Spectral Clustering - Analysis

Eric Walton

1 Background

1.1 Block Diagonal matrices

Define a block diagonal matrix M to be a square matrix with all zeros, except square matrices M_i placed corner-to-corner along the diagonal:

$$M = \begin{bmatrix} M_1 & 0 & \cdots & 0 \\ 0 & M_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & M_n \end{bmatrix}$$

Note that each M_i can be a different size.

Theorem 1.1 *The eigenvalues of M are exactly the union of the eigenvalues, with multiplicities, from each M_i.*

Proof If λ is an eigenvalue of one of the matrices M_i, then for some v, $M_i v = \lambda v$. Taking X and Y to be arbitrary square matrices, we have:

$$\begin{bmatrix} X & 0 & 0 \\ 0 & M_i & 0 \\ 0 & 0 & Y \end{bmatrix} \begin{bmatrix} 0 \\ v \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ M_i v \\ 0 \end{bmatrix} = \lambda \begin{bmatrix} 0 \\ v \\ 0 \end{bmatrix}$$

Therefore λ is an eigenvalue of M as well. This holds for all λ, thus all eigenvalues of any of the M_i are also eigenvalues of M. Since the width of M is exactly the sum of the widths of each M_i, there can be no additional eigenvalues for M. ∎

1.2 Order doesn't matter

This will formally show what is intuitively obvious: that the order we choose data points doesn't matter.

We first choose an index for each data point. The weighted adjacency graph A reflects this, as $A_{i,j}$ is the similarity between the ith and jth data points.

Suppose we have two different orderings of data. From those we produce A and $\hat{A}$. Suppose permutation matrix P transforms A into $\hat{A}$ as follows:

$$\hat{A} = PAP^T$$

Why two permutation matrices? Suppose P flips rows 2 and 3. Then PA flips those two rows. But then we need to flip columns 2 and 3. So, transpose the result, flips rows 2 and 3, and transpose it back. In other words, compute this:

$$(P(PA)^T)^T = PAP^T$$

Likewise, let:

$$\hat{D} = PDP^T$$

Compute the graph LaPlacian:

$$\hat{L} = PDP^T - PAP^T = P(D - A)P^T = PLP^T$$

How do the eigenvalues and eigenvectors of L and $\hat{L}$ differ? Let x be an eigenvector of L, with eigenvalue λ.

$$Lx = \lambda x$$

Since every permutation matrix is orthogonal, $P^T P = I$:

$$LP^T Px = \lambda x$$

$$PLP^T Px = P\lambda x$$

$$\hat{L}(Px) = \lambda(Px)$$

For every eigenvalue/eigenvector pair for L, above holds. That is, Px is an eigenvector of $\hat{L}$ with the same eigenvalue λ. Px is simply the same permutation of the rows of vector x.

To conclude, changing order of data points in A makes only one difference: the vectors u_k will change their ordering in the same way.

2 Perturbation theory

In this section we will show how small changes in a matrix make only small changes to the eigenvalues and eigenvectors.

Let M be an $n \times n$ matrix. Let λ be an eigenvalue with eigenvector x.

2.1 Eigenvalues: Weyl's inequality

Let A be an $n \times n$ hermitian matrix, with eigenvalues $\alpha_1 \dots \alpha_n$. This is our "actual" matrix.

Let E be an $n \times n$ hermitian matrix, whose values are small. This is our "error" matrix. Let $\epsilon_1 \dots \epsilon_n$ be the eigenvalues of E.

Let $I = A + E$. This is our "ideal" matrix. Let $\iota_1 \dots \iota_n$ be the eigenvalues of $A+$.

Order the eigenvalues of A, E, and I so that they are in ascending order.

