matrix Y^TWY,

$$
\begin{aligned}
Y^TWY &= \{y_1, y_2, \cdots, y_n\}_{K\times N} \begin{Bmatrix} W_{11} & W_{12} & \cdots & W_{1N} \\ W_{21} & W_{22} & \cdots & W_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ W_{N1} & W_{N2} & \cdots & W_{NN} \end{Bmatrix}_{N\times N} \begin{Bmatrix} y_1^T \\ y_2^T \\ \vdots \\ y_n^T \end{Bmatrix}_{N\times K} \\
&= \left\{ \sum_{i=1}^{N} y_i W_{i1} \quad \sum_{i=1}^{N} y_i W_{i2} \quad \cdots \quad \sum_{i=N}^{N} y_i W_{iN} \right\}_{K\times N} \begin{Bmatrix} y_1^T \\ y_2^T \\ \vdots \\ y_n^T \end{Bmatrix}_{N\times K} \\
&= \begin{Bmatrix} \sum\limits_{i\in A_1}\sum\limits_{j\in A_1} w_{ij} & \sum\limits_{i\in A_1}\sum\limits_{j\in A_2} w_{ij} & \cdots & \sum\limits_{i\in A_1}\sum\limits_{j\in A_K} w_{ij} \\ \sum\limits_{i\in A_2}\sum\limits_{j\in A_1} w_{ij} & \sum\limits_{i\in A_2}\sum\limits_{j\in A_2} w_{ij} & \cdots & \sum\limits_{i\in A_2}\sum\limits_{j\in A_K} w_{ij} \\ \vdots & \vdots & \ddots & \vdots \\ \sum\limits_{i\in A_K}\sum\limits_{j\in A_1} w_{ij} & \sum\limits_{i\in A_K}\sum\limits_{j\in A_2} w_{ij} & \cdots & \sum\limits_{i\in A_K}\sum\limits_{j\in A_K} w_{ij} \end{Bmatrix}_{K\times K} \\
\because\ W(A_i, A_j) &= \sum_{i\in A_K}\sum_{j\in A_K} w_{ij} \\
\therefore\ Y^TWY &= \begin{Bmatrix} W(A_1,A_1) & W(A_1,A_2) & \cdots & W(A_1,A_K) \\ W(A_2,A_1) & W(A_2,A_2) & \cdots & W(A_2,A_K) \\ \vdots & \vdots & \ddots & \vdots \\ W(A_K,A_1) & W(A_K,A_2) & \cdots & W(A_K,A_K) \end{Bmatrix}_{K\times K} \\
\therefore\ tr(b) &= tr(Y^TWY) \\
\therefore\ O' &= Y^TDY - Y^TWY \\
\therefore\ tr(O) &= tr(O') \\
\therefore\ tr(OP^{-1}) &= tr(O'P^{-1})
\end{aligned} \tag{11}
$$

Therefore, the loss function of spectral clustering is

$$
\hat{Y} = \arg\min_{Y} tr(Y^T \underbrace{(D-W)}_{\text{laplacian matrix}} Y \cdot (Y^TDY)^{-1}) \tag{12}
$$

Normalized Cut

Given a partition of nodes of a graph $G = (V, E)$ into two sets A and B, let x be an $N = |A|$ dimensional indicator vector, $x_i = 1$ if node i is in A and -1, otherwise. Let $d(i) = \sum_j w(i, j)$ represent the degree of node i. With the definition x and d, we can write the $NCut(A, B)$ as:

$$
\begin{aligned}
NCut(A,B) &= \frac{cut(A,B)}{assoc(A,V)} + \frac{cut(B,A)}{assoc(B,V)} \\
&= \frac{\sum\limits_{x_i>0, x_j<0} -w_{ij}x_ix_j}{\sum\limits_{x_i>0} d_i} + \frac{\sum\limits_{x_i<0, x_j>0} -w_{ij}x_ix_j}{\sum\limits_{x_i<0} d_i}
\end{aligned} \tag{13}
$$

