Graphs in Machine Learning

Michal Valko

Inria Lille - Nord Europe, France

TA: Daniele Calandriello

Partially based on material by: Ulrike von Luxburg, Gary Miller, Doyle & Schnell, Daniel Spielman
Previous Lecture

- similarity graphs
  - different types
  - construction
  - sources of graphs
  - practical considerations
- spectral graph theory
- Laplacians and their properties
  - symmetric and asymmetric normalization
- random walks
- recommendation on a bipartite graph
- resistive networks
  - recommendation score as a resistance?
  - Laplacian and resistive networks
  - resistance distance and random walks
Statistical Machine Learning in Paris!

https://sites.google.com/site/smileinparis/sessions-2016--17

**Speaker:** Isabelle Guyon - LRI (équipe TAO), UPSud  
**Topic:** Network Reconstruction  
**Date:** Monday, October 17, 2016  
**Time:** 13:30 - 14:30 (this is pretty soon)  
**Place:** Institut Henri Poincaré — salle 314
This Lecture

- geometry of the data and the connectivity
- spectral clustering
- manifold learning with Laplacians eigenmaps
- Gaussian random fields and harmonic solution
- graph-based semi-supervised learning and manifold regularization
- transductive learning
- inductive and transductive semi-supervised learning
Next Class: Lab Session

- 24. 10. 2016 by Daniele Calandriello
- cca. 10h30-11h help with setup (optional), 11h-13: TD
- Salle Condorcet
- Download the image and set it up BEFORE the class
- Matlab/Octave
- Short written report (graded)
- All homeworks together account for 40% of the final grade
- Content
  - Graph Construction
  - Test sensitivity to parameters: $\sigma$, $k$, $\varepsilon$
  - Spectral Clustering
  - Spectral Clustering vs. $k$-means
  - Image Segmentation
How to rule the world?

Let’s make Sokovia great again!
How to rule the world?

One reason you're seeing this ad is that

Donald J. Trump wants to reach people who are part of an audience called "Likely To Engage in Politics (Liberal)". This is based on your activity on Facebook and other apps and websites, as well as where you connect to the internet.

There may be other reasons you're seeing this ad, including that Donald J. Trump wants to reach people ages 25 and older who live near Boston, Massachusetts. This is information based on your Facebook profile and where you've connected to the internet.

Was this explanation useful? Yes No
How to rule the world: “AI” is here

https://www.washingtonpost.com/opinions/obama-the-big-data-president/2013/06/14/1d71fe2e-d391-11e2-b05f-3ea3f0e7bb5a_story.html


Talk of Rayid Ghaniy: https://www.youtube.com/watch?v=gDM1GuszM_U
Application of Graphs for ML: Clustering
Application: Clustering - Recap

What do we know about the clustering in general?
- ill defined problem (different tasks $\rightarrow$ different paradigms)
- “I know it when I see it”
- inconsistent (wrt. Kleinberg’s axioms)
- number of clusters $k$ need often be known
- difficult to evaluate

What do we know about $k$-means?
- “hard” version of EM clustering
- sensitive to initialization
- optimizes for compactness
- yet: algorithm-to-go
Spectral Clustering: Cuts on graphs
Spectral Clustering: Cuts on graphs

Defining the cut objective we get the clustering!
**Spectral Clustering: Cuts on graphs**

\[ \text{MinCut: } \text{cut}(A, B) = \sum_{i \in A, j \in B} w_{ij} \]

Can be solved efficiently, but maybe not what we want . . . .

*Are we done?*
Spectral Clustering: Balanced Cuts

Let’s balance the cuts!

**MinCut**

\[
cut(A, B) = \sum_{i \in A, j \in B} w_{ij}
\]

**RatioCut**

\[
\text{RatioCut}(A, B) = \sum_{i \in A, j \in B} w_{ij} \left( \frac{1}{|A|} + \frac{1}{|B|} \right)
\]

**Normalized Cut**

\[
\text{NCut}(A, B) = \sum_{i \in A, j \in B} w_{ij} \left( \frac{1}{\text{vol}(A)} + \frac{1}{\text{vol}(B)} \right)
\]
Spectral Clustering: Balanced Cuts

\[
\text{RatioCut}(A, B) = \text{cut}(A, B) \left( \frac{1}{|A|} + \frac{1}{|B|} \right)
\]

\[
\text{NCut}(A, B) = \text{cut}(A, B) \left( \frac{1}{\text{vol}(A)} + \frac{1}{\text{vol}(B)} \right)
\]

Can we compute this? RatioCut and NCut are NP hard :(
Spectral Clustering: Relaxing Balanced Cuts

Relaxation for (simple) balanced cuts for 2 sets

\[ \min_{A,B} \text{cut}(A, B) \quad \text{s.t.} \quad |A| = |B| \]

