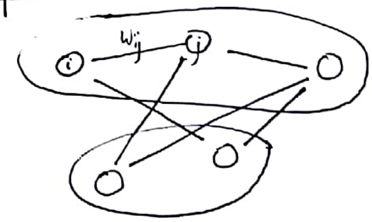
SPECTRAL CLUSTERING LECTURE

Core problem: mincut & variants



goal: minimize the weight of cut edges (subject to additional constraints)

Motivation: clustering of data

→ often, we get datapoints $x_1, \ldots, x_n$ with some notion of similarity: x_i & x_j similarity score is w_{ij}.

→ half of clustering is about MINIMIZING INTERCLUSTER WEIGHTS

· the other half is MAXIMIZING INTRACLUSTER WEIGHTS.

Note: many ways to construct this similarity matrix

- k-NN graph
- ε-neighborhood graph
- Gaussian kernel for Euclidean spaces $\quad w_{ij} = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right)$

Setting & Notation,

- $G = (V, E)$ undirected graph
- W weights matrix

w_{ij} is weight of an edge $(i,j) \in E$

$w_{ij} = 0$ if $(i,j) \notin E$

- Properties: - W is symmetric
  - if $w_{ij} = 1 \Leftrightarrow (i,j) \in E$, then W is the usual adjacency matrix.

- D degree matrix $= \operatorname{diag}(d_1, \ldots, d_n)$

$$d_i := \sum_{j \in V} w_{ij}$$

- if $S \subset V$, $\bar{S} := V \setminus S$
- if $S, T \subset V$, $E(S,T) := \{(i,j) \in E : i \in S, j \in T\}$

- $\partial S = E(S, \bar{S})$
- $\text{weight}(F) = \sum_{e \in F} w_e$ for $F \subset E$
- Measure of size $S \subset V$:
  - $|S| :=$ cardinality
  - $\text{vol}(S) := \sum_{i \in S} d_i = \sum_{\substack{i \in S \\ j \in V}} w_{ij} = \text{weight}(E(S, V))$.

MINCUT problem

$$\min_{\emptyset \subsetneq S \subsetneq V} \text{weight}(E(S,\bar{S})) = \text{weight}(\partial S).$$

Fact (Stoer-Wagner, 1995). There is an efficient algorithm that solves MINCUT.

QUESTION: is this good for clustering?

(No ...)

→ we can try to "balance" the size of each partition

→ turns out the problem becomes NP-hard (Wagner & Wagner 1993).

→ approach through spectral graph theory.

SPECTRAL GRAPH THEORY

At its heart, the graph Laplacian, L. Many connections

L

Algebraic Topology

- graph seen as 1-simplicial complex

↓

Space of vertex functions $\mathbb{R}^V$

↓

derivative

↓

L

Energy / Optimization

- functions on $\mathbb{R}^V$, with particular positive semidefinite quadratic form "energy" (cf harmonic analysis)

$$E(f) = \int \|\nabla f\|^2$$

→ isoperimetric problems, differential geometry, balance $E(f)$ & $\|f\|^2$

→ spectral theory

Probabalistic / random walks

- random walks on graph
- analysis of markov chains

REVIEW of LAPLACIAN

$G = (V, E)$

Dictionary between discrete calculus & vector calculus:

- V is space — M is space
- $\mathbb{R}^V$ is space of vertex functions — $C^\infty(M;\mathbb{R})$ space of functions
  - inner product space $\langle f, g\rangle_V := \sum_{i\in V} f(i)g(i)$ — inner product space $\langle f, g\rangle_M := \int_M f\cdot g$
- $\mathbb{R}^E$ is space of alternating vertex functions — $\mathfrak{X}(M)$ space of vector fields
  - $X([i,j]) = -X([j,i])$ for $(i,j)\in E$
  - inner product space
$$\langle X, Y\rangle_E := \sum_{e\in E} w_e X(e)Y(e)$$
- d the derivative operator $d: \mathbb{R}^V \to \mathbb{R}^E$ — ∇ gradient operator
$$df([i,j]) = f(j) - f(i)$$
- Dirichlet energy, measure of smoothness
$$E(f) := \langle df, df\rangle_E = \sum_{\substack{i,j\\ i<j}} w_{ij}(f(i)-f(j))^2 \qquad \mathcal{E}(f) = \int_M \|\nabla f\|^2$$

Recall the adjoint:

If $T: A \to B$ map of finite-dim inner product spaces, then

$\exists!\, T^*: B \to A$ s.t. $\forall a\in A, b\in B$, $\langle b, Ta\rangle_B = \langle T^*b, a\rangle_A$

- Laplacian L defined so that:
$$\langle df, df\rangle_E = \langle f, Lf\rangle_V$$

Why consider L? Optimization problem:

eg. $\min \langle df, df\rangle_E$ s.t. $\|f\|^2 = 1$ $\xrightarrow{\text{Lagrange multipliers}}$ $Lf = \lambda f$.

Properties of Laplacian:

① $L = D - W$

② $\mathbb{1} \in \ker(L)$

③ $L \in \mathrm{Sym}(\mathbb{R}^{n\times n})$

④ let $A\subset V$. $\mathbb{1}_A^T L \mathbb{1}_A = \sum_{\substack{(i,j)\in E\\ i<j}} w_{ij}(\mathbb{1}_A(i) - \mathbb{1}_A(j))^2 = \text{weight}(E(A,\bar{A}))$

REVIEW of SPECTRAL THEORY: geometric & variational characterization of eigenvalues

Def. Let $M \in \mathbb{R}^{n\times n}$. An eigenvector $v \in \mathbb{R}^n$ satisfies $v \neq 0$, $\exists \lambda \in \mathbb{R}$ s.t.

$$Mv = \lambda v.$$

The eigenvalues of a matrix form its spectrum.

Thm (Spectral Theorem for Sym. Mat).

Geometric characterization:

Def. Let $M \in \mathbb{R}^{n\times n}$. Let $v \in \mathbb{R}^n \setminus \{0\}$. Rayleigh quotient

$$R(v) = \frac{v^\top M v}{v^\top v}$$

notice: $R(\alpha v) = R(v)$ $\forall \alpha \in \mathbb{R}\setminus\{0\}$, can assume $v \in S^{n-1}$.

Thm (Variational Characterization). Let $M \in \mathrm{Sym}(\mathbb{R}^{n\times n})$

$\lambda_1 \leq \cdots \leq \lambda_n$ eigenvalues of M.

$$\lambda_1 = \min_{v\in S^{n-1}} R(v) \quad , \text{let } v_1 = \operatorname{argmin}$$

$$\vdots$$

$$\lambda_i = \min_{\substack{v\in S^{n-1} \\ v \perp \{v_1,\dots,v_{i-1}\}}} R(v)$$

Pf. minimize $v^\top M v$ s.t. $\|v\|^2 = 1$.

$\Rightarrow$ $Mv = \lambda v$, and $v^\top M v = \lambda v^\top v = \lambda$. $\square$

Exercise Let $M \in \mathrm{Sym}(\mathbb{R}^{n\times n})$. Let $X = \begin{bmatrix} | & & | \\ x_1 & \cdots & x_k \\ | & & | \end{bmatrix}$ $x_i \in \mathbb{R}^n$

$$\min_X \sum_{i=1}^{k} x_i^\top M x_i \quad \text{s.t. } x_i \perp x_j \text{ when } i \neq j, \ \|x_i\|^2 = 1$$

$$\Leftrightarrow \quad \min_X \operatorname{Tr}(X^\top M X) \quad \text{s.t. } X^\top X = \mathrm{Id}_{k\times k}.$$

QUESTION: is X unique?

No. Let $\mathcal{U}$ be unitary matrix in $\mathbb{R}^{k\times k}$.

<u>Fact</u>. $\mathrm{Tr}(ABC) = \mathrm{Tr}(BCA)$.

$$\mathrm{Tr}\left((X\mathcal{U})^T M (X\mathcal{U})\right) = \mathrm{Tr}\left(M(X\mathcal{U})(X\mathcal{U})^T\right)$$
$$= \mathrm{Tr}(M X \mathcal{U}\mathcal{U}^T X^T)$$
$$= \mathrm{Tr}(M X X^T) = \mathrm{Tr}(X^T M X).$$

$$(X\mathcal{U})^T (X\mathcal{U}) = \mathcal{U}^T X^T X \mathcal{U} = \mathcal{U}^T\mathcal{U} = \mathrm{Id}_{k\times k}.$$

$\Rightarrow$ Equivalently,

$$\min_X \mathrm{Tr}\left(X^T \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{bmatrix} X\right) \qquad (\lambda_1 \le \cdots \le \lambda_n).$$

s.t. $X^T X = \mathrm{Id}_{k\times k}$

$\rightarrow$ minimized at $\sum_{i=1}^{k} \lambda_i$

$$\Rightarrow \quad X^* = \begin{bmatrix} | & & | \\ v_1 & \cdots & v_k \\ | & & | \end{bmatrix} \qquad v_i \text{ corresponding eigenvectors}$$

<u>Exercise</u> (Generalized Eigenvector Problem). Let $M, N \in \mathrm{Sym}(\mathbb{R}^{n\times n})$, $N \succeq 0$.

What is solution to eigenvalue problem?

$$Mv = \lambda N v$$
$$\Updownarrow$$
$$\underbrace{N^{-1/2} M N^{1/2}}\ \underbrace{N^{1/2} v} = \lambda \underbrace{N^{1/2} v}$$

Thus, equivalent to solving the eigenvalue problem for $N^{-1/2} M N^{-1/2}$.

<u>Preview of Laplacians</u>

We saw that $L = D - W$.

- hint: we want to solve optimization problems ~~that~~ with constraints
$$Lv = \lambda D v$$
$\Rightarrow$ define $D^{-1}L =: L_{rw}$ the <u>random walk</u> Laplacian, $L_{rw} = I - D^{-1}W$

- hint: eigenvalues will be equivalent to solving eigenvalue problems for
$$D^{-1/2} L D^{-1/2} =: L_{sym} \quad \text{the } \underline{\text{symmetric/normalized Laplacian}}$$
$$= I - D^{-1/2} W D^{-1/2}$$

RETURN TO GRAPH CUT : the probabalistic version

- give graph a random walk interpretation.
  - a random walker transitions from a vertex i to adjacent vertex j w.p.

$$\frac{w_{ij}}{\sum_k w_{ik}} = \frac{w_{ij}}{d_i}$$

Exercise let $\pi^{(t)} = [\pi_1 \cdots \pi_n]$ probability mass of a random walker at time t. After 1 time step, what is the new $\pi^{(t+1)}$?

$$\pi^{(t+1)}_j = \sum_i \pi_i \cdot \frac{w_{ij}}{d_i} = (\pi D^{-1} W)_j$$

Thus, $P = D^{-1}W$ is the transition matrix.

Spectral Clustering

Goal: Cut V into two sets $A \sqcup \bar{A}$. Let random walker on graph. Can we minimize

$$\Pr[A \to \bar{A} \text{ or } \bar{A} \to A]$$

the probability the random walker crosses from A into $\bar{A}$ or vice versa?

Exercise: Compute $\Pr[A \to \bar{A}]$

$$= \frac{\text{weight}(E(A,\bar{A}))}{\text{weight}(E(A,V))} = \frac{\mathbb{1}_A^T L \mathbb{1}_A}{\sum_{i\in A} d_i} = \frac{\mathbb{1}_A^T L \mathbb{1}_A}{\text{vol}(A)}$$

Thus, $\Pr[A \to \bar{A} \text{ or } \bar{A} \to A] = \frac{E(A,\bar{A})}{\text{vol}(A)} + \frac{E(\bar{A},A)}{\text{vol}(\bar{A})}$.

Define. $\text{Ncut}(A_1,\ldots,A_k) = \sum_{i=1}^{k} \frac{E(A_i,\bar{A}_i)}{\text{vol}(A_i)}$.

$V = A_1 \sqcup \cdots \sqcup A_k$

Problem: $\min_A \text{Ncut}(A,\bar{A}), \quad \emptyset \subsetneq A \subsetneq V$.

Fact. This is NP-hard.

Optimization Problem

$$\min_A \quad \underbrace{\frac{\mathbb{1}_A^T L \mathbb{1}_A}{vol(A)}} + \frac{\mathbb{1}_{\bar{A}}^T L \mathbb{1}_{\bar{A}}}{vol(\bar{A})}$$

$$\underbrace{\frac{\mathbb{1}_A^T}{\sqrt{vol(A)}}}_{f_A} L \frac{\mathbb{1}_A}{\sqrt{vol(A)}} \qquad f_{\bar{A}} := \frac{\mathbb{1}_{\bar{A}}^T}{\sqrt{vol(\bar{A})}}$$

Observe:

$$A \sqcup \bar{A} \Leftrightarrow \mathbb{1}_A \perp \mathbb{1}_{\bar{A}}$$

$$f_A^T D f_A = \sum_{i \in A} d_i \left(\frac{1}{\sqrt{vol(A)}}\right)^2 = \frac{1}{vol(A)} \sum_{i \in A} d_i = 1.$$

$$\Rightarrow \begin{bmatrix} f_A & f_{\bar{A}} \end{bmatrix}^T D \begin{bmatrix} f_A & f_{\bar{A}} \end{bmatrix} = Id.$$

Let $X = \begin{bmatrix} f_A & f_{\bar{A}} \end{bmatrix}$

$\Rightarrow$ Optimize:

$$\min_{X \in \{[f_A\ f_{\bar{A}}] : A \subset V\}} Tr(X^T L X) \qquad s.t. \quad X^T D X = Id$$

$\downarrow$ Relax

$$\min_X \quad Tr(X^T L X) \qquad s.t. \quad X^T D X = Id.$$

We saw that the solution corresponds to

$$Lx = \lambda D x$$

$$\updownarrow$$

$$L_{sym}(D^{1/2} x) = \lambda (D^{1/2} x)$$

$\Rightarrow$ if v is eigenvector of L_{sym}, then

$D^{-1/2} v$ is solution to generalized eigenvalue problem.

$$\Rightarrow X = \begin{bmatrix} | & | \\ D^{-1/2} v_1 & D^{-1/2} v_2 \\ | & | \end{bmatrix} \qquad (\text{note: } D^{-1/2} v_1 = \mathbb{1} \text{ for } L_{sym}).$$

Graph Clustering (Shi & Malik 2000) Ncut.

Input $\quad$ n data points $\bar{x}_1, \dots, \bar{x}_n$
$\quad$ W similarity matrix

Output $\quad$ $A_1 \sqcup \cdots \sqcup A_k = V$ partition of $\bar{x}_i$'s

Goal $\quad$ $\min\limits_{X \in \{0,1\}^{n \times k}} \operatorname{Tr}(X^T L X)$ s.t. $X^T D X = Id$

Relax $\quad$ $\min\limits_{X \in \mathbb{R}^{n \times k}} \operatorname{Tr}(X^T L X)$ s.t. $X^T D X = Id$

$\iff$ $L x_i = \lambda_i D x_i$ $\iff$ $x_i = D^{-1/2} v_i$
$v_i \in$ Spectrum (L_{sym}).

Algorithm.

1. Compute the k generalized eigenvectors $x_1, \dots, x_k$
$$L x_i = \lambda_i D x_i .$$

2. Represent $\bar{x}_i$ by ith row of X

$$X = \begin{bmatrix} \text{---} & \text{---} & \text{---} \\ x_1 & \cdots & x_k \\ | & & | \end{bmatrix} \leftarrow \bar{x}_i \text{ represented in } \mathbb{R}^k$$

3. Run k-means++ on the rows of X.

Note. HW1's spring problem is equivalent to the RatioCut problem.

$$\text{RatioCut}: \min_{A_1 \sqcup \cdots \sqcup A_k} \sum \frac{\mathbb{1}_{A_i}^T L \mathbb{1}_{A_i}}{|A_i|} \leftarrow \text{instead of } vol(A_i).$$

Prop. (Guattery & Miller 1998) $\forall c > 0, \exists G$ s.t.
$$c \cdot OPT(G) \leq OPT_{Approx}(G)$$

where $OPT(G) = \min\limits_{\emptyset \subsetneq A \subsetneq V} \text{Ratio Cut}(A, \bar{A})$

$OPT_{Approx} = \min$ Convex Relaxation.

"Cockroach graph" constructive example.

Prop. (Bui & Jones 1992). Approx. problem NP-hard.