Define D be an $N \times N$ diagonal matrix with d on its diagonal, W be an $N \times N$ symmetrical matrix with $W(i,j) = w_{ij}$,

$$k = \frac{\sum\limits_{x_i>0} d_i}{\sum\limits_i d_i}$$

and $\mathbf{1}$ be an $N \times 1$ vector of all ones. Using the fact $\frac{\mathbf{1}+x}{2}$ and $\frac{\mathbf{1}-x}{2}$ are indicator vectors for $x_i > 0$ and $x_i < 0$, respectively. For the case of $\frac{\mathbf{1}+x}{2}$, every item (position i, for instance) in $\frac{(\mathbf{1}+x)^T}{2} W$ represents the weights sum between the node i and all nodes in the subset A. Therefore, $\frac{(\mathbf{1}+x)^T}{2} W \frac{(\mathbf{1}-x)}{2}$ denotes the weights sum between nodes in subset A and B. Therefore,

$$\begin{aligned}
&\frac{(\mathbf{1}+x)^T}{2} W \frac{(\mathbf{1}-x)}{2} \\
&= \frac{(\mathbf{1}+x)^T}{2} W \left(\mathbf{1} - \frac{(\mathbf{1}+x)}{2}\right) \\
&= \frac{(\mathbf{1}+x)^T}{2} W\mathbf{1} - \frac{(\mathbf{1}+x)^T}{2} W \frac{(\mathbf{1}+x)}{2} \\
\because\ &\frac{(\mathbf{1}+x)^T}{2} diag(W\mathbf{1}) \frac{(\mathbf{1}+x)}{2} \\
&= \{1 \quad 0 \quad \cdots \quad 1\} \begin{Bmatrix} d_1 & 0 & \cdots & 0 \\ 0 & d_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & d_n \end{Bmatrix} \begin{Bmatrix} 1 \\ 0 \\ \vdots \\ 1 \end{Bmatrix} \\
&= \{1 \quad 0 \quad \cdots \quad 1\} \begin{Bmatrix} d_1 \\ 0 \\ \vdots \\ d_n \end{Bmatrix} \\
\because\ &\{1 \quad 0 \quad \cdots \quad 1\} \begin{Bmatrix} d_1 \\ d_2 \\ \vdots \\ d_n \end{Bmatrix} = \{1 \quad 0 \quad \cdots \quad 1\} \begin{Bmatrix} d_1 \\ 0 \\ \vdots \\ d_n \end{Bmatrix} = \frac{(\mathbf{1}+x)^T}{2} W\mathbf{1} \\
\therefore\ &\frac{(\mathbf{1}+x)^T}{2} W \frac{(\mathbf{1}-x)}{2} \\
&= \frac{(\mathbf{1}+x)^T}{2} diag(W\mathbf{1}) \frac{(\mathbf{1}+x)}{2} - \frac{(\mathbf{1}+x)^T}{2} W \frac{(\mathbf{1}+x)}{2} \\
&= \frac{(\mathbf{1}+x)^T}{2} (D-W) \frac{(\mathbf{1}+x)}{2}
\end{aligned} \tag{14}$$

Rewrite the $Ncut$ as below,

$$\begin{aligned}
Ncut(A,B) &= \frac{\frac{(\mathbf{1}+x)^T}{2}(D-W)\frac{(\mathbf{1}+x)}{2}}{k\mathbf{1}^T D\mathbf{1}} + \frac{\frac{(\mathbf{1}-x)^T}{2}(D-W)\frac{(\mathbf{1}-x)}{2}}{(1-k)\mathbf{1}^T D\mathbf{1}} \\
&= \frac{1}{4}\left[\frac{(\mathbf{1}+x)^T(D-W)(\mathbf{1}+x)}{k\mathbf{1}^T D\mathbf{1}} + \frac{(\mathbf{1}-x)^T(D-W)(\mathbf{1}-x)}{(1-k)\mathbf{1}^T D\mathbf{1}}\right] \\
&= \frac{1}{4}\left[\frac{x^T(D-W)x + \mathbf{1}^T(D-W)\mathbf{1}}{k(1-k)\mathbf{1}^T D\mathbf{1}} + \frac{2(1-2k)\mathbf{1}^T(D-W)\mathbf{1}}{k(1-k)\mathbf{1}^T D\mathbf{1}}\right]
\end{aligned} \tag{15}$$

