Session 8: SVD, spectral clustering on graphs

Optimization and Computational Linear Algebra for Data Science

Léo Miolane

Contents

Midterm next week

- Thu. Oct. 29, the questions have to be downloaded from Gradescope between 00:01 AM and 9:59 PM.
- **Duration:** 1 hour and 40 minutes to work on the problems + 20 minutes to scan and upload your work.
- Upload your work **as a single PDF**.
- In case the upload does not work for you, **email me your work**.

Singular Value Decomposition

Singular Value decomposition

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

$$A = U\Sigma V^\mathsf{T}.$$

Remarks

Low-rank approximation

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

Graphs and Graph Laplacian

Graphs

Graph Laplacian

Definition

The Laplacian matrix of G is defined as

$$L = D - A.$$

Graph Laplacian

Definition

The Laplacian matrix of G is defined as

$$L = D - A.$$

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

Properties of the Laplacian

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

Properties of the Laplacian

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

Algebraic connectivity

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 $\lambda_2 > 0$.

- 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. $\lambda_2 > 0$.

Exercise: show that λ_2 increases when one adds edges to G.

Spectral clustering with the Laplacian

Spectral clustering algorithm

Input: Graph Laplacian L, number of clusters k

1. Compute the first k orthonormal eigenvectors $v_1, \ldots, v_k$ of the Laplacian matrix L.
2. Associate to each node i the vector $x_i = (v_2(i), \ldots, v_k(i))$.
3. Cluster the points $x_1, \ldots, x_n$ with (for instance) the k-means algorithm.
4. Deduce a clustering of the nodes of the graph.

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:

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

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

Cut of a partition

Minimal cut problem

« Min-Cut » is NP-Hard

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

Spectral clustering as a «relaxation»

Idea: We first solve the « relaxed » problem:

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

Questions?

Questions?