Spectral Clustering

Chunpai Wang

March 2016

Spectral Clustering can identify groups based on the *similarity* between all pairs of data points[Von07]. Given a set of data points $x_1, ..., x_n$, and some notion of similarity $s_{ij} \geq 0$ between all pairs of data points x_i and x_j, a nice way of representing the data is in form of the similarity graph (matrix) $G = (V, E)$. Two vertices are connected if the similarity s_{ij} between the corresponding data points x_i and x_j is positive (or larger than a certain threshold), and the edge is weighted by s_{ij}.

1 Basic Notations

Let $G = (V, E)$ be an undirected graph with vertex set $V = \{v_1, ..., v_n\}$. We assume that the undirected graph is weighted with $w_{ij} = w_{ji} \geq 0$ for edge between vertices v_i and v_j. The weighted adjacency matrix of the graph is denoted as $W = (w_{ij})_{i,j=1,...,n}$.

- If $w_{ij} = 0$, this means that vertices v_i and v_j are not connected. The degree of a vertex $v_i \in V$ is defined as

$$d_i = \sum_{j=1}^{n} w_{ij}$$

- Note that, in fact, this sum only runs over all vertices adjacent to v_i, for which $w_{ij} > 0$ and $w_{ij} \neq 0$. Also, $w_{ii} = 0$ for all i. Hence, the diagonal of weighted matrix W are $0s$.
- The degree matrix D is defined as the diagonal matrix with the degrees $d_1, ..., d_n$ on the diagonal.
- Given a subset of vertices $A \subset V$, we denote its complement $V \backslash A$ by $\bar{A}$. We define the indicator vector $\mathbb{1}_A = (f_1, ..., f_n)' \in \mathbb{R}^n$ as the vector with entries $f_i = 1$ if $v_i \in A$ and $f_i = 0$ otherwise. For convenience, we introduce shorthand notation $i \in A$ for the set of indices $\{i | v_i \in A\}$.
- We consider two different ways of measuring the "size" of a subset $A \subset V$:

$$|A| := \text{the number of vertices in} A$$

$$vol(A) := \sum_{i \in A} d_i \quad \text{(total weight of its edges)}$$

- A subset $A \subset V$ of a graph is *connected* if any two vertices in A can be joined by a path such that all intermediate nodes also lie in A.

- We can call A a *connected component* if it is connected and if there are no connections between vertices in A and $\bar{A}$. The sets $A_1, ..., A_k$ form a partition of the graph if $A_i \cap A_j = \emptyset$ and $A_1 \cap ... \cap A_k = V$. Here, k is the number of connected components.

2 Different Similarity Graphs

First, for spectral clustering, we need to construct similarity matrix for all pair of data points. There are several popular constructions to transform a given vertex set with pairwise similarity or pairwise distance into a graph. Usually, we put a threshold on the similarity or distance.

- **The ϵ-neighborhood graph**, which we connect all points whose pairwise distances are smaller than ϵ.

- **k-nearest neighbor graphs**, which we connect the vertex v_i with vertex v_j if v_j is among the k nearest neighbors of v_i. However, this definition leads to a directed graph, as the k-nearest neighborhood relationship is not symmetric. First choice, ignore the direction. Second choice, only connect the v_i and v_j if both are k-nearest neighbors of each other.

- **The fully connected graph**, which we simply connect all points with positive similarity with each other. We may use the Gaussian similarity function

$$s(x_i, x_j) = exp(-\frac{||x_i - x_j||^2}{2\sigma^2})$$

  the parameter σ controls the width of the neighborhoods, similarly to the parameter ϵ in case of the ϵ-neighborhood graph.

3 Graph Laplacians and Their Basic Properties

After we construct the similarity matrix, we need to convert it to a graph Laplacian matrix, which is the main tool for spectral clustering, due to its many useful properties. However, there are several different Laplacian matrix, and we will discuss their properties respectively. When we talk about the eigenvector of a matrix here, the constant vector $\mathbb{1}$ and a multiple $a \cdot \mathbb{1}$ for some $a \neq 0$ are considered as the same eigen-vectors. Eigenvalues will always be ordered increasingly, respecting multiplicities. By "the first k eigenvectors", we refer to the eigen-vectors corresponding to the k-smallest eigenvalues.

3.1 The Unnormalized Graph Laplacian

The unnormalized graph Laplacian matrix is defined as

$$L = D - W$$

where D is the diagonal matrix and W is the weighted matrix.

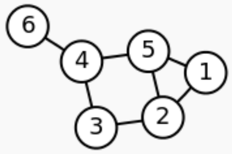
Labeled graph	Degree matrix	Adjacency matrix	Laplacian matrix
	$\begin{pmatrix} 2&0&0&0&0&0\\ 0&3&0&0&0&0\\ 0&0&2&0&0&0\\ 0&0&0&3&0&0\\ 0&0&0&0&3&0\\ 0&0&0&0&0&1 \end{pmatrix}$	$\begin{pmatrix} 0&1&0&0&1&0\\ 1&0&1&0&1&0\\ 0&1&0&1&0&0\\ 0&0&1&0&1&1\\ 1&1&0&1&0&0\\ 0&0&0&1&0&0 \end{pmatrix}$	$\begin{pmatrix} 2&-1&0&0&-1&0\\ -1&3&-1&0&-1&0\\ 0&-1&2&-1&0&0\\ 0&0&-1&3&-1&-1\\ -1&-1&0&-1&3&0\\ 0&0&0&-1&0&1 \end{pmatrix}$

Figure 1: The Unnormalized Laplacian Matrix

Proposition 1. *The Laplacian matrix L satisfies the following properties:*

1. *For every vector $f \in \mathbb{R}^n$ we have*

$$f^\top L f = \frac{1}{2}\sum_{i,j=1}^{n} w_{ij}(f_i - f_j)^2$$

2. *L is symmetric and positive semi-definite.*

3. *The smallest eigenvalue of L is 0, the corresponding eigenvector is the constant one vector $\mathbb{1}$.*

4. *L has n non-negative, real-valued eigenvalues $0 \le \lambda_1 \le \lambda_2 \le \ldots \le \lambda_n$.*

Proof. 1. By the definition of $d_i = \sum_{j=1}^n w_{ij}$,

$$\begin{aligned} f^\top L f &= f^\top D f - f^\top W f \\ &= \sum_{i=1}^{n} f_i d_i f_i - \sum_{i,j=1}^{n} f_i w_{ij} f_j \\ &= \frac{1}{2}\Big(\sum_{i=1}^{n} d_i f_i^2 - 2\sum_{i,j=1}^{n} f_i f_j w_{ij} + \sum_{j=1}^{n} d_j f_j^2\Big) \\ &= \frac{1}{2}\Big(\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij} f_i^2 - 2\sum_{i,j=1}^{n} f_i f_j w_{ij} + \sum_{j=1}^{n}\sum_{i=1}^{n} w_{ji} f_j^2\Big) \\ &= \frac{1}{2}\Big(\sum_{i=1,j=1}^{n} w_{ij} f_i^2 - 2\sum_{i,j=1}^{n} f_i f_j w_{ij} + \sum_{i=1,j=1}^{n} w_{ij} f_j^2\Big) \\ &= \frac{1}{2}\sum_{i,j=1}^{n} w_{ij}(f_i - f_j)^2 \end{aligned}$$

2. L is symmetric, because both D and W are symmetric. L is positive semi-definiteness, because $f^\top L f \ge 0$ for all $f \in \mathbb{R}^n$ according to (1.).

3. Since for all $f \in \mathbb{R}^n$, $f^\top L f \ge 0$, for all eigenvector v, $v^T L v = \lambda v^T v \ge 0$. Hence, all eigenvalues of L are non-negative, and smallest eigenvalue is 0. As $Lv = \lambda v$, if $\lambda = 0$, then $Lv = \lambda v = 0$, which means $v = \mathbb{1}$, because the sum of each row of L is the $d_i - \sum_{j=1}^n w_{ij} = 0$.

4. All $n \times n$ symmetric matrice have n eigenvalues. According to part 3. we have the smallest eigenvalue is 0, therefore, $0 = \lambda_1 \le \lambda_2 \le \lambda_3 \ldots \le \lambda_n$ $\square$

Proposition 2. (Number of Connected Components) *Let G be an undirected graph with non-negative weights. Then the multiplicity k of the eigenvalues 0 of L equals the number of connected components $A_1, ..., A_k$ in the graph. The eigenspace of eigenvalue 0 is spanned by the indicator vector $\mathbb{1}_{A_1}, ..., \mathbb{1}_{A_k}$ of those components.*

Proof. We start with the case $k = 1$, that is the graph is connected, and there is only one connected component. Assume that f is an eigenvector with eigenvalue 0, and we will show that this eigen-pair is an indicator of connected component. First, we know that

$$0 = f^\top L f = \sum_{i,j=1}^{n} w_{ij}(f_i - f_j)^2$$

As the weights w_{ij} are non-negative, this sum can only vanish if all terms $w_{ij}(f_i - f_j)^2$ vanish. Thus, if two vertices v_i and v_j are connected $(w_{ij} > 0)$, then f_i and f_j must be equal. With this argument, we can see that f needs to be constant for all vertices which can be connected by a path in the graph, such as v_k connected with v_i and v_j. Moreover, as all vertices of a connected component in an undirected graph can be connected by a path, f needs to be constant on the whole connected component. Therefore, we can conclude that in a graph consisting of only one connected component we only have the constant one vector $\mathbb{1}$ as eigenvector with eigenvalue 0, and this is obviously the indicator vector of the connected component.

If there are k connected components in a graph, we can find that L will be following with vertices are ordered according to the connected components they belong to. Note that, each of the

$$L = \begin{pmatrix} L_1 & & & \\ & L_2 & & \\ & & \ddots & \\ & & & L_k \end{pmatrix}$$

Figure 2: Laplacian Matrix with k Connected Component

blocks L_i is a proper graph Laplacian on its own. We know that the spectrum of L is given by the union of the spectra of L_i, and the corresponding eigenvectors of L are the eigenvectors of L_i, filled with 0 at the positions of the other blocks. As each L_i is a graph Laplacian of a connected graph, we know that every L_i has eigenvalue 0 with multiplicity 1, and the corresponding eigenvector is the constant one vector on the i^{th} connected component. Thus, the matrix L has as many eigenvalues 0 as there are connected component, and the corresponding eigenvectors are the indicator vectors of the connected components. $\square$

3.2 The Normalized Graph Laplacians

Two matrices are called normalized graph Laplacians

- $L_{sym} := D^{-1/2} L D^{-1/2} = I - D^{-1/2} W D^{-1/2}$
- $L_{rw} := D^{-1} L = I - D^{-1} W$
- L_{sym} is a symmetric matrix, and L_{rw} is closely related to a random walk.

Proposition 3. The normalized Laplacians satisfy the following properties:

1. For every $f \in \mathbb{R}^n$ we have

$$f^\top L_{sym} f = \frac{1}{2} \sum_{i,j=1}^{n} w_{ij} \Big(\frac{f_i}{\sqrt{d_i}} - \frac{f_j}{\sqrt{d_j}} \Big)^2$$

2. (λ, u) is an eigen pair of L_{rw} if and only if $(\lambda, D^{1/2} u)$ is an eigen pair of L_{sym}.
3. λ is an eigenvalue of L_{rw} with eigenvector u if and only if λ and u solve the generalized eigen-problem $Lu = \lambda D u$.
4. 0 is an eigenvalue of L_{rw} with the constant one vector $\mathbb{1}$ as eigenvector. 0 is an eigenvalue of L_{sym} with eigenvector $D^{1/2}\mathbb{1}$.
5. L_{sym} and L_{rw} are positive semi-definite and have n non-negative real-valued eigenvalues $0 = \lambda_1 \leq ... \leq \lambda_n$

Proof. 1. Since $d_i = \sum_{j=1} w_{ij}$, we have $1 = \frac{\sum_{j=1} w_{ij}}{d_i}$.

$$\begin{aligned}
f^\top L_{sym} f &= f^\top I f - f^\top D^{-1/2} W D^{-1/2} f \\
&= \sum_{i=1}^{n} f_i f_i - \sum_{i,j=1}^{n} \frac{f_i w_{ij} f_j}{\sqrt{d_i}\sqrt{d_j}} \\
&= \frac{1}{2} \Big(\sum_{i=1}^{n} f_i^2 - 2 \sum_{i,j=1}^{n} \frac{f_i w_{ij} f_j}{\sqrt{d_i}\sqrt{d_j}} + \sum_{j=1}^{n} f_j^2 \Big) \\
&= \frac{1}{2} \Big(\sum_{i=1}^{n} \sum_{j=1}^{n} \frac{w_{ij}}{d_i} f_i^2 - 2 \frac{f_i w_{ij} f_j}{\sqrt{d_i}\sqrt{d_j}} + \sum_{j=1}^{n} \sum_{i=1}^{n} \frac{w_{ji}}{d_j} f_j^2 \Big) \\
&= \frac{1}{2} \Big(\sum_{i=1,j=1}^{n} w_{ij} \big(\frac{f_i}{\sqrt{d_i}} \big)^2 - 2 \sum_{i=1,j=1}^{n} \frac{f_i w_{ij} f_j}{\sqrt{d_i}\sqrt{d_j}} + \sum_{i=1,j=1}^{n} w_{ij} \big(\frac{f_j}{\sqrt{d_j}} \big)^2 \Big) \\
&= \frac{1}{2} \sum_{i,j=1}^{n} w_{ij} \Big(\frac{f_i}{\sqrt{d_i}} - \frac{f_j}{\sqrt{d_j}} \Big)^2
\end{aligned}$$

2.

$$\begin{aligned}
L_{sym}D^{1/2}u &= \lambda(D^{1/2}u) \\
(D^{-1/2}LD^{-1/2})D^{1/2}u &= D^{1/2}\lambda u \\
D^{-1/2}(D^{-1/2}LD^{-1/2})D^{1/2}u &= \lambda u \\
D^{-1}LIu &= \lambda u \\
(D^{-1}L)u &= \lambda u \\
L_{rw}u &= \lambda u
\end{aligned}$$

3.

$$\begin{aligned}
L_{rw}u &= \lambda u \\
D^{-1}Lu &= \lambda u \\
DD^{-1}Lu &= D\lambda u \\
Lu &= \lambda Du
\end{aligned}$$

4.

$$\begin{aligned}
L_{rw}\mathbb{1} &= 0\mathbb{1} \\
D^{-1}L\mathbb{1} &= 0\mathbb{1} \\
D^{-1/2}D^{-1/2}L\mathbb{1} &= 0\mathbb{1} \\
D^{1/2}D^{-1/2}D^{-1/2}L\mathbb{1} &= 0D^{1/2}\mathbb{1} \\
ID^{-1/2}L\mathbb{1} &= 0D^{1/2}\mathbb{1} \\
D^{-1/2}LI\mathbb{1} &= 0D^{1/2}\mathbb{1} \\
D^{-1/2}LD^{-1/2}D^{1/2}\mathbb{1} &= 0D^{1/2}\mathbb{1} \\
L_{sym}D^{1/2}\mathbb{1} &= 0D^{1/2}\mathbb{1}
\end{aligned}$$

5. Both L_{sym} and L_{rw} are symmetric, therefore they have n eigenvalues. Since $f^\top L_{sym} f \geq 0$ for every vector f, L_{sym} is positive semi-definite.
Because for all eigenvector v, $v^\top L_{sym} v = \lambda v^\top v \geq 0$. Hence, all eigenvalues of L_{sym} are non-negative, and smallest eigenvalue is 0 and $0 \leq \lambda_1 \leq \lambda_2 \leq ... \leq \lambda_n$.
For L_{rw}, since $f^\top L_{rw} f = f^\top (D^{-1/2} L_{sym} D^{1/2}) f \geq 0$, L_{rw} is PSD. According to (2), L_{rw} has the same eigenvalues with L_{sym}. □

Proposition 4.(Number of Connected Components) *Let G be an undirected graph with non-negative weights. Then the multiplicity k of the eigenvalue 0 of both L_{rw} and L_{sym} equals the number of connected components $A_1, A_2, ..., A_k$ in the graph. For L_{rw}, the eigenspace of 0 is spanned by the indicator vectors $\mathbb{1}_{A_i}$ of those components. For L_{sym}, the eigenspace of 0 is spanned by the vectors $D^{1/2}\mathbb{1}_{A_i}$.*

Proof. The proof is analogous to the one of Proposition 2, using Proposition 3. □

4 Spectral Clustering Algorithm

Algorithm 1 Unnormalized Spectral Clustering

1: **Input:** Similarity matrix $S \in \mathbb{R}^{n \times n}$, number k of clusters to construct;
2: Construct a similarity graph by one of the ways described in section2. Let W be its weighted adjacency matrix;
3: Compute the unnormalized Laplacian L.
4: Compute the first k eigenvectors $v_1, ..., v_k$ of L.
5: Let $V \in \mathbb{R}^{n \times k}$ be the matrix containing the vectors $v_1, ..., v_k$ as columns.
6: For $i = 1, ..., n$, let $y_i \in \mathbb{R}^k$ be the vector corresponding to the i-th row of V.
7: Cluster the points $(y_i)_{i=1,...,n}$ in $\mathbb{R}^k$ with the k-means algorithms into clusters $C_1, ..., C_k$.
8: **Output:** Clusters $A_1, ..., A_k$ with $A_i = \{j | y_j \in C_i\}$;

Algorithm 2 Normalized Spectral Clustering with L_{rw}

1: **Input:** Similarity matrix $S \in \mathbb{R}^{n \times n}$, number k of clusters to construct;
2: Construct a similarity graph by one of the ways described in section2. Let W be its weighted adjacency matrix;
3: Compute the unnormalized Laplacian L.
4: Compute the first k eigenvectors $v_1, ..., v_k$ of the generalized eigenproblem $Lv = \lambda Dv$.
5: Let $V \in \mathbb{R}^{n \times k}$ be the matrix containing the vectors $v_1, ..., v_k$ as columns.
6: For $i = 1, ..., n$, let $y_i \in \mathbb{R}^k$ be the vector corresponding to the i-th row of V.
7: Cluster the points $(y_i)_{i=1,...,n}$ in $\mathbb{R}^k$ with the k-means algorithms into clusters $C_1, ..., C_k$.
8: **Output:** Clusters $A_1, ..., A_k$ with $A_i = \{j | y_j \in C_i\}$;

Algorithm 3 Normalized Spectral Clustering with L_{sym}

1: **Input:** Similarity matrix $S \in \mathbb{R}^{n \times n}$, number k of clusters to construct;
2: Construct a similarity graph by one of the ways described in section2. Let W be its weighted adjacency matrix;
3: Compute the normalized Laplacian L_{sym}.
4: Compute the first k eigenvectors $v_1, ..., v_k$ of L_{sym}.
5: Form the matrix $U \in \mathbb{R}^{n \times k}$ from V by normalizing the row sums to have norm 1, that is $u_{ij} = v_{ij} \ (\sum_k v_{ik}^2)^{1/2}$.
6: For $i = 1, ..., n$, let $y_i \in \mathbb{R}^k$ be the vector corresponding to the i-th row of V.
7: Cluster the points $(y_i)_{i=1,...,n}$ in $\mathbb{R}^k$ with the k-means algorithms into clusters $C_1, ..., C_k$.
8: **Output:** Clusters $A_1, ..., A_k$ with $A_i = \{j | y_j \in C_i\}$;

Github Code: https://github.com/Chunpai/SpectralClustering

5 Graph Cut Point of View

Spectral clustering can be derived as an approximation to graph partitioning problems. For two disjoint subsets $A, B \in V$ we define

$$cut(A, B) = \sum_{i \in A, j \in B} w_{ij}$$

Rather than solve the mincut problem, we also explicitly request that the partitions are all "reasonably large". The two most common objective functions which encode this are RatioCut and the normalized cut NCut.

$$RatioCut(A_1, ..., A_k) = \sum_{i=1}^{k} \frac{cut(A_i, \bar{A}_i)}{|A_i|}$$

$$NCut(A_1, ..., A_k) = \sum_{i=1}^{k} \frac{cut(A_i, \bar{A}_i)}{vol(A_i)}$$

5.1 Approximating RatioCut for $k = 2$

Formulation

When $k = 2$, we have the objective function

$$\min_{A \subset V} RatioCut(A, \bar{A}) = \min_{A \subset V}(\frac{cut(A, \bar{A})}{|A|} + \frac{cut(A, \bar{A})}{|\bar{A}|})$$

Problem Reformulation

Now, we define the vector $f = (f_1, ..., f_n)^\top \in \mathbb{R}^n$ with entries

$$f_i = \begin{cases} \sqrt{|\bar{A}|/|A|} & \text{if} v_i in A \\ -\sqrt{|A|/|\bar{A}|} & \text{if} v_i in \bar{A} \end{cases} \tag{1}$$

Then, we can rewrite the objective function using the unnormalized graph Laplacian,

$$\min_{A \subset V} f^\top L f \quad \text{subject to} \quad f \perp \mathbb{1}, \text{ and } \|f\| = \sqrt{n} \tag{2}$$

Note that, matrix L is graph Laplacian, which is used by spectral clustering. Refers to Proposition1.3, we can find that $\mathbb{1}$ is an eigenvector, and we would like to find a vector f subject to $f \perp \mathbb{1}$, which means f is an eigenvector as well.

Proof. Since if we define f as above, we can have

$$\begin{aligned}
f^\top Lf &= \sum_{i,j=1}^{n} w_{ij}(f_i - f_j)^2 \\
&= \sum_{i\in A, j\in \bar{A}} w_{ij}\big(\sqrt{\frac{|\bar{A}|}{|A|}} + \sqrt{\frac{|A|}{|\bar{A}|}}\big)^2 + \sum_{i\in \bar{A}, j\in A} w_{ij}\big(-\sqrt{\frac{|\bar{A}|}{|A|}} - \sqrt{\frac{|A|}{|\bar{A}|}}\big)^2 \\
&= \sum_{i\in A, j\in \bar{A}} w_{ij}\big(\frac{|\bar{A}|}{|A|} + 2 + \frac{|A|}{|\bar{A}|}\big) + \sum_{i\in \bar{A}, j\in A} w_{ij}\big(\big) \\
&= \big(\frac{|\bar{A}|}{|A|} + 2 + \frac{|A|}{|\bar{A}|}\big)\big(\sum_{i\in A, j\in \bar{A}} w_{ij} + \sum_{i\in \bar{A}, j\in A} w_{ij}\big) \\
&= \big(\frac{|\bar{A}|}{|A|} + 2 + \frac{|A|}{|\bar{A}|}\big) \cdot 2 \cdot cut(A, \bar{A}) \\
&= 2 \cdot cut(A, \bar{A})\big(\frac{|\bar{A}|}{|A|} + 1 + 1 + \frac{|A|}{|\bar{A}|}\big) \\
&= 2 \cdot cut(A, \bar{A})\big(\frac{|\bar{A}| + |A|}{|A|} + \frac{|A| + |\bar{A}|}{|\bar{A}|}\big) \\
&= 2 \cdot cut(A, \bar{A})\big(\frac{|V|}{|A|} + \frac{|V|}{|\bar{A}|}\big) \\
&= 2 \cdot |V|\big(\frac{cut(A, \bar{A})}{|A|} + \frac{cut(A, \bar{A})}{|\bar{A}|}\big) \\
&= 2 \cdot |V| \cdot RatioCut(A, \bar{A})
\end{aligned}$$

Additionally, we have

$$f^T\mathbb{1} = \sum_{i=1}^{n} f_i = \sum_{i\in A}\sqrt{\frac{|\bar{A}|}{|A|}} - \sum_{i\in \bar{A}}\sqrt{\frac{|A|}{|\bar{A}|}} = |A|\sqrt{\frac{|\bar{A}|}{|A|}} - |\bar{A}|\sqrt{\frac{|A|}{|\bar{A}|}} = 0 \tag{3}$$

In other words, the vector f as defined above is orthogonal to the constant one vector $\mathbb{1}$. Finally, f also satisfies

$$\|f\|^2 = \sum_{i=1}^{n} f_i^2 = |A|\frac{|\bar{A}|}{|A|} + |\bar{A}|\frac{|A|}{|\bar{A}|} = |V| = n \tag{4}$$

□

Problem Relaxation

However, the new formulation is still a NP-hard discrete optimization problem as the entries of the solution vector f are only allowed to take two particular values. Therefore, we would like to find a relaxation of problem, such as discarding the condition on the discrete values for f_i and instead allow $f_i \in \mathbb{R}$. This leads to the relaxed optimization problem

$$\min_{f\in\mathbb{R}^n} f^\top Lf \ \text{ subject to } \ f \perp \mathbb{1}, \text{ and } \|f\| = \sqrt{n} \tag{5}$$

By the Rayleigh-Ritz theorem and the Proposition1.3, it can be be seen immediately that the solution of this problem is given by the vector f which is the eigenvector corresponding to the second smallest eigenvalue of L. So we can approximate a minimizer of RatioCut by the second eigenvector of L.

K-Means

In order to obtain a partition of the graph we need to re-transform the real-valued solution vector f of the relaxed problem into a discrete indicator vector. The simplest way to do this is to use the sign of f as indicator function, that is to choose

$$\begin{cases} v_i \in A & \text{if } \ f_i \geq 0 \\ v_i \in \bar{A} & \text{if } \ if_i < 0 \end{cases} \tag{6}$$

However, when $k > 2$, this heuristic is too simple. What most spectral clustering algorithms do instead is to consider the coordinates f_i as points in $\mathbb{R}$ and cluster them into two groups $C, \bar{C}$ by the k-means clustering algorithm.Then we carry over the resulting clustering to the underlying data points, that is we choose

$$\begin{cases} v_i \in A & \text{if } \ f_i \in C \\ v_i \in \bar{A} & \text{if } \ f_i \in \bar{C} \end{cases} \tag{7}$$

This is exactly the unnormalized spectral clustering algorithm for the case of $k = 2$. However, there is nothing principled about using k-means algorithm to construct discrete partitions from the real-valued representation vector y_i. Any other algorithm which can solve this problem could be used instead.

5.2 Approximating NCut for $k = 2$

Formulation

When $k = 2$, we have the objective function

$$\min_{A \subset V} NCut(A, \bar{A}) = \min_{A \subset V} \left(\frac{cut(A, \bar{A})}{vol(A)} + \frac{cut(A, \bar{A})}{vol(\bar{A})} \right) \tag{8}$$

where

$$vol(A) := \sum_{i \in A} d_i; \quad vol(\bar{A}) = \sum_{i \in \bar{A}} d_i \tag{9}$$

Problem Reformulation

We define the cluster indicator vector f by

$$f_i = \begin{cases} \sqrt{\frac{vol(\bar{A})}{vol(A)}} & \text{if } \ i \in A \\ -\sqrt{\frac{vol(A)}{vol(\bar{A})}} & \text{if } \ i \in \bar{A} \end{cases} \tag{10}$$

Then, we can rewrite the objective function using the unnormalized graph Laplacian,

$$\min_{A \subset V} f^\top L f \ \text{ subject to } \ Df \perp \mathbb{1}, \text{ and } f \top Df = vol(V) \tag{11}$$

Proof.

$$
\begin{aligned}
Df &= \sum_{i=1}^{n} d_i f_i \\
&= \sum_{i \in A} d_i \sqrt{\frac{vol(\bar{A})}{vol(A)}} - \sum_{i \in \bar{A}} d_i \sqrt{\frac{vol(A)}{vol(\bar{A})}} \\
&= \sum_{i \in A} d_i \sqrt{\frac{\sum_{i \in \bar{A}} d_i}{\sum_{i \in A} d_i}} - \sum_{i \in \bar{A}} d_i \sqrt{\frac{\sum_{i \in A} d_i}{\sum_{i \in \bar{A}} d_i}} \\
&= \sqrt{\sum_{i \in \bar{A}} d_i * \sum_{i \in A} d_i} - \sqrt{\sum_{i \in A} d_i * \sum_{i \in \bar{A}} d_i} \\
&= 0
\end{aligned}
$$

Therefore, the vector Df is orthogonal to the constant one vector $\mathbb{1}$.

$$
\begin{aligned}
f^T Df &= \sum_{i=1}^{n} f_i d_i f_i \\
&= \sum_{i \in A} \sqrt{\frac{vol(\bar{A})}{vol(A)}} * d_i * \sqrt{\frac{vol(\bar{A})}{vol(A)}} + \sum_{i \in \bar{A}} \sqrt{\frac{vol(A)}{vol(\bar{A})}} * d_i * \sqrt{\frac{vol(A)}{vol(\bar{A})}} \\
&= \sum_{i \in A} d_i \frac{vol(\bar{A})}{vol(A)} + \sum_{i \in \bar{A}} d_i \frac{vol(A)}{vol(\bar{A})} \\
&= \sum_{i \in A} d_i \frac{\sum_{i \in \bar{A}} d_i}{\sum_{i \in A} d_i} + \sum_{i \in \bar{A}} d_i \frac{\sum_{i \in A} d_i}{\sum_{i \in \bar{A}} d_i} \\
&= \sum_{i \in \bar{A}} d_i + \sum_{i \in A} d_i \\
&= \sum_{i=1}^{n} d_i \\
&= vol(V)
\end{aligned}
$$

$$
\begin{aligned}
f^T L f &= \frac{1}{2}\sum_{i,j=1}^{n} w_{ij}(f_i - f_j)^2 \\
&= \frac{1}{2}\sum_{i\in A, j\in \bar{A}} w_{ij}\Big(\sqrt{\frac{vol(\bar{A})}{vol(A)}} + \sqrt{\frac{vol(A)}{vol(\bar{A})}}\Big)^2 + \frac{1}{2}\sum_{i\in \bar{A}, j\in A} w_{ij}\Big(-\sqrt{\frac{vol(\bar{A})}{vol(A)}} - \sqrt{\frac{vol(A)}{vol(\bar{A})}}\Big)^2 \\
&= \frac{1}{2}\sum_{i\in A, j\in \bar{A}} w_{ij}\Big(\frac{vol(\bar{A})}{vol(A)} + \frac{vol(A)}{vol(\bar{A})} + 2\Big) + \frac{1}{2}\sum_{i\in \bar{A}, j\in A} w_{ij}\Big(\frac{vol(\bar{A})}{vol(A)} + \frac{vol(A)}{vol(\bar{A})} + 2\Big) \\
&= \sum_{i\in A, j\in \bar{A}} w_{ij}\Big(\frac{vol(\bar{A})}{vol(A)} + \frac{vol(A)}{vol(\bar{A})} + 2\Big) \\
&= cut(A,\bar{A})\Big(\frac{vol(\bar{A})}{vol(A)} + \frac{vol(A)}{vol(\bar{A})} + 2\Big) \\
&= cut(A,\bar{A})\Big(\frac{vol(\bar{A})}{vol(A)} + \frac{vol(A)}{vol(A)} + \frac{vol(A)}{vol(\bar{A})} + \frac{vol(\bar{A})}{vol(\bar{A})}\Big) \\
&= cut(A,\bar{A})\Big(\frac{vol(\bar{A}) + vol(A)}{vol(A)} + \frac{vol(A) + vol(\bar{A})}{vol(\bar{A})}\Big) \\
&= vol(V)\Big(\frac{cut(A,\bar{A})}{vol(A)} + \frac{cut(A,\bar{A})}{vol(\bar{A})}\Big) \\
&= vol(V) Ncut(A,\bar{A})
\end{aligned}
$$

$\square$

Problem Relaxation

Again we relax the problem by allowing f to be real valued:

$$\min_{f\in\mathbb{R}^n} f^\top L f \;\; \text{subject to} \;\; Df \perp \mathbb{1}, \text{ and } f \top D f = vol(V) \tag{12}$$

Now we substitute $g := D^{1/2} f$, and the problem becomes

$$\min_{g\in\mathbb{R}^n} g^\top D^{-1/2} L D^{-1/2} g \;\; \text{subject to} \;\; g \perp D^{1/2}\mathbb{1}, \text{ and } \|g\|^2 = vol(V) \tag{13}$$

Observe that $D^{-1/2} L D^{-1/2} = L_{sym}$, and $D^{1/2}\mathbb{1}$ is the first eigenvector of L_{sym}, and $vol(V)$ is a constant. Hence, the problem is in the form of the standard Rayleigh-Ritz theorem, and its solution g is given by the second eigenvector of L_{sym}. Re-substituting $f = D^{-1/2} g$ and using Proposition 3, we can see that f is the second eigenvector of L_{rw}, or equivalently the generalized eigenvector of $Lv = \lambda D v$.

5.3 Comments On The Relaxation Approach

Of course, the relaxation we discussed above is not unique. For example, a completely different relaxation which leads to a semi-definite program is derived in Bie and Cristianini (2006), and there might be many other useful relaxations. The reason why the spectral relaxation is so appealing is not that it leads to particularly good solutions, but is the fact that it results in a very simple to solve standard linear algebra problem.

References

[Von07] Ulrike Von Luxburg. "A tutorial on spectral clustering". In: *Statistics and computing* 17.4 (2007), pp. 395–416.