EE381 V: Large Scale Machine Learning | Spring 2013

Lecture 5 — January 29

Lecturer: Constantine Caramanis and Sujay Sanghavi | *Scribe: Suriya Gunasekar*

5.1 Recap

Let $\{x_1, x_2, \ldots, x_n\}$ be the set of data points that need to be clustered. A graph, $G = (V, E)$, can be defined over the points such that V is the set of data points and $E = \{(i, j) : A_{ij} > 0\}$, where A_{ij} is a measure of similarity or "closeness" between points x_i and x_j. An example of a similarity measure is $\exp(-\frac{1}{2\sigma^2}\|x_i - x_j\|^2)$.

- The similarity matrix, A, is defined such that $A_{ij} = \exp(-\frac{1}{2\sigma^2}\|x_i - x_j\|^2)$.
- The degree matrix, D, is a diagonal matrix with the diagonal entries given by, $D_{ii} = d_i = \sum_{j\in[n]} A_{ij}$.
- Finally, the Laplacian, L, and the normalized Laplacian, L_n, of the graph, G, are defined as, $L \triangleq D - A$ and $L_n \triangleq I - D^{-1/2}AD^{-1/2}$ respectively.

It is easy to see that:

$$L_n = D^{-1/2}LD^{-1/2} \tag{5.1}$$

5.2 Spectral Clustering Algorithm

1. Compute the normalized Laplacian, L_n
2. Let, $u_1, u_2, \ldots, u_k$ be the bottom k eigenvectors (corresponding to the k smallest eigenvalues) of L_n
3. Define $U = \begin{bmatrix} \uparrow & \uparrow & & \uparrow \\ u_1 & u_2 & \ldots & u_k \\ \downarrow & \downarrow & & \downarrow \end{bmatrix}$
4. For $i = \{1, 2, \ldots, n\}$, let $y_i \in \mathbb{R}^k$ be the rows of matrix E and $\hat{x}_i = \frac{y_i}{\|y_i\|_2}$.
5. Run k-means clustering (or any distance based clustering) on $\{\hat{x}_i\}$

5.3 General definitions and results

5.3.1 Spectral Theorem

If $M \in S^n$ is an $n \times n$ symmetric matrix, then:

1. M has an orthogonal basis of eigenvectors $T = \begin{bmatrix} \uparrow & \uparrow & & \uparrow \\ u_1 & u_2 & \dots & u_n \\ \downarrow & \downarrow & & \downarrow \end{bmatrix}$

2. $M = T\Lambda T^* = \sum_i \lambda_i u_i u_i^*$

5.3.2 Singular Value Decomposition

Any rank-k matrix $M \in \mathbb{R}^{m\times n}$ can be written as:

$$M = [U_1 \mid U_2] \begin{bmatrix} \Sigma & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} [V_1 \mid V_2]^* = U_1 \Sigma V_1^*$$

where $U_1 \in \mathbb{R}^{m\times k}$, $V_1 \in \mathbb{R}^{n\times k}$ and $\Sigma \in \mathbb{R}^{k\times k}$. U_1 and V_1 are orthonormal, (i.e. $U_1^* U_1 = V_1^* V_1 = I$) and Σ is the diagonal matrix with diagonal entries $\sigma_1 \geq \sigma_2 \geq \dots \sigma_k > 0$

5.3.3 Matrix Norms

For a rank-k matrix M with singular values $\sigma_1 \geq \sigma_2 \geq \dots \sigma_k > 0$, the following norms are defined:

1. Frobenius Norm: $\|M\|_F = \left(\sum_{ij} M_{ij}^2\right)^{1/2} = \left(\sum_{i=1}^k \sigma_i^2\right)^{1/2}$

2. Operator Norm: $\|M\|_2 = \max\limits_{u:\|u\|_2=1} \|Mu\|_2 = \max\limits_{u,v:\|u\|_2=\|v\|_2=1} v^T M u = \sigma_1$

Exercise If $M \in \mathbb{R}^{m\times n}$ and if $T_1 \in \mathbb{R}^{m\times m}$ and $T_2 \in \mathbb{R}^{n\times n}$ are orthonormal matrices, i.e. $T_1^* T_1 = T_1 T_1^* = I_m$ and $T_2^* T_2 = T_2 T_2^* = I_n$, then $\|T_1 M T_2\|_2 = \|M\|_2$.

Proof:

$$\|T_1 M T_2\|_2 = \max_{u,v} \frac{v^T T_1 M T_2 u}{\|v\|_2 \|u\|_2} \tag{5.2}$$

$$= \max_{u,v} \frac{(T_1 v)^T M (T_2 u)}{\|T_1 v\|_2 \|T_2 u\|_2} \tag{5.3}$$

$$= \max_{x,y} \frac{x^T M y}{\|x\|_2 \|y\|_2} = \|M\|_2 \tag{5.4}$$

where, Equation 5.3 follows as for any orthonormal matrix Q with $Q^* Q = I$, we have $\|Qx\|_2 = (Qx)^*(Qx) = x^* Q^* Q x = x^* x = \|x\|_2$; and Equation 5.4 follows by redefining variables as $x = T_1 v$ and $y = T_2 u$. □

5.4 Goodness of the spectral clustering algorithm

Theorem 5.1. *L and L_n are positive semi-definite. If G is a fully connected graph, the smallest eigenvalue of L_n, $\lambda_1(L_n) = 0$ and the eigenvector corresponding to this eigenvalue is given by* $u_1 = \begin{pmatrix} \sqrt{d_1} \\ \sqrt{d_2} \\ \vdots \\ \sqrt{d_n} \end{pmatrix}$.

Proof: Consider $L = D - A$. For $v \in \mathbb{R}^n$, we have:

$$\begin{aligned} v^T L v = &\ v^T D v - v^T A v \\ = &\ \sum_i v_i^2 d_i - \sum_{ij} v_i A_{ij} v_j \\ = &\ \sum_i v_i^2 \left[\sum_j A_{ij} \right] - \sum_{ij} v_i A_{ij} v_j \\ = &\ \sum_{ij} A_i j (v_i^2 - v_i v_j) \\ = &\ \frac{1}{2} \sum_{ij} A_{ij} (v_i^2 + v_j^2 - 2 v_i v_j) \\ = &\ \frac{1}{2} \sum_{ij} (v_i - v_j)^2 A_{ij} \end{aligned} \tag{5.5}$$

From Equation 5.1, for $v \in \mathbb{R}^n$, we have:

$$v^T L_n v = (D^{-1/2} v)^T L (D^{-1/2} v) = \frac{1}{2} \sum_{ij} \left(\frac{v_i}{\sqrt{d_i}} - \frac{v_j}{\sqrt{d_j}} \right)^2 A_{ij} \geq 0 \tag{5.6}$$

From Equations 5.5 and 5.6, we have $L, L_n \succeq 0$.

Further it can be verified that with $u_1 = \begin{pmatrix} \sqrt{d_1} \\ \sqrt{d_2} \\ \vdots \\ \sqrt{d_n} \end{pmatrix}$, we have $Au_1 = 0$. This implies that 0 is an eigenvalue of L_n and as $L_n \succeq 0$, it is also the smallest eigenvalue, i.e. $\lambda_1(L_n) = 0$. □

Theorem 5.2. *If G is disconnected with k connected components, then the spectral clustering algorithm returns exact clustering.*

Proof: If G is disconnected, the nodes can be rearranged such that A and hence L_n is block diagonal, with k blocks.

$$L_n = \begin{bmatrix} L_n^{(1)} & & & \\ & L_n^{(2)} & & \\ & & \ddots & \\ & & & L_n^{(k)} \end{bmatrix}$$

The eigenvalues of L_n are the union of the eigenvalues of $L_n^{(i)}$. $\Lambda(L_n) = \Lambda\left(L_n^{(1)}\right) \cup \Lambda\left(L_n^{(2)}\right) \cup \ldots \cup \Lambda\left(L_n^{(k)}\right)$. Similarly, eigenvectors of L_n are the union of the appropriately zero padded eigenvectors of $L_n^{(i)}$. $spec(L_n) = spec\left(L_n^{(1)}\right) \cup spec\left(L_n^{(2)}\right) \cup \ldots \cup spec\left(L_n^{(k)}\right)$

Each diagonal block, $L_n^{(i)}$, is a completely connected component and hence by Theorem 5.1, $\lambda_1\left(L_n^{(i)}\right) = 0\ \forall\ i \in [k]$ and the corresponding eigenvectors are $u_1\left(L_n^{(1)}\right) = \begin{pmatrix} \sqrt{d_1^{(1)}} \\ \vdots \\ \sqrt{d_{|S_1|}^{(1)}} \end{pmatrix}$, $u_1\left(L_n^{(2)}\right) = \begin{pmatrix} \sqrt{d_1^{(2)}} \\ \vdots \\ \sqrt{d_{|S_2|}^{(2)}} \end{pmatrix}$ and so on. Thus, the bottom k eigenvectors form the matrix:

$$U = \begin{bmatrix} u_1 & 0 & \ldots & 0 \\ 0 & u_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & u_k \end{bmatrix}$$

In this case, for $i \in [n]$, $\hat{x}^{(i)} = \frac{y_i}{\|y_1\|_2} = e_a$, where, $x_i \in S_a$ (S_a denotes the cluster indexed by a), where e_a is a standard basis vector.
However for a general matrix A (which is typically not block diagonal as assumed above), the matrix with the bottom k eigenvectors can be written as $\hat{U} = UQ$, the new $\hat{x}_{new}^{(i)} = Q\hat{x}^{(i)}$. However as unitary transformations preserve the distances (i.e $\langle x, y \rangle = \langle Qx, Qy \rangle$), any distance based clustering algorithm would perfectly retrieve the clusters. $\square$

5.4.1 Perturbation of symmetric matrices

Consider a symmetric perturbation of a symmetric matrix $M \in \mathbb{R}^{n\times n}$ given by $M+\Delta$ ($M+\Delta$ is also symmetric). Let $E_0 \in \mathbb{R}^{n\times k}$ be the matrix formed with the bottom k eigenvectors of M as the columns and F_0 be the corresponding matrix for $M+\Delta$. We define the divergence between the subspaces spanned by E_0 and F_0 as follows:

$$d_p(E_0, F_0) = \|E_0E_0^* - F_0F_0^*\|_2 \tag{5.7}$$

Lemma 5.3. *If the basis from E_0 and F_0 is completed as $E = [E_0|E_1]$ and $F = [F_0|F_1]$ respectively. Then,*

$$d_p(E_0, F_0) = \|F_1^* E_0\|_2 = \|E_1^* F_0\|_2 = \|\sin\Theta\|_2$$

where, $\Theta = diag(\boldsymbol{\theta})$ and $\boldsymbol{\theta}$ is a vector of principal angles between subspaces spanned by columns of E_0 and F_0. The principal angles, θ_i, are defined such that $\cos\theta_i = \sigma_i(E_0^* F_0)$. Thus, we have the singular value decomposition of $E_0^* F_0$ as $E_0^* F_0 = U \cos\Theta V^*$, $U, V \in \mathbb{R}^{k\times k}$

Exercise $A = \begin{pmatrix} 0 & 0 \\ 0 & \epsilon \end{pmatrix}$ and $\Delta = \begin{pmatrix} 0 & \beta \\ \beta & \beta \end{pmatrix}$. Thus, $A + \Delta = \begin{pmatrix} 0 & \beta \\ \beta & \beta + \epsilon \end{pmatrix}$. The eigenvector corresponding to the smallest eigenvalues of A and $A + \Delta$ are given by $E_0 = e_0 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ and $F_0 = f_0 = \frac{1}{\sqrt{2D+2D^2}} \begin{pmatrix} 1+D \\ -\frac{2\beta}{\epsilon} \end{pmatrix}$ respectively, where $D = \sqrt{1 + \frac{4\beta^2}{\epsilon^2}}$.

$$d_p(e_0, f_0) = \|\sin\Theta\|_2 = \sqrt{1 - \langle e_0, f_0\rangle^2}$$

Also,

$$\langle e_0, f_0\rangle = \frac{1+D}{\sqrt{2D+2D^2}} = \sqrt{\frac{1+D}{2D}} = \frac{1}{\sqrt{2}}\sqrt{1 + \left(1 + \frac{4\beta^2}{\epsilon^2}\right)^{-1/2}} = \sqrt{1 - \frac{\beta^2}{\epsilon^2} + \mathcal{O}(\beta^4)} = 1 - \frac{\beta}{2\epsilon} + \mathcal{O}(\beta^2)$$

Thus, $d_p(e_0, f_0) = \frac{\beta}{\epsilon} + H.O.T$

Theorem 5.4. $\sin\Theta$ *theorem*:
If

$$M = [E_0 \mid E_1] \begin{bmatrix} M_0 & \mathbf{0} \\ \mathbf{0} & M_1 \end{bmatrix} [E_0 \mid E_1]^*$$

and

$$M + \Delta = [F_0 \mid F_1] \begin{bmatrix} \hat{M}_0 & \mathbf{0} \\ \mathbf{0} & \hat{M}_1 \end{bmatrix} [F_0 \mid F_1]^*$$

where, $M_0, M_1, \hat{M}_0, \hat{M}_1$ are diagonal matrices with the appropriate eigenvalues along the diagonal. If $\exists a, b, \delta$, such that $M_0(i,i) \in [a, b]$, $\forall i$ and $\hat{M}_1(i,i) \in (-\infty, a-\delta) \cup (b+\delta, \infty)$ $\forall i$ then, $d_p(E_0, F_0) \leq \frac{1}{\delta}\|\Delta\|_2$

Reference

[1] Ulrike Luxburg. 2007. `A tutorial on spectral clustering`. Statistics and Computing.