Diffusion Maps and Spectral Clustering

Uri Shaham

1 Spectral Clustering

1.1 Graph Laplacians

Spectral clustering is a clustering algorithm which is far more expressive than k-means, and is not limited to convex clusters.

Definition 1.1 (Undirected weighted graph)**.** *A weighted graph is an undirected graph $G = (V, E)$ where $V = \{v_1, \ldots, v_n\}$ and each to vertices v_i, v_j are connected with an edge with non-negative weight $w_{ij} \geq 0$. We denote the weight matrix as W.*

Remark 1.2. *When data lies in Euclidean space and the weights are not given, it is a common practice to use Gaussian kernel to compute the weights*

$$w_{ij} = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right),$$

for such bandwidth parameter σ.

Remark 1.3. *A common practice is to connect each point only to its k nearest neighbors, which is known as knn graph. The graph can be easily made symmetric, for example by using $W \leftarrow \frac{1}{2}\left(W + W^T\right)$.*

Definition 1.4 (Degree and Degree matrix)**.** *The degree d_i of a vertex v_i and the graph degree matrix D are defined as*

- *d_i is the sum of weights on the edges of v_i, i.e., $d_i = \sum_j w_{ij}$.*
- *D is a diagonal matrix with elements $D_{ii} = d_i$.*

Definition 1.5 (Unnormalized graph Laplacian)**.** *The unnormalized graph Laplacian is defined as $L_{un} = D - W$.*

Observe that L_{un} is symmetric.

Proposition 1.6. *For every vector $f \in \mathbb{R}^n$, $f^T L_{un} f = \frac{1}{2}\sum_{i,j} w_{ij}\left(f_i - f_j\right)^2$.*

Proof. Exercise. □

It follows that L_{un} is positive semi-definite

Proposition 1.7. *The smallest eigenvalue of of L_{un} is 0, and its corresponding eigenvector is the constant vector $\frac{1}{\sqrt{n}}\mathbb{1}$.*

Proof. Each row of L_{un} sums to zero, so $L_{\text{un}}\mathbf{1} = 0$. As it is PSD, zero is its smallest eigenvalue. □

Proposition 1.8. *The multiplicity of the zero eigenvalue equals the number k of connected components of G, and the corresponding eigenvectors are indicator vectors of the components.*

Proof. We know that 0 is an eigenvalue. Let f be a corresponding eigenvector. Then since $f^T L_{\text{un}} f = 0$, we have that $f_i = f_j$ whenever $w_{ij} > 0$. For $k > 1$ connected components, L_{un} is block diagonal, so its spectrum is the union of the spectra of all blocks. □

Definition 1.9 (Normalized graph Laplacian)**.** *Assume $d_i > 0$ for each vertex (or handle isolated vertices separately). We define two versions of normalized Laplacians:*

- *The random walk graph Laplacian is $L_{rw} = D^{-1} L_{un} = I - D^{-1} W$.*
- *The Symmetric graph Laplacian is $L_{sym} = D^{-\frac{1}{2}} L_{un} D^{-\frac{1}{2}} = I - D^{-\frac{1}{2}} W D^{-\frac{1}{2}}$.*

Remark 1.10. *Observe that L_{rw} and L_{sym} are similar matrices, and hence share the same spectrum. In particular, all eigenvalues of L_{rw} are real.*

Proposition 1.11. *L_{sym} is positive semi-definite (and consequently L_{rw} has nonnegative real eigenvalues, although L_{rw} itself is not symmetric)*

Proof. Exercise. □

Proposition 1.12. *0 is an eigenvalue of L_{rw} with eigenvector $\mathbf{1}$, and of L_{sym} with eigenvector $D^{1/2}\mathbf{1}$.*

Proof. Exercise. □

Proposition 1.13. *Assuming every degree is positive, the multiplicity of the zero eigenvalue of both normalized Laplacians is the number of connected components. For L_{rw} the kernel is spanned by component indicators; for L_{sym} it is spanned by $D^{1/2}$ times those indicators.*

Proof. Analogous to the proof of proposition 1.8. □

1.2 Spectral Clustering

We have seen that eigenvectors corresponding to the zero eigenvalues (to whom we refer from now the first eigenvectors) of the Laplacian indicate the connected components. It therefore makes sense to use these eigenvectors to represent the data and identify the cluster structure.

Spectral clustering works by representing the data using the first k eigenvectors, and running k-means on this representation. Unlike k-means it can handle non-convex data (see Figure 1. Spectral clustering

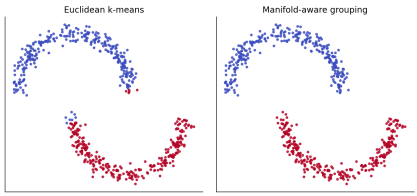


Figure 1: Euclidean k-means (left) and the desired non-convex grouping (right), shown on a two-moons toy dataset.

can interestingly be motivated by a graph cut point of view (see section 5 in Von Luxburg's excellent tutorial).

Remark 1.14. *Since the first eigenvector is constant in L_{un} and L_{rw}, it is often omitted.*

Further Reading

For further reading see von Luxburg, *A Tutorial on Spectral Clustering*.

2 Diffusion Maps

Diffusion map is a dimensionality reduction technique, that captures geometrical properties of data. In order to do so, it utilizes weighted graphs, which encode local similarity between pairs of points. This local interaction then allows to obtain global representations of the entire data.

Observe that $P := D^{-1}W$ is a Markov matrix, and hence encodes transition probabilities between the vertices. P is a diffusion (averaging) operator, and defines directions of propagation. powers of P corresponds to multi-step walks.

Proposition 2.1. *P has stationary distribution π, with $\pi_i = \frac{d_i}{\sum_j d_j}$, i.e., $\pi^T P = \pi^T$.*

Proof. Exercise. □

Corollary 2.2. *a stationary distribution π is a left eigenvector of P, with eigenvalue 1. Since left and right eigenvalues are the same, P has a 1 eigenvalue.*

Corollary 2.3. *Let u be an arbitrary vector of length n. Each entry of Pu is a convex combination of the entries of u, and hence cannot be larger than the maximal one. Thus every eigenvalue of P has absolute value at most 1 (apply the sup norm to an eigenvector).*

Order the real eigenvalues of P as $\lambda_0 = 1$ and $|\lambda_1| \geq |\lambda_2| \geq \cdots$; write ψ_ℓ for the corresponding right eigenvectors, normalized by $\psi_\ell^T D \psi_j = \delta_{\ell j}$.

Definition 2.4 (Diffusion distance)**.** *The (non-squared) diffusion distance at integer time $t \geq 0$ is*

$$D_t(i,j) = \left(\sum_{k=1}^{n} \frac{[P_{ik}^t - P_{jk}^t]^2}{d_k} \right)^{1/2}.$$

An alternative conventional definition weights by $1/\pi_k$, yielding the same distance up to the common factor $\sqrt{\sum_k d_k}$.

A small diffusion distance means that the random walks starting from i and j have similar distributions after t steps.

Theorem 2.5 (Diffusion-distance identity)**.** *With $\psi_\ell^T D \psi_j = \delta_{\ell j}$,*

$$D_t(i,j)^2 = \sum_{\ell=1}^{n-1} \lambda_\ell^{2t} [\psi_\ell(i) - \psi_\ell(j)]^2.$$

The constant eigenvector ψ_0 contributes nothing to differences.

Proof. Let $S = D^{-1/2} W D^{-1/2} = \sum_\ell \lambda_\ell \phi_\ell \phi_\ell^T$ with orthonormal ϕ_ℓ, and $\psi_\ell = D^{-1/2}\phi_\ell$. Then $P^t = \sum_\ell \lambda_\ell^t \psi_\ell \phi_\ell^T D^{1/2}$. Therefore

$$\frac{P_{ik}^t - P_{jk}^t}{\sqrt{d_k}} = \sum_\ell \lambda_\ell^t [\psi_\ell(i) - \psi_\ell(j)] \phi_\ell(k).$$

Square, sum over k, and use orthonormality of the ϕ_ℓ. □

Proposition 2.6. *Let $P = D^{-1}W$ and $A = D^{-\frac{1}{2}} W D^{-\frac{1}{2}}$. Let (λ, ϕ) be an eigenpair of A. Then $(\lambda, D^{-\frac{1}{2}}\phi)$ are an eigenpair of P.*

Proof. Exercise. □

2.1 Representation

The above suggests that we can represent a node v_i using a feature vector

$$\Psi_t(v_i) = \left(\lambda_1^t \psi_1(i), \ldots, \lambda_L^t \psi_L(i)\right)^T,$$

for some $L \leq n$.

Recall that by proposition 1.11, L_{sym} is positive semi-definite. The entire spectrum of L_{rw} lies between 0 and 2, and consequently, the spectrum of $P = I - L_{\text{rw}}$ lies between -1 and 1. For $t > 0$, eigenmodes with small $|\lambda_\ell|^t$ contribute little to the diffusion distance. A truncation is accurate when the neglected terms are small. Note that eigenvalues near -1 do not decay rapidly (as on bipartite graphs); a lazy random walk removes this periodicity.