Session 8: Linear Algebra for Graphs (& SVD)

Optimization and Computational Linear Algebra for Data Science

SINGULAR VALUE DECOMPOSITION: STRANG CHAP 7

SPECTRAL CLUSTERING: ....

Marylou Gabrié (based on material by Léo Miolane)

Contents

1. Singular Value Decomposition

PCA

- Data matrix $A \in \mathbb{R}^{n \times m}$

  n data points in $\left.\begin{array}{l} m \text{ dimensions} \\ m \text{ features} \end{array}\right.$

- "Covariance matrix" $S = A^{\mathsf{T}} A \in \mathbb{R}^{m \times m}$.

- S is symmetric (positive semi-definite).

- **Spectral Theorem:** there exists an orthonormal basis $v_1, \ldots, v_m$ of $\mathbb{R}^m$ such that the v_i's are eigenvectors of S associated with the eigenvalues $\lambda_1 \geq \cdots \geq \lambda_m \geq 0$ (positive semi-definite).

Definitions: Singular values/vectors

$A \in \mathbb{R}^{n \times m}$

For $i = 1, \ldots, m$:

- We define $\sigma_i = \sqrt{\lambda_i}$, called the i^{th} **singular value** of A .

eigenvalues of A^TA.

Let $r = \text{rank}(A)$ = number of non-zero λ_i's (exercise!).

For $i = 1, \ldots, r$:

eigenvectors v_i of A^TA

- We call $u_i = \frac{1}{\sigma_i} A v_i$ the i^{th} **left singular vector** of A.
- $u_1, \cdots u_r$ are orhonormal.
- If $r < n$, we add $u_{r+1}, \cdots u_n$ such that $u_1, \cdots u_n$ is an orthonormal basis of $\mathbb{R}^n$.

For $i = 1, \ldots, m$:

- Observe that we have $Av_i = \sigma_i u_i$ and $A^T u_i = \sigma_i v_i$.
- We call v_j the i^{th} **right singular vector** of A.

Singular Value decomposition

$A^T A =$

Theorem

Let $A \in \mathbb{R}^{n\times m}$. Then there exists two orthogonal matrices $U \in \mathbb{R}^{n\times n}$ and $V \in \mathbb{R}^{m\times m}$ and a matrix $\Sigma \in \mathbb{R}^{n\times m}$ such that $\Sigma_{1,1} \geq \Sigma_{2,2} \geq \cdots \geq 0$ and $\Sigma_{i,j} = 0$ for $i \neq j$, that verify

$$A = U\Sigma V^\top.$$

$$A = \underbrace{\overset{n\times n}{\begin{bmatrix} | & & | \\ u_1 & \cdots & u_n \\ | & & | \end{bmatrix}}}_{U} \underbrace{\overset{n}{\begin{bmatrix} \sigma_1 & & & & (0) & \\ & \sigma_2 & & & & \\ & & \ddots & & & \\ & & & \sigma_n & & \\ & (0) & & & 0 & \\ & & & & & 0 \; 0 \; 0 \end{bmatrix}}}_{\Sigma} \underbrace{\begin{bmatrix} - & v_1^T & - \\ & \vdots & \\ - & v_m^T & - \end{bmatrix}}_{V^T}$$

Remark: While eigendecomposition is for some square matrices, singular value decomposition exists for all rectangular matrices.

Remarks

- Right singular vectors v_i's are eigenvectors of $A^\top A \in \mathbb{R}^{m\times m}$, with eigenvalues $\lambda_i = \sigma_i^2$.

$$A^T A = (V\Sigma^T U^T)(U\Sigma V^T) = V\Sigma^T\Sigma V^T = V \underbrace{\begin{pmatrix} \sigma_1^2 & & & & & \\ & \ddots & & & & \\ & & \sigma_r^2 & & & \\ & & & 0 & & \\ & & & & \ddots & \\ & & & & & 0 \end{pmatrix}}_{m\times m} V^T$$

- Left singular vectors u_i's are eigenvectors of $AA^\top \in \mathbb{R}^{n\times n}$, with eigenvalues $\lambda_i = \sigma_i^2$.

$$AA^T = U\Sigma V^T V \Sigma U^T = U\Sigma^T\Sigma U^T = U \underbrace{\begin{pmatrix} \sigma_1^2 & & \\ & \ddots & \\ & & \sigma_r^2 \end{pmatrix}}_{n\times n} U^T$$

Rk: A^TA and AA^T have same number of non zero eigenvalues.