Then for all $i = 1, \dots, n$:

$$\iota_i + \epsilon_1 \leq \alpha_i \leq \iota_i + \epsilon_n$$

In plain terms, each eigenvalue α_i differs from its "ideal" counterpart ι_i by at most the largest eigenvalue of E, ϵ:

$$\iota_i \leq \alpha_i \leq \iota_i + \epsilon_n$$

The matrix of all zeros has all zero eigenvalues. Thus as $E \to 0$ $\epsilon_n \to 0$.

2.2 Eigenvectors

Suppose that A is diagonalizable matrix as follows:

$$A = X\Lambda X^{-1}$$

$$AX = X\Lambda$$

Suppose, further, that A (and therefore X and Λ), depend on some real-valued parameter p. Take the derivative of both sides with respect to p, using the product rule:

$$A'X + AX' = X'\Lambda + X\Lambda'$$

$$A'X - X\Lambda' = -AX' + X'\Lambda$$

Left-multiply both sides by X^{-1}:

$$X^{-1}A'X - \Lambda' = -X^{-1}AX' + X^{-1}X'\Lambda$$

Introduce C such that $X' = XC$.

$$X^{-1}A'X - \Lambda' = -X^{-1}AXC + X^{-1}XC\Lambda$$

Simplify:

$$X^{-1}A'X - \Lambda' = -\Lambda C + C\Lambda$$

Parse the entries of each side:

$$(-\Lambda C + C\Lambda)_{ij} = -\lambda_i c_{ij} + c_{ij}\lambda_j = c_{ij}(\lambda_j - \lambda_i)$$

$$(X^{-1}A'X - \Lambda')_{ij} = x_{i-}^{-1}A'x_{-j} - \Lambda'_{ij}$$

Where $i \neq j$:

$$c_{ij} = \frac{x_{i-}^{-1}A'x_{-j}}{\lambda_j - \lambda_i}$$

Trust me, when A' is small, the entries c_{ij} are also small. Therefore X' is small. In other words, small changes to A make only small changes to X.

3 Problem statement

Given a data set $x_1, x_2, \ldots, x_n$.

The goal is to create a clustering that groups data according to proximity. In particular, we want to group data that is "clustered" in non-linear ways, by modeling proximity between near datapoints.

4 Similarity matrix

Model the data set as a massive edge-weighted, undirected graph. The graph will have n vertices, one for each data point.

Choose a method of modeling similarity between vertices. Assign each edge weight to be the similarity measure between the data points. Similarity is between 0 and 1. Create an $n \times n$ matrix S. Let $S_{i,j}$ be the similarity measure between points x_i and x_j.

Given this matrix S, we find clustering through a graph partitioning algorithm.

4.1 Similarity measures: examples

Neighborhood:

$$S_{i,j} = \begin{cases} 1 & \text{if } \|x_i - x_j\| < \epsilon \\ 0 & \text{if otherwise} \end{cases}$$

Guassian:

$$S_{i,j} = \exp\left[\frac{-\|x_i - x_j\|^2}{2\sigma^2}\right]$$

5 Graph LaPlacian

5.1 Definition

Suppose an undirected graph G, with vertices $V = \{v_1, v_2, ...\}$ and edges $E = \{e_1, e_2, ...\}$. Assume G has no self-loops, and no parallel edges.

Suppose the adjacency matrix of the graph is:

$$A = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \end{bmatrix}$$

The degree matrix is a diagonal matrix whose entries are the sum of each row:

$$D = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

The Laplacian matrix is formed by taking $D - A$. In this case:

$$L = \begin{bmatrix} 1 & 0 & 0 & -1 & 0 & 0 \\ 0 & 2 & 0 & -1 & -1 & 0 \\ 0 & 0 & 1 & 0 & 0 & -1 \\ -1 & -1 & 0 & 2 & 0 & 0 \\ 0 & -1 & 0 & 0 & 1 & 0 \\ 0 & 0 & -1 & 0 & 0 & 1 \end{bmatrix}$$

The entries of L are:

$$L_{i,j} = \begin{cases} deg(v_i) & \text{if } i = j \\ -1 & \text{if } i \neq j \text{ and } v_i \text{ is adjacent to } v_j \\ 0 & \text{otherwise.} \end{cases}$$