Graph function \( f \) for cluster membership: \( f_i = \begin{cases} 1 & \text{if } V_i \in A, \\ -1 & \text{if } V_i \in B. \end{cases} \)

What is the cut value with this definition?

\[ \text{cut}(A, B) = \sum_{i \in A, j \in B} w_{i,j} = \frac{1}{4} \sum_{i,j} w_{i,j}(f_i - f_j)^2 = \frac{1}{2} f^T L f \]

What is the relationship with the smoothness of a graph function?
Spectral Clustering: Relaxing Balanced Cuts

\[
\text{cut}(A, B) = \sum_{i \in A, j \in B} w_{i,j} = \frac{1}{4} \sum_{i,j} w_{i,j} (f_i - f_j)^2 = \frac{1}{2} f^T L f
\]

\[|A| = |B| \implies \sum_i f_i = 0 \implies f \perp 1_N\]

\[\|f\| = \sqrt{N}\]

objective function of spectral clustering

\[
\min_f f^T L f \quad \text{s.t.} \quad f_i = \pm 1, \quad f \perp 1_N, \quad \|f\| = \sqrt{N}
\]

Still NP hard :( \implies Relax even further!

\[f_i = \pm 1 \implies f_i \in \mathbb{R}\]
Spectral Clustering: Relaxing Balanced Cuts

Objective function of spectral clustering

\[
\min_{f} f^T L f \quad \text{s.t.} \quad f_i \in \mathbb{R}, \quad f \perp 1_N, \quad \|f\| = \sqrt{N}
\]

Rayleigh-Ritz theorem

If \( \lambda_1 \leq \cdots \leq \lambda_N \) are the eigenvectors of real symmetric \( L \) then

\[
\lambda_1 = \min_{x \neq 0} \frac{x^T L x}{x^T x} = \min_{x^T x = 1} x^T L x
\]

\[
\lambda_N = \max_{x \neq 0} \frac{x^T L x}{x^T x} = \max_{x^T x = 1} x^T L x
\]

\[
\frac{x^T L x}{x^T x} \equiv \text{Rayleigh quotient}
\]

How can we use it?
Spectral Clustering: Relaxing Balanced Cuts

Objective function of spectral clustering

\[
\min_{f} f^T L f \quad \text{s.t.} \quad f_i \in \mathbb{R}, \quad f \perp 1_N, \quad \|f\| = \sqrt{N}
\]

Generalized Rayleigh-Ritz theorem (Courant-Fischer-Weyl)

If \( \lambda_1 \leq \cdots \leq \lambda_N \) are the eigenvectors of real symmetric \( L \) and \( v_1, \ldots, v_N \) the corresponding orthogonal eigenvalues, then for \( k = 1 : N - 1 \)

\[
\lambda_{k+1} = \min_{x \neq 0, x \perp v_1, \ldots, v_k} \frac{x^T L x}{x^T x} = \min_{x_1=1, x \perp v_1, \ldots, v_k} x^T L x
\]

\[
\lambda_{N-k} = \max_{x \neq 0, x \perp v_n, \ldots, v_{N-k+1}} \frac{x^T L x}{x^T x} = \max_{x_1=1, x \perp v_N, \ldots, v_{N-k+1}} x^T L x
\]
Rayleigh-Ritz theorem: Quick and dirty proof

When we reach the extreme points?

\[
\frac{\partial}{\partial x} \left( \frac{x^T L x}{x^T x} \right) = \frac{\partial}{\partial x} \left( \frac{f(x)}{g(x)} \right) = 0 \iff f'(x)g(x) = f(x)g'(x)
\]

By matrix calculus (or just calculus):

\[
\frac{\partial x^T L x}{\partial x} = 2Lx \quad \text{and} \quad \frac{\partial x^T x}{\partial x} = 2x
\]

When \( f'(x)g(x) = f(x)g'(x) \)?

\[
Lx (x^T x) = (x^T L x)x \iff Lx = \frac{x^T L x}{x^T x} x \iff Lx = \lambda x
\]

Conclusion: Extremes are the eigenvectors with their eigenvalues
Spectral Clustering: Relaxing Balanced Cuts

Objective function of spectral clustering:

\[
\min_{f} f^T L f \quad \text{s.t.} \quad f_i \in \mathbb{R}, \quad f \perp 1_N, \quad \|f\| = \sqrt{N}
\]

Solution: second eigenvector

How do we get the clustering?

The solution may not be integral.