Let,

$$
\begin{aligned}
\alpha(x) &= x^T(D-W)x \\
\beta(x) &= \mathbf{1}^T(D-W)x \\
\gamma &= \mathbf{1}^T(D-W)\mathbf{1} \\
M &= \mathbf{1}^T D\mathbf{1}
\end{aligned} \tag{16}
$$

$NCut(A,B)$ can change to,

$$
\begin{aligned}
NCut(A,B) &= \frac{1}{4}\left[\frac{(\alpha(x)+\gamma)+2(1-2k)\beta(x)}{k(1-k)M}\right] \\
&\rightarrow \frac{(\alpha(x)+\gamma)+2(1-2k)\beta(x)}{k(1-k)M} - \frac{2(\alpha(x)+\gamma)}{M} + \frac{2\alpha(x)}{M} + \underbrace{\frac{2\gamma}{M}}_{Constant} \\
&\rightarrow \frac{(\alpha(x)+\gamma)+2(1-2k)\beta(x)}{k(1-k)M} - \frac{2(\alpha(x)+\gamma)}{M} + \frac{2\alpha(x)}{M} \\
\therefore \quad &= \frac{(1-2k+k^2)(\alpha(x)+\gamma)+2(1-2k)\beta(x)}{k(1-k)M} + \frac{2\alpha(x)}{M} \\
&= \frac{\frac{(1-2k+2k^2)}{(1-k)^2}(\alpha(x)+\gamma)+\frac{2(1-2k)}{(1-k)^2}\beta(x)}{\frac{k}{1-k}M} + \frac{2\alpha(x)}{M} \\
&\text{Let } b = \frac{k}{1-k} \\
\therefore \quad &= \frac{(1+b^2)(\alpha(x)+\gamma)+2(1-b^2)\beta(x)}{bM} + \frac{2b\alpha(x)}{bM} \\
\because \quad &\frac{\gamma}{M} \text{ is a constant, and it doesn't affect the optimization.} \\
\therefore \quad &= \frac{(1+b^2)(\alpha(x)+\gamma)+2(1-b^2)\beta(x)}{bM} + \frac{2b\alpha(x)}{bM} - \frac{2b\gamma}{bM} \\
&= \frac{(1+b^2)(x^T(D-W)x+\mathbf{1}^T(D-W)\mathbf{1})}{b\mathbf{1}^T D\mathbf{1}} + \frac{2(1-b^2)\mathbf{1}^T(D-W)x}{b\mathbf{1}^T D\mathbf{1}} + \\
&\frac{2bx^T(D-W)x}{b\mathbf{1}^T D\mathbf{1}} - \frac{2b\mathbf{1}^T(D-W)\mathbf{1}}{b\mathbf{1}^T D\mathbf{1}} \\
\because \quad &(\mathbf{1}+x)^T(D-W)(\mathbf{1}+x) = \mathbf{1}^T(D-W)\mathbf{1} + 2(\mathbf{1}^T(D-W)x) + x^T(D-W)x \\
\because \quad &(\mathbf{1}-x)^T(D-W)(\mathbf{1}-x) = \mathbf{1}^T(D-W)\mathbf{1} - 2(\mathbf{1}^T(D-W)x) + x^T(D-W)x \\
\because \quad &(\mathbf{1}-x)^T(D-W)(\mathbf{1}+x) = \mathbf{1}^T(D-W)\mathbf{1} - x^T(D-W)x \\
\therefore \quad &= \frac{(\mathbf{1}+x)^T(D-W)(\mathbf{1}+x)}{b\mathbf{1}^T D\mathbf{1}} + \frac{b^2(\mathbf{1}-x)^T(D-W)(\mathbf{1}-x)}{b\mathbf{1}^T D\mathbf{1}} - \frac{2b(\mathbf{1}-x)^T(D-W)(\mathbf{1}+x)}{b\mathbf{1}^T D\mathbf{1}} \\
&\text{Look at } (\mathbf{1}-x)^T(D-W)(\mathbf{1}-x) = \mathbf{1}^T(D-W)\mathbf{1} - 2(\mathbf{1}^T(D-W)x) + x^T(D-W)x \\
\therefore \quad &(A-B)^T(D-W)(A-B) = A^T(D-W)A - 2(B^T(D-W)A) + B^T(D-W)B \\
&\text{Let } A = (\mathbf{1}+x), B = b(\mathbf{1}-x), b \text{ is a scalar.} \\
\therefore \quad &[(\mathbf{1}+x)-b(\mathbf{1}-x)]^T(D-W)[(\mathbf{1}+x)-b(\mathbf{1}-x)] \\
&= (\mathbf{1}+x)^T(D-W)(\mathbf{1}+x) - 2b(\mathbf{1}-x)^T(D-W)(\mathbf{1}+x) + b^2(\mathbf{1}-x)^T(D-W)(\mathbf{1}-x)
\end{aligned} \tag{17}
$$

$$\therefore\ \ NCut(A,B) = \frac{[(\mathbf{1}+x) - b(\mathbf{1}-x)]^T(D-W)[(\mathbf{1}+x) - b(\mathbf{1}-x)]}{b\mathbf{1}^T D\mathbf{1}}$$

$$\begin{aligned}
&\text{Setting } y = (\mathbf{1}+x) - b(\mathbf{1}-x), b = \frac{k}{1-k} = \frac{\sum_{x_i>0} d_i}{\sum_{x_i<0} d_i} \\
&\therefore\ y^T D\mathbf{1} = \frac{1}{2}\sum_{x_i>0} d_i - b\frac{1}{2}\sum_{x_i<0} d_i = 0 \\
&\therefore\ y^T Dy = (\mathbf{1}+x)^T D(\mathbf{1}+x) + b^2(\mathbf{1}-x)^T D(\mathbf{1}-x) - 2b(\mathbf{1}-x)^T D(\mathbf{1}+x) \\
&\because\ 2b(\mathbf{1}-x)^T D(\mathbf{1}+x) = 2b(\mathbf{1}-x)^T D(\mathbf{1}+x) \\
&\because\ (\mathbf{1}-x)^T D \text{ is the indicator vector for } x_i < 0 \\
&\because\ (\mathbf{1}+x) \text{ is the indicator vector for } x_i > 0 \\
&\therefore\ 2b(\mathbf{1}-x)^T D(\mathbf{1}+x) = 0 \\
&\therefore\ y^T Dy = (\mathbf{1}+x)^T D(\mathbf{1}+x) + b^2(\mathbf{1}-x)^T D(\mathbf{1}-x) \\
&\quad = \sum_{x_i>0} d_i + b^2 \sum_{x_i<0} d_i \\
&\quad = b\sum_{x_i<0} d_i + b^2 \sum_{x_i<0} d_i \\
&\quad = b(\sum_{x_i<0} d_i + b\sum_{x_i<0} d_i) \\
&\quad = b\mathbf{1}D\mathbf{1}
\end{aligned} \tag{18}$$

In the end, the loss of normalized cut is

$$NCut = \min_y \frac{y^T(D-W)y}{y^T Dy} \tag{19}$$

with the condition $y(i) \in \{1, -b\}$ and $y^T D\mathbf{1} = 0$. The laplacian matrix $L = D - W$ is a semi-positive definite matrix, therefore the eigenvalue $\lambda_i \geq 0$. Recall a fact about *Reyleigh quotient*: Let A be the real symmetric matrix. Under the constraint that x is orthogonal to the j-th smallest eigenvectors $x_1, x_2, \cdots, x_{j-1}$, the quotient $\frac{x^T Ax}{x^T x}$ is minimized by the next smallest eigenvector x_j and its minimum value is the corresponding eigenvalue λ_j. As a result, we obtain

$$z_1 = \arg\min_{z^T z=0} \frac{z^T D^{-\frac{1}{2}}(D-W)D^{-\frac{1}{2}}}{z^T z}$$

and, consequently,

$$y_1 = \arg\min_{y^T D\mathbf{1}=0} \frac{y^T(D-W)y}{y^T Dy}$$