Matrix L can also be formed from the incidence matrix. Let M be the $|E| \times |V|$ incidence matrix of G. For each edge e that connects vertex i and j, such that $i < j$, assign:

$$M_{e,i} := -1 \quad \text{and} \quad M_{e,j} := 1$$

All other entries are 0. Note that the edges can be listed in any order.
The incidence matrix from G could be written:

$$M = \begin{bmatrix} 0 & -1 & 0 & 0 & 1 & 0 \\ -1 & 0 & 0 & 1 & 0 & 0 \\ 0 & -1 & 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 & 1 \end{bmatrix}$$

It can be shown that for every graph:

$$L = M^T M$$

Therefore, L is at least positive semi-definite. In fact, since every row of M contains exactly one 1 and one -1, then M times the vector of all 1s is 0. Let $\vec{1}$ be a vector of all 1s:

$$M\vec{1} = \vec{0}$$
$$M^T M\vec{1} = \vec{0}$$
$$L\vec{1} = \vec{0}$$

These vectors $\vec{0}$ are each different sizes, but this shows that $\vec{1}$ is also in the nullspace of L, and that therefore L is positive semi-definite.

5.2 Further property

As shown above, L is positive semi-definite.

L also satisfies, for all $v \in \mathbf{R}^n$:

$$v^T L v = \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} A_{i,j}(v_i - v_j)^2$$

5.3 Proof

Remember the items of D are the row-wise sums of A. So:

$$v^T D v = v_1 \left(\sum_{i=1}^{n} A_{1,i} \right) v_1 + v_2 \left(\sum_{i=1}^{n} A_{2,i} \right) v_2 + \cdots + v_n \left(\sum_{i=1}^{n} A_{n,i} \right) v_n$$

$$= \sum_{j=1}^{n} \sum_{i=1}^{n} v_j^2 A_{j,i}$$

Breaking down $v^T A v$, we get:

$$v^T A v = \begin{bmatrix} v_1 & v_2 & \dots & v_n \end{bmatrix} \begin{bmatrix} A_{1,1} & A_{1,2} & \dots & A_{1,n} \\ A_{2,1} & A_{2,2} & \dots & A_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ A_{n,1} & A_{n,2} & \dots & A_{n,n} \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix}$$

$$= \begin{bmatrix} v_1 & v_2 & \dots & v_n \end{bmatrix} \begin{bmatrix} \sum_{i=1}^{n} v_i A_{1,i} \\ \sum_{i=1}^{n} v_i A_{2,i} \\ \vdots \\ \sum_{i=1}^{n} v_i A_{n,i} \end{bmatrix}$$

$$= v_1 \sum_{i=1}^{n} v_i A_{1,i} + v_2 \sum_{i=1}^{n} v_i A_{2,i} + \cdots + v_n \sum_{i=1}^{n} v_i A_{n,i}$$

$$= \sum_{j=1}^{n} \sum_{i=1}^{n} v_j v_i A_{j,i}$$

Putting the two together:

$$v^T L v = v^T (D - A) v = v^T D v - v^T A v$$

$$= \sum_{j=1}^{n}\sum_{i=1}^{n} v_j^2 A_{j,i} - \sum_{j=1}^{n}\sum_{i=1}^{n} v_j v_i A_{j,i}$$

$$= \sum_{j=1}^{n}\sum_{i=1}^{n} A_{j,i}\left(v_j^2 - v_j v_i\right)$$

$$= \frac{1}{2}\sum_{j=1}^{n}\sum_{i=1}^{n} A_{j,i}\left(v_j^2 - v_j v_i\right) + \frac{1}{2}\sum_{j=1}^{n}\sum_{i=1}^{n} A_{j,i}\left(v_j^2 - v_j v_i\right)$$

Swap i and j in the second term. Since A is symmetric, $A_{j,i} = A_{i,j}$ so no need to swap A's indices:

$$= \frac{1}{2}\sum_{j=1}^{n}\sum_{i=1}^{n} A_{j,i}\left(v_j^2 - v_j v_i\right) + \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} A_{j,i}\left(v_i^2 - v_i v_j\right)$$

Exchange the two sums in the second term and combine:

$$= \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} A_{j,i}\left(v_i^2 - 2v_i v_j + v_j^2\right)$$

$$= \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} A_{j,i}\left(v_i - v_j\right)^2$$

6 Key Theorems

Let graph G have connected components $A_1, A_2, \ldots, A_K$. We define:

$$\vec{1}_k = \begin{bmatrix} \gamma_1 \\ \vdots \\ \gamma_n \end{bmatrix} \text{ where } \gamma_i = \begin{cases} 1, & \text{if } x_i \in A_k \\ 0, & \text{if } x_i \notin A_k \end{cases}$$

Theorem 6.1 *The nullspace of L has dimension K and is spanned by* $\vec{1}_1, \vec{1}_2, \ldots, \vec{1}_K$.

Proof

Case 1: $K = 1$

Since there is only one cluster, $\vec{1}_1$ is all 1s. By the properties of the graph LaPlacian, $\vec{1}_1 \in \mathbf{N}(L)$.

Next, we need to show that if $Lf = 0$, then $f = \alpha\vec{1}_1$.

$$Lf = 0$$

$$0 = f^T Lf = \sum_i \sum_j A_{i,j}(f_i - f_j)^2$$

Since we have only one cluster, $A_{i,j} > 0$ for all i, j. Therefore, for all i and j:

$$f_i = f_j$$

Thus, every f is a multiple of $\vec{1}_1$. Therefore, $dim(\mathbf{N}(L)) = 1$.

Case 2: $K > 1$

Reorder the vertices by connected component. This can be done as the order we list vertices is arbitrary. L becomes a block diagonal matrix:

$$L = \begin{bmatrix} L_1 & 0 & . & 0 \\ 0 & L_2 & . & 0 \\ . & . & . & . \\ 0 & 0 & . & L_K \end{bmatrix}$$

Each L_k is a graph LaPlacian matrix. They are each of different sizes, but because of Case 1 above, $L_k\vec{1} = 0$.

Therefore, in a similar case to the block diagonal discussion above::

$$L\vec{1}_k = \begin{bmatrix} X & 0 & 0 \\ 0 & L_k & 0 \\ 0 & 0 & Y \end{bmatrix} \begin{bmatrix} 0 \\ \vec{1} \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

Thus, every $\vec{1}_k$ is in the nullspace of L. Note, in addition, that the vectors $\vec{1}_k$ are mutually independent.

Since $\mathbf{N}(L_k)$ has dimension 1 for all k, then $\mathbf{N}(L)$ has dimension K. But we've already found K independent vectors in $\mathbf{N}(L)$, namely $\{\vec{1}_1, \vec{1}_2, \ldots, \vec{1}_K\}$. Therefore there are no others needed and the vectors $\vec{1}_k$ span $\mathbf{N}(L)$.

7 Ideal Case

7.1 Key Theorem

Suppose all similarities between clusters are 0. Compute basis $u_1, u_2, \ldots, u_K$ for $\mathbf{N}(L)$. Because of above theorem, $span\{u_1, u_2, \ldots, u_K\} = span\{\vec{1}_1, \vec{1}_2, \ldots, \vec{1}_K\}$.