What to do?

\[
\text{cluster}_i = \begin{cases} 
1 & \text{if } f_i \geq 0, \\
-1 & \text{if } f_i < 0.
\end{cases}
\]

Works but this heuristics is often too simple. In practice, cluster \( f \) using \( k \)-means to get \( \{C_i\}_i \) and assign:

\[
\text{cluster}_i = \begin{cases} 
1 & \text{if } i \in C_1, \\
-1 & \text{if } i \in C_{-1}.
\end{cases}
\]
Spectral Clustering: Approximating RatioCut

Wait, but we did not care about approximating mincut!

**RatioCut**

$$\text{RatioCut}(A, B) = \sum_{i \in A, j \in B} w_{ij} \left( \frac{1}{|A|} + \frac{1}{|B|} \right)$$

Define graph function \( f \) for cluster membership of RatioCut:

$$f_i = \begin{cases} \sqrt{\frac{|B|}{|A|}} & \text{if } V_i \in A, \\ -\sqrt{\frac{|A|}{|B|}} & \text{if } V_i \in B. \end{cases}$$

$$f^T L f = \frac{1}{2} \sum_{i,j} w_{i,j} (f_i - f_j)^2 = (|A| + |B|) \text{RatioCut}(A, B)$$
Spectral Clustering: Approximating RatioCut

Define graph function $f$ for cluster membership of RatioCut:

$$f_i = \begin{cases} 
\sqrt{\frac{|B|}{|A|}} & \text{if } V_i \in A, \\
-\sqrt{\frac{|A|}{|B|}} & \text{if } V_i \in B.
\end{cases}$$

$$\sum_i f_i = 0$$

$$\sum_i f_i^2 = N$$

objective function of spectral clustering (same - it’s magic!)

$$\min_{f} f^T L f \quad \text{s.t.} \quad f_i \in \mathbb{R}, \quad f \perp 1_N, \quad \|f\| = \sqrt{N}$$
Spectral Clustering: Approximating NCut

Normalized Cut

\[
\text{NCut}(A, B) = \sum_{i \in A, j \in B} w_{ij} \left( \frac{1}{\text{vol}(A)} + \frac{1}{\text{vol}(B)} \right)
\]

Define graph function \( f \) for cluster membership of NCut:

\[
f_i = \begin{cases} 
\sqrt{\frac{\text{vol}(A)}{\text{vol}(B)}} & \text{if } V_i \in A, \\
-\sqrt{\frac{\text{vol}(B)}{\text{vol}(A)}} & \text{if } V_i \in B.
\end{cases}
\]

\[
(Df)^T 1_n = 0 \quad f^T Df = \text{vol}(\mathcal{V}) \quad f^T Lf = \text{vol}(\mathcal{V}) \text{NCut}(A, B)
\]

objective function of spectral clustering (NCut)

\[
\min_f f^T Lf \quad \text{s.t.} \quad f_i \in \mathbb{R}, \quad Df \perp 1_N, \quad f^T Df = \text{vol}(\mathcal{V})
\]
Spectral Clustering: Approximating NCut

Objective function of spectral clustering (NCut)
\[
\min_{f} f^T L f \quad \text{s.t.} \quad f_i \in \mathbb{R}, \quad Df \perp 1_N, \quad f^T D f = \text{vol}(\mathcal{V})
\]

Can we apply Rayleigh-Ritz now? Define \( w = D^{1/2} f \)

Objective function of spectral clustering (NCut)
\[
\min_{w} w^T D^{-1/2} L D^{-1/2} w \quad \text{s.t.} \quad w_i \in \mathbb{R}, \quad w \perp D^{1/2} 1_N, \quad \|w\|^2 = \text{vol}(\mathcal{V})
\]

Objective function of spectral clustering (NCut)
\[
\min_{w} w^T L_{\text{sym}} w \quad \text{s.t.} \quad w_i \in \mathbb{R}, \quad w \perp v_1, L_{\text{sym}}, \quad \|w\|^2 = \text{vol}(\mathcal{V})
\]
Objective function of spectral clustering (NCut)

\[
\min_{\mathbf{w}} \mathbf{w}^T \mathbf{L}_{\text{sym}} \mathbf{w} \quad \text{s.t.} \quad w_i \in \mathbb{R}, \quad \mathbf{w} \perp \mathbf{v}_1, \mathbf{L}_{\text{sym}}, \quad \|\mathbf{w}\| = \text{vol}(\mathcal{V})
\]

Solution by Rayleigh-Ritz?

\[
\mathbf{w} = \mathbf{v}_2, \mathbf{L}_{\text{sym}} \quad \mathbf{f} = \mathbf{D}^{-1/2} \mathbf{w}
\]

\(\mathbf{f}\) is the second eigenvector of \(\mathbf{L}_{\text{rw}}\)!

tl;dr: Get the second eigenvector of \(\mathbf{L}/\mathbf{L}_{\text{rw}}\) for RatioCut/NCut.
**Spectral Clustering: