Spectral Clustering - MAT180

Prepared by Shuyang Ling

May 6, 2017

1 Spectral clustering

Spectral clustering is a graph-based method which uses the eigenvectors of the graph Laplacian derived from the given data to partition the data. It outperforms K-means since it can capture the geometry of data.

The spectral clustering algorithm takes two steps in general:

1. Construct an undirected weighted graph based on the data. Given $\{x_i\}_{i=1}^n$ with pairwise affinities w_{ij}, we obtain the unnormalized Laplacian $L = D - W$ (or the normalized graph Laplacian $L_S = I - D^{-1/2}WD^{-1/2}$). Then we compute the eigenvectors $\phi_1, \cdots, \phi_k$, corresponding to the first smallest k eigenvectors of L, i.e.,

$$L\phi_l = \lambda_l \phi_l,$$

where $\{\phi_l\}_{l=1}^n$ is an orthonormal basis in $\mathbb{R}^n$ and the eigenvalues are

$$\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n.$$

Let $\Phi \in \mathbb{R}^{n\times k}$ be a matrix consisting of $\{\phi_l\}_{l=1}^k$,

$$\Phi(x_i) = (\phi_1(x_i), \cdots, \phi_k(x_i))^T \in \mathbb{R}^k.$$

In other words, we transform the original data x_i from $\mathbb{R}^n$ to $\mathbb{R}^k$ through the eigenvectors of L.

2. Apply K-means clustering to Φ to group the data into k clusters.

One thing we need to discuss here is the choice of the affinities w_{ij}. There are several common ways to choose w_{ij} (the weight on the edge between node x_i and x_j). The rule of thumb is: the larger w_{ij} means more similarities/associations between x_i and x_j. Here are several useful examples:

- The ϵ-neighborhood graph:

$$w_{ij} = \begin{cases} 1, & \|x_i - x_j\| \leq \epsilon \\ 0, & \text{otherwise.} \end{cases}$$

If x_i and x_j are close to each other, then $w_{ij} = 1$; otherwise, $w_{ij} = 0$.

- The fully connected graph: $w_{ij} = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right)$ where σ is the width of the neighborhoods (just like ϵ in the previous example). If x_i is close to x_j, then w_{ij} is close to 1. On the other hand, if x_i is far away from x_j (what "far away" means depends strongly on the choice of σ), then w_{ij} would be close to 0.

1.1 Applying spectral clustering to a disconnected graph

To warm up, we apply the spectral clustering to a disconnected graph. Suppose we have an undirected graph with weight w_{ij} and k connected components $S_1, \cdots, S_k$.

Based on the spectral clustering, we first construct its graph Laplacian. After rearranging data points, the graph Laplacian L can be written as a block-diagonal matrix,

$$L = \begin{bmatrix} L_1 & 0 & \cdots & 0 \\ 0 & L_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & L_k \end{bmatrix}$$

where each $L_i \in \mathbb{R}^{|S_i| \times |S_i|}$ is the graph Laplacian for each connected component.

Since $L_i \mathbf{1}_{|S_i|} = 0$ for all $1 \leq i \leq k$ where $\mathbf{1}_{|S_i|}$ is a $|S_i| \times 1$ vector with all entries equal to 1. Therefore, the k smallest eigenvalues of L equal 0. Moreover, the null space of this matrix is spanned by k indicator functions such as

$$\phi_1 = \frac{1}{\sqrt{|S_1|}} \begin{bmatrix} \mathbf{1}_{|S_1|} \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \quad \cdots\cdots, \quad \phi_k = \frac{1}{\sqrt{|S_k|}} \begin{bmatrix} 0 \\ 0 \\ \vdots \\ \mathbf{1}_{|S_k|} \end{bmatrix}.$$

This representation is unique up to an orthogonal transformation.

Hence the Φ is in the form of

$$\Phi = \begin{bmatrix} \frac{1}{\sqrt{|S_1|}} \mathbf{1}_{|S_1|} & 0 & \cdots & 0 \\ 0 & \frac{1}{\sqrt{|S_2|}} \mathbf{1}_{|S_2|} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{\sqrt{|S_k|}} \mathbf{1}_{|S_k|} \end{bmatrix} \in \mathbb{R}^{n \times k}.$$

Each row of Φ is a vector with only one nonzero entry and it is equal to $\frac{1}{\sqrt{S_i}}$. In other words, this spectral mapping transforms all points in the i-th connected component S_i into a single point $\frac{1}{\sqrt{|S_i|}} e_i$ where $\{e_i\}_{i=1}^k$ is the canonical basis in $\mathbb{R}^k$.

1.2 Graph cut point of view on spectral clustering

The intuition of clustering is to separate points in different groups according to their similarities. For data given in an undirected graph with weight W, we want to find a partition such that the connection between each group is small. Such a partition corresponds to a cut on the graph. In particular, as discussed before, for a disconnected graph with weight W, we can find partition such that the connection between two components is zero. Spectral clustering can be explained and derived from the notion of a cut on graph.

First we define what a cut means here for two sets. Suppose we have two disjoint sets of vertices on the graph S and $\overline{S}$ s.t. $S \cap \overline{S} = \varnothing$ and $S \cup \overline{S} = V$. We define a cut w.r.t. $S \subseteq V$ as

$$\text{cut}(S) = \sum_{i \in S, j \in \overline{S}} w_{ij} = \text{cut}(\overline{S}),$$

is the total summation of the edge weights whose two vertices are in different sets. In particular, $\text{cut}(V) = 0$.

Here, $\text{cut}(S)$ measures how much S and $\overline{S}$ are associated/connected. A smaller value for $\text{cut}(S)$ means fewer connections between S and $\overline{S}$.

But would minimizing $\text{cut}(S)$ over all $S \subseteq V$ give us a good partition of the graph? It does not always work. For example, the cut can return unwanted and meaningless results: declaring a single vertex as a cluster. In other words, minimizing $\text{cut}(S)$ is likely to give us unbalanced clusters/partitions. Therefore, we define the ratio cut and normalized cut:

$$\text{Rcut}(S) = \frac{\text{cut}(S)}{|S|} + \frac{\text{cut}(\overline{S})}{|\overline{S}|}, \tag{1}$$

$$\text{Ncut}(S) = \frac{\text{cut}(S)}{\text{vol}(S)} + \frac{\text{cut}(\overline{S})}{\text{vol}(\overline{S})} \tag{2}$$

where $\text{vol}(S)$ is the volume of S and is defined as $\text{vol}(S) = \sum_{i \in S} d_i = \sum_{i \in S} \sum_{j \in V} w_{ij}$.

Now let us consider the strength and weakness of $\text{Rcut}(S)$ and $\text{Ncut}(S)$:

1. Pros: given the same value of $\text{cut}(S)$, we will have smaller $\text{Rcut}(S)$ and $\text{Ncut}(S)$ if S and $\overline{S}$ are more or less of the same size. Due to this property, we can avoid unbalanced clusters by using $\text{Rcut}(S)$ and $\text{Ncut}(S)$ as new criteria.

2. Cons: both two criteria are difficult to minimize, i.e.,

$$\min_{S \subset V} \text{Rcut}(S) \tag{3}$$

$$\min_{S \subset V} \text{Ncut}(S) \tag{4}$$

   are NP-hard in general since they are discrete/combinatorial optimization. Using brute force requires us to search over $2^{|V|}$ (the total number of subsets in V) choices.

1.3 Relaxation of the ratio cut

We find it difficult to minimize this function over all subsets of S. However, it does not mean that we cannot approximate (1) and (2) by other optimization programs which are easier to solve. We hope the solutions to the alternative programs are able to give us good approximations of the original ones.

In this section, we will show that the ratio cut (1) can be approximated by a continuous optimization program (eigenvalue/vector problem of L) which exactly matches the first step of spectral clustering. We start with relating the ratio cut to the quadratic form of L.

Let us consider the following function $f_S : V \to \mathbb{R}$ (depending on the set S) as

$$f_S(i) = \begin{cases} \sqrt{\frac{|\overline{S}|}{|V||S|}}, & i \in S, \\ -\sqrt{\frac{|S|}{|V||\overline{S}|}}, & i \in \overline{S}. \end{cases} \tag{5}$$

It is a simple step function over V and in particular, $f_S(i) = -f_{\overline{S}}(i)$.

Lemma 1. *There holds*

$$f_S^T L f_S = \text{Rcut}(S).$$

for all f_S defined in (5).

Proof: The proof is quite straightforward,

$$\begin{aligned}
f_S^T L f_S &= \frac{1}{2}\sum_{i\in V, j\in V} w_{ij}(f_S(i)-f_S(j))^2 \\
&= \frac{1}{2}\left(\sum_{i\in S, j\in \overline{S}} + \sum_{i\in \overline{S}, j\in S}\right) w_{ij}(f_S(i)-f_S(j))^2 && \text{by (5)} \\
&= \sum_{i\in S, j\in \overline{S}} w_{ij}(f_S(i)-f_S(j))^2 && \text{by symmetry of } i,j \\
&= \sum_{i\in S, j\in \overline{S}} w_{ij}\left(\sqrt{\frac{|\overline{S}|}{|V||S|}} + \sqrt{\frac{|S|}{|V||\overline{S}|}}\right)^2 && \text{by (5)} \\
&= \frac{1}{|V|}\sum_{i\in S, j\in \overline{S}} w_{ij}\left(\frac{|\overline{S}|}{|S|} + \frac{|S|}{|\overline{S}|} + 2\right) \\
&= \sum_{i\in S, j\in \overline{S}} w_{ij}\left(\frac{1}{|S|} + \frac{1}{|\overline{S}|}\right) = \mathrm{Rcut}(S) && \text{by } |V| = |S| + |\overline{S}|
\end{aligned}$$

□

So far, we have not yet made (3) any easier since we just replaced the expression in (3) with $\min_{f_S} f_S^T L f_S$. Now the idea of the next step is to enlarge the constraint set by investigating what properties all $\{f_S\}_{S\subseteq}$ satisfy.

There are two common properties about f_S which hold for all S:

Lemma 2. *For all $S \subseteq V$ and f_S defined in* (5)*, there hold*

$$\|f_S\| = 1, \quad f_S^T \mathbf{1} = 0 \tag{6}$$

Proof: By definition of f_S in (5),

$$\|f_S\|^2 = \sum_{i\in V} |f_S(i)|^2 = \sum_{i\in S}\frac{|\overline{S}|}{|V||S|} + \sum_{i\in \overline{S}}\frac{|S|}{|V||\overline{S}|} = \frac{|\overline{S}|}{|V||S|}|S| + \frac{|S|}{|V||\overline{S}|}|\overline{S}| = 1.$$

For $f_S^T\mathbf{1}$, we have

$$f_S^T\mathbf{1} = \sum_{i\in V} f_S(i) = \sum_{i\in S} f_S(i) - \sum_{i\in \overline{S}} f_S(i) = \sqrt{\frac{|S||\overline{S}|}{|V|}} - \sqrt{\frac{|S||\overline{S}|}{|V|}} = 0.$$

□

We replace the original program (3) by

$$\min f^T L f, \quad s.t. \quad \|f\|^2 = 1, f^T\mathbf{1} = 0. \tag{7}$$

Note that $\{f_S\}_{S\subseteq V} \subseteq \{f : \|f\| = 1, f^T\mathbf{1} = 0\}$; in fact, (7) minimizes $f^T L f$ over a larger set than $\{f_S : S \subseteq V\}$, and we hope to get an approximate solution from this alternative approach. Moreover, (7) is much easier to solve since it directly corresponds to the eigenvalue/vector problem of L, which matches exactly the first step of the spectral clustering. We call (7) the *relaxation* of (3).

Lemma 3. *The minimizer of* (7) *is* ϕ_2*, the eigenvector of the Laplacian* L *w.r.t. the second smallest eigenvalue* λ_2*. In other words, the minimum of* (7) *is* λ_2*.*

Proof: Note that $\mathbf{1}$ is inside the null space of L. For any f satisfying $\|f\| = 1$ and $f^T\mathbf{1} = 0$, we can rewrite f as $f = \sum_{l=1}^n \alpha_l \phi_l$ where $\{\phi_l\}_{l=1}^n$ is an orthonormal basis in $\mathbb{R}^n$ and $\alpha_l = f^T\phi_l$. Moreover, $\alpha_1 = 0$ follows from $f^T\mathbf{1} = 0$ and $\sum_{l=2}^n \alpha_l^2 = 1$. Therefore, $f = \sum_{l=2}^n \alpha_l\phi_l$ for all f satisfying $\|f\| = 1$ and $f^T\mathbf{1} = 0$.

$$f^T L f = \sum_{l=2}^n \lambda_l \alpha_l^2 \geq \lambda_2 \sum_{l=2}^n \alpha_l^2 = \lambda_2.$$

In particular, if $\alpha_2 = 1$ and $\alpha_l = 0$ for all $l \geq 3$, i.e., $f = \phi_2$, the minimum is attained. □

Compared with (3), we know that $\{f_S\}_{S\subseteq V} \subseteq \{f : \|f\| = 1, f^T\mathbf{1} = 0\}$ and hence

$$\mathrm{Rcut}(S) \geq \lambda_2, \quad \min_{S\subseteq V} \mathrm{Rcut}(S) \geq \lambda_2.$$

How to get clustering (or get f_S over $2^{|V|}$ choices) based on ϕ_2? We can just run k-means, which is equivalent to setting a threshold r such that

$$\phi_2(i) > r \Longleftrightarrow i \in S, \quad \phi_2(i) \leq r \Longleftrightarrow i \in \overline{S}.$$

This is also called "rounding" procedure.

1.4 Relaxation of the normalized cut

Like ratio cut (1), normalized cut (2) also has a relaxation form. Let $f_S(i)$ be a function from $V \to \mathbb{R}$,

$$f_S(i) := \begin{cases} \sqrt{\frac{\mathrm{vol}(\overline{S})}{\mathrm{vol}(V)\,\mathrm{vol}(S)}}, & i \in S, \\ -\sqrt{\frac{\mathrm{vol}(S)}{\mathrm{vol}(V)\,\mathrm{vol}(\overline{S})}}, & i \in \overline{S}. \end{cases} \tag{8}$$

It is easy to see that (8) can be obtained by replacing $|S|$, $|\overline{S}|$ and $|V|$ in (5) by $\mathrm{vol}(S)$, $\mathrm{vol}\,\overline{S}$ and $\mathrm{vol}(V)$ respectively.

Lemma 4. *For* f_S *defined in* (8)*, there holds*

$$f_S^T L f_S = \mathrm{Ncut}(S). \tag{9}$$

Proof: The proof is similar to what we have done in the previous section.

$$\begin{aligned} f_S^T L f_S &= \frac{1}{2} \sum_{i\in V, j\in V} w_{ij}(f_S(i) - f_S(j))^2 \\ &= \sum_{i\in S, j\in \overline{S}} w_{ij}(f_S(i) - f_S(j))^2 \\ &= \sum_{i\in S, j\in \overline{S}} w_{ij} \left(\sqrt{\frac{\mathrm{vol}(\overline{S})}{\mathrm{vol}(V)\,\mathrm{vol}(S)}} + \sqrt{\frac{\mathrm{vol}(S)}{\mathrm{vol}(V)\,\mathrm{vol}(\overline{S})}} \right)^2 \\ &= \frac{1}{\mathrm{vol}(V)} \left(\frac{\mathrm{vol}(\overline{S})}{\mathrm{vol}\,(S)} + \frac{\mathrm{vol}(S)}{\mathrm{vol}(\overline{S})} + 2 \right) \sum_{i\in S, j\in \overline{S}} w_{ij} \\ &= \left(\frac{1}{\mathrm{vol}(S)} + \frac{1}{\mathrm{vol}(\overline{S})} \right) \mathrm{cut}(S) = \mathrm{Ncut}(S). \end{aligned}$$

□

All $\{f_S\}$ in the form of (8) share two common properties.

Lemma 5. *For all $S \subseteq V$ and f_S defined in (8), there hold*

$$f_S^T D f_S = 1, \quad f_S^T D\mathbf{1} = 0. \tag{10}$$

Proof: For $f_S^T D f_S$,

$$\begin{aligned} f_S^T D f_S &= \sum_{i\in V} d_i f_S^2(i) = \frac{\mathrm{vol}(\overline{S})}{\mathrm{vol}(V)\,\mathrm{vol}(S)} \sum_{i\in S} d_i + \frac{\mathrm{vol}(S)}{\mathrm{vol}(V)\,\mathrm{vol}(\overline{S})} \sum_{i\in \overline{S}} d_i \\ &= \frac{\mathrm{vol}(\overline{S}) + \mathrm{vol}(S)}{\mathrm{vol}(V)} = 1. \end{aligned}$$

For $f_S^T D\mathbf{1}$, we have

$$\begin{aligned} f_S^T D\mathbf{1} &= \sum_{i\in V} f_S(i) d_i = \sum_{i\in S} f_S(i) d_i + \sum_{i\in \overline{S}} f_S(i) d_i \\ &= \sqrt{\frac{\mathrm{vol}(\overline{S})}{\mathrm{vol}(V)\,\mathrm{vol}(S)}} \sum_{i\in S} d_i - \sqrt{\frac{\mathrm{vol}(S)}{\mathrm{vol}(V)\,\mathrm{vol}(\overline{S})}} \sum_{i\in \overline{S}} d_i \\ &= \sqrt{\frac{\mathrm{vol}(\overline{S})}{\mathrm{vol}(V)\,\mathrm{vol}(S)}}\, \mathrm{vol}(S) - \sqrt{\frac{\mathrm{vol}(S)}{\mathrm{vol}(V)\,\mathrm{vol}(\overline{S})}}\, \mathrm{vol}(\overline{S}) = 0. \end{aligned}$$

□

Just like the relaxation (7) for ratio cut (3), we propose the following program as a relaxation for (4).

$$\min f^T L f, \quad s.t. \quad f^T D f = 1, f^T D\mathbf{1} = 0. \tag{11}$$

Actually (11) is equivalent to the eigenvalue/vector problem of L_S, the normalized graph Laplacian $L_S = I - D^{-1/2} W D^{-1/2} = D^{-1/2} L D^{-1/2}$. Letting $x = D^{1/2} f$ and substitute it into (11), we get

$$\min x^T L_S x, \quad s.t. \quad \|x\| = 1, x^T D^{1/2}\mathbf{1} = 0. \tag{12}$$

Remember $D^{1/2}\mathbf{1}$ satisfies $L_S D^{1/2}\mathbf{1} = D^{-1/2} L\mathbf{1} = 0$. This means $D^{1/2}\mathbf{1}$ is in the null space of L_S. Therefore, the minimizer of this program is actually ϕ_2, the eigenvector corresponding to the second smallest eigenvalue of L_S. The proof can be easily adapted from the proof of Lemma 3.

How to get clustering (or get f_S over $2^{|V|}$ choices) based on the ϕ_2? We still can apply k-means to ϕ_2, which is equivalent to setting a threshold r such that

$$\phi_2(i) > r \Longleftrightarrow i \in S, \quad \phi_2(i) \le r \Longleftrightarrow i \in \overline{S}.$$

1.5 Comparison

In conclusion,

- The relaxation of the ratio cut = the eigenvalue and eigenvector problem of $L = D - W$.

- The relaxation of the normalized cut = the eigenvalue and eigenvector problem of $L_S = I - D^{-1/2}WD^{-1/2}$.
- Moreover, all those results can be easily extended to the multi-cluster scenario.

Which algorithm is better:

- Empirically, Ncut is a better choice. If the graph is regular (all d_i are the same), they are same.
- The goal of clustering is not just to minimize the cut, but also to maximize the in-cluster association.

$$\text{assoc}(S, S) = \sum_{i \in S, j \in S} w_{ij} = \sum_{i \in S, j \in V} w_{ij} - \sum_{i \in S, j \in \overline{S}} = \text{vol}(S) - \text{cut}(S).$$

  Therefore, we want to minimize $\text{cut}(S)$ and maximize $\text{vol}(S)$, which is exactly what normalized cut does.

This note is based on Amit Singer's notes and Ulrike von Luxburg's tutorial on spectral clustering.