Define a matrix Y such that:

$$Y = \begin{bmatrix} u_1 & u_2 & \cdots & u_K \end{bmatrix}$$

Theorem 7.1 *If x_i and x_j are in the same cluster, then:*

$$row\ i\ of\ Y = row\ j\ of\ Y$$

Proof By above theorem, for each u_k:

$$u_k = \alpha_1^{(k)}\vec{1}_1 + \alpha_2^{(k)}\vec{1}_2 + \cdots + \alpha_K^{(k)}\vec{1}_K$$

for constants $\alpha_i^{(k)}$. Suppose, as above proof, the datapoints are listed by cluster. We can write above equation with block vectors:

$$u_k = \alpha_1^{(k)} \begin{bmatrix} \vec{1} \\ 0 \\ 0 \\ \vdots \end{bmatrix} + \alpha_2^{(k)} \begin{bmatrix} 0 \\ \vec{1} \\ 0 \\ \vdots \end{bmatrix} + \ldots + \alpha_K^{(k)} \begin{bmatrix} 0 \\ 0 \\ \vdots \\ \vec{1} \end{bmatrix} = \begin{bmatrix} \alpha_1^{(k)}\vec{1} \\ \alpha_2^{(k)}\vec{1} \\ \vdots \\ \alpha_K^{(k)}\vec{1} \end{bmatrix}$$

Thus:

$$Y = \begin{bmatrix} \alpha_1^{(1)}\vec{1} & \alpha_1^{(2)}\vec{1} & \cdots & \alpha_1^{(K)}\vec{1} \\ \alpha_2^{(1)}\vec{1} & \alpha_2^{(2)}\vec{1} & \cdots & \alpha_2^{(k)}\vec{1} \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_K^{(1)}\vec{1} & \alpha_K^{(2)}\vec{1} & \cdots & \alpha_K^{(K)}\vec{1} \end{bmatrix}$$

If x_i and x_j belong to the same cluster, rows i and j would be within the same row-block above, and have the same values. If we change the order of our data, then A, D, and L all change the order they list data, and our matrix Y would be unchanged, except with rows in a different order. ∎

7.2 Example

Using example from earlier. We have two clusters in that case.

$$u_1 = \alpha_1\vec{1}_1 + \beta_1\vec{1}_2$$

$$u_2 = \alpha_2\vec{1}_1 + \beta_2\vec{1}_2$$

Using our data from before:

$$\vec{1}_1 = \begin{bmatrix} 1 \\ 1 \\ 0 \\ 1 \\ 1 \\ 0 \end{bmatrix} \qquad \vec{1}_2 = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 1 \end{bmatrix}$$

Therefore:

$$Y = \begin{bmatrix} u_1 & u_2 \end{bmatrix} = \begin{bmatrix} \alpha_1 & \alpha_2 \\ \alpha_1 & \alpha_2 \\ \beta_1 & \beta_2 \\ \alpha_1 & \alpha_2 \\ \alpha_1 & \alpha_2 \\ \beta_1 & \beta_2 \end{bmatrix}$$

Notice that rows 1,2,4 and 5, that belong to the same component, have the same coefficients. Likewise with rows 3 and 6.

8 General case

In general, we may not have any 0-valued eigenvalues of L. Perhaps each similarity between clusters is simply a very low, positive, number. However, because of perturbation theory of eigenvectors and eigenvalues, the eigenvectors and eigenvalues will be close to what the would have been in the ideal case. For matrix E with small values:

$$L_{actual} = L_{ideal} + E$$

Therefore, the smallest eigenvalues of L_{actual} will correspond to the zero eigenvalues of L_{ideal}. Instead of solving the nullspace, choose the k smallest eigenvalues of L_{actual}.

9 Acknowledgements

This was originally an assignment from Gopal Nataraj, and portions of this were adapted from his lectures. I freely consulted, and used the wikipedia articles on the topic:

- `http://en.wikipedia.org/wiki/Eigenvalue_perturbation`
- `http://en.wikipedia.org/wiki/Estimation_of_covariance_matrices`

As well as this scholarly article from the Electronic Journal of Linear Algebra by N.P. van der Aa, H.G. Term Morsche, and R.R.M. Maathiej:

`http://www.math.technion.ac.il/iic/ela/ela-articles/articles/vol16_pp300-314.pdf`