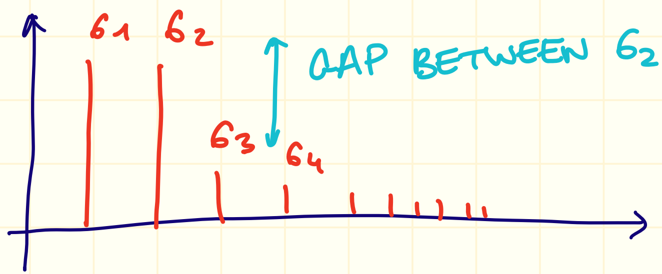
Low-rank approximation

How can we approximate a matrix A by a matrix of "small" rank?

$A \in \mathbb{R}^{n \times m}$

① COMPUTE THE SVD: $A = U \Sigma V^T = U \begin{pmatrix} \sigma_1 & & \\ & \sigma_2 & \\ & & \ddots \end{pmatrix} V^T$

② LOOK AT THE SINGULAR VALUES:



GAP BETWEEN σ_2 and σ_3.

③ TAKE $\tilde{A} = U \begin{pmatrix} \sigma_1 & & & & & & \\ & \sigma_2 & & & & & \\ & & 0 & & & & \\ & & & 0 & & & \\ & & & & 0 & & \\ & & & & & 0 & \\ & & & & & & \ddots \end{pmatrix} V^T$

"Good" rank-2 approximation of A

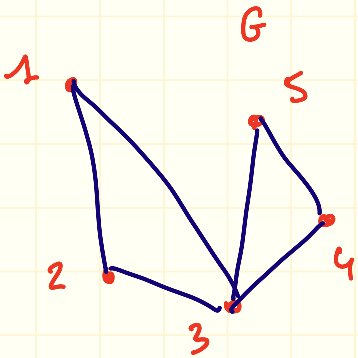
Questions?

Questions?

2. Graphs and Graph Laplacian

2.1 Definitions: Graphs

Consider a graph G made of n **nodes** with some **edges**:



Definition

The **adjacency matrix** A of G is the $n \times n$ matrix with entries

$i \sim j$

$$A_{i,j} = \begin{cases} 1 \text{ if edge between nodes } i \text{ and } j \\ 0 \text{ otherwise} \end{cases}$$

Definition

The **degree matrix** $D \in \mathbb{R}^{n \times n}$ of G is the diagonal matrix with

$$D_{i,i} = \#\{\text{neighbors of } i\} = \deg(i)$$

Graph Laplacian

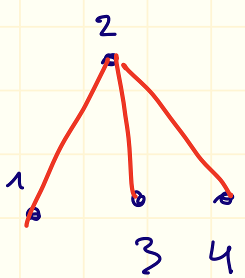
Rk: L, D and A are symmetric matrices.

Definition

The Laplacian matrix of G is defined as

$$L = D - A.$$

Example:



$$A = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix} \qquad D = \begin{pmatrix} 1 & & (0) & \\ & 3 & & \\ (0) & & 1 & \\ & & & 1 \end{pmatrix}$$

$$L = D - A = \begin{pmatrix} 1 & -1 & 0 & 0 \\ -1 & 3 & -1 & -1 \\ 0 & -1 & 1 & 0 \\ 0 & -1 & 0 & 1 \end{pmatrix}.$$

Graph Laplacian

Definition

The Laplacian matrix of G is defined as

$$L = D - A.$$

For all $x \in \mathbb{R}^n$, $\qquad x^\mathsf{T} L x = \sum_{i \sim j} (x_i - x_j)^2.$

$i \sim j \to$ i "connected to" j.

Proof.

$n \times n$

rest is all 0

$$L = \sum_{i \sim j} \underbrace{\begin{array}{cc} \begin{pmatrix} & & & \\ & 1 & -1 & \\ & -1 & 1 & \\ & & & \end{pmatrix} & \begin{array}{l} \leftarrow i \\ \leftarrow j \end{array} \\ \begin{array}{cccc} \quad & \uparrow & \uparrow & \quad \\ & i & j & \end{array} & \end{array}}_{B^{(i,j)}} = \sum_{i \sim j} B^{(i,j)}$$

For any $x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$ in $\mathbb{R}^n$

$$x^\mathsf{T} B^{(ij)} x = x_i^2 + x_j^2 - x_i x_j - x_j x_i$$
$$= (x_i - x_j)^2$$

2.2 Properties of the Laplacian

For all $x \in \mathbb{R}^n$, $\qquad x^\mathsf{T} L x = \sum_{i \sim j} (x_i - x_j)^2. \geq 0$

$\Rightarrow$ L is positive semi-definite (PSD) (by definition)

$\Rightarrow$ $0 \leq \lambda_1 \leq \cdots \leq \lambda_n$ eigenvalues

BY THEOREM in PCA.

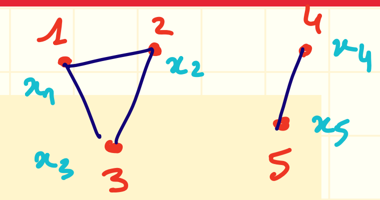
Recall: $\lambda_1 = \min_{\|v\|=1} v^T L v = \min_x \frac{x^T L x}{\|x\|} \geq 0$

$\Rightarrow$ Take $x^T L x$ for $x = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}$ in $\mathbb{R}^n$

$x^T L x = 0 \Rightarrow \lambda_1 = 0$ associated with $\begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$.

2.2 Properties of the Laplacian

G:



2 connect components

For all $x \in \mathbb{R}^n$, $$x^{\mathsf{T}} L x = \sum_{i \sim j} (x_i - x_j)^2.$$

since L is PSD

$$x \in \mathrm{Ker}(L) \iff x^T L x = 0$$

$$\iff x_i = x_j \text{ for all } i \sim j$$

$$\iff x_i = x_j \text{ for all } i \text{ and } j \text{ in the same connected component.}$$

Ex: For G. $x = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 0 \\ 0 \end{pmatrix}$ and $y = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 1 \end{pmatrix}$ are a basis of $\mathrm{Ker}(L)$

$\dim \mathrm{Ker}(L) = 2$.

Algebraic connectivity

$0 = \lambda_1 \leqslant \lambda_2 \leqslant \cdots \leqslant \lambda_n$

Proposition

- The multiplicity of the eigenvalue 0 of L (i.e. the number of i such that $\lambda_i = 0$) is equal to the number of connected components of G.
- In particular, G is connected if and only if $\lambda_2 > 0$.

[no isolated set of nodes]

- λ_2 is sometimes called the «algebraic connectivity» of G and measures somehow how well G is connected.
- From now, we assume that G is connected, i.e. $\lambda_2 > 0$.

3. Application: Spectral graph clustering

3.1 Spectral clustering algorithm

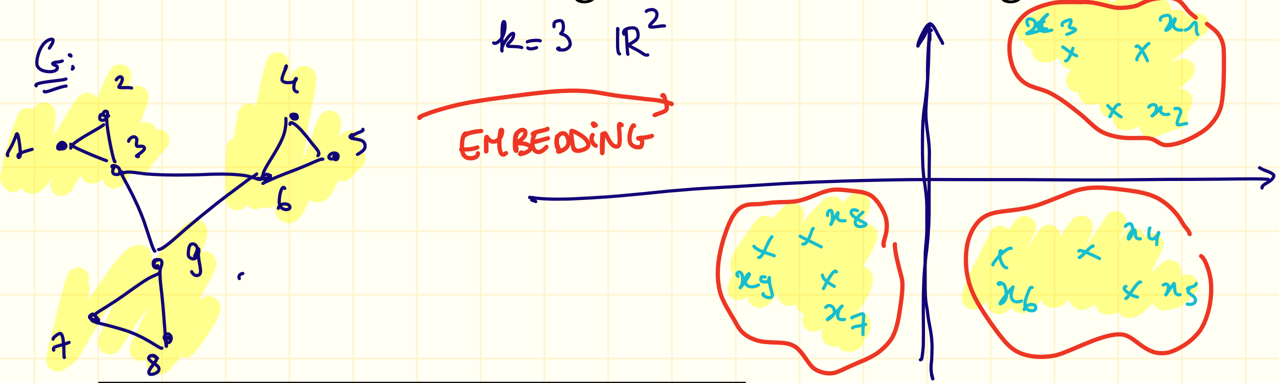
$0 = \lambda_1 < \lambda_2 \leq \dots$

Input: Graph Laplacian L, number of clusters k

1. Compute the first k orthonormal eigenvectors $v_1, \dots, v_k$ of the Laplacian matrix L.

i-th coordinate of eigenvector associated with λ_2.

2. Associate to each node i the vector $x_i = (v_2(i), \dots, v_k(i))$. $\in \mathbb{R}^{k-1}$
3. Cluster the points $x_1, \dots, x_n$ with (for instance) the k-means algorithm.[1]
4. Deduce a clustering of the nodes of the graph.



[1]Chap 13 - Elements of Statistical Learning (Hastie, Tibshirani, and Friedman (2001)) https://web.stanford.edu/~hastie/Papers/ESLII.pdf

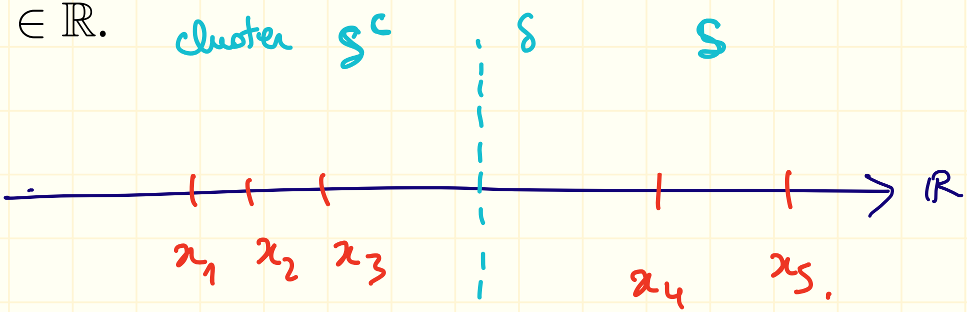
3.2 The case of two groups

For $k = 2$ groups:

1. Compute the second eigenvector v_2 of the Laplacian matrix L.
2. Associate to each node i the number $x_i = v_2(i)$.
3. Cluster the nodes in:

$$S = \{i \mid v_2(i) \geq \delta\} \qquad \text{and} \qquad S^c = \{i \mid v_2(i) < \delta\},$$

for some $\delta \in \mathbb{R}$.



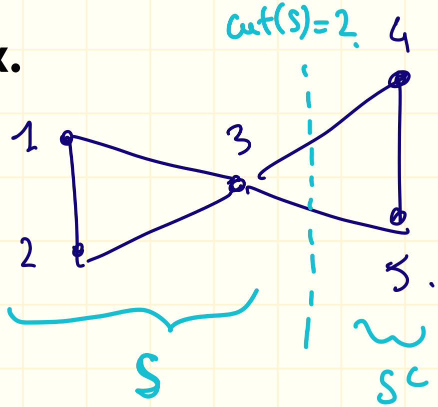
How does this work?

Let $S \subset \{1, 2, \cdots n\}$.

Definition

The cut of S, denoted $\mathrm{cut}(S)$ is defined as the number of edges between S and S^C.

Ex.



$S = \{1, 2, 3\}.$ $\qquad S^C = \{4, 5\}$

$\mathrm{cut}(S) =$

- We encode S by a vector $x \in \{+1, -1\}^n$ defined by

$$x = \begin{pmatrix} 1 \\ 1 \\ 1 \\ -1 \\ -1 \end{pmatrix} \begin{matrix} \left.\vphantom{\begin{matrix}1\\1\\1\end{matrix}}\right\} \text{for nodes in } S \\ \left.\vphantom{\begin{matrix}1\\1\end{matrix}}\right\} \text{for nodes in } S^C \end{matrix}$$

$$x_i = \begin{cases} 1 & \text{if } i \in S \\ -1 & \text{if } i \in S^C. \end{cases}$$

Minimal cut problem

Recall $$x^\top L x = \sum_{i\sim j} (x_i - x_j)^2.$$

Proposition

For $x \in \{+1, -1\}^n$ representing the subset of nodes S,

$$\mathrm{cut}(S) = \frac{1}{4} x^\top L x$$

graph laplacian

Minimal cut problem

Recall $\quad x^\top L x = \sum_{i \sim j} (x_i - x_j)^2.$

Proposition

For $x \in \{+1, -1\}^n$ representing the subset of nodes S,

$$\mathrm{cut}(S) = \frac{1}{4} x^\top L x$$

Goal. Find S (or equivalently $x \in \{+1, -1\}^n$) such that

- $\mathrm{cut}(S)$ is small $\iff x^\top L x$ is small.
- S and S^C have same number of nodes $\quad \# S = \# S^C$

i.e $\quad x \perp \begin{pmatrix} 1 \\ 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}$

« Min-Cut » is NP-Hard

Goal: minimize $x^\mathsf{T} L x$ subject to $\begin{cases} x \in \{-1,1\}^n \\ x \perp (1,\dots,1). \end{cases}$

Basically: Have to try all the $x \in \{-1,+1\}^n$

How many elements?

2^n

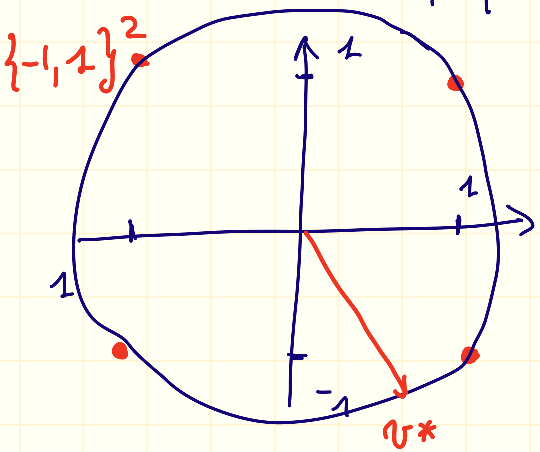
Spectral clustering as a «relaxation»

Idea: We first solve the « relaxed » problem:

$$\text{minimize} \quad v^{\mathsf{T}} L v \quad \text{subject to} \quad \begin{cases} \|v\| = \sqrt{n} \\ v \perp \underbrace{(1, \ldots, 1)}_{v_1}. \end{cases}$$

$n = 2$

$\{ v \mid \|v\| = \sqrt{2} \}$



Solution is $v^* = \sqrt{n}\, v_2$

$v_* = \begin{pmatrix} 0.7 \\ -1.3 \end{pmatrix}$

Define $\tilde{x} \in \{-1, 1\}^n$

$\tilde{x}_i = \begin{cases} +1 & \text{if } v(i) \geq 0 \\ -1 & \text{otherwize} \end{cases}$

$\delta = 0$

[In practice $\tilde{x}$ often leads small cut]

Questions?

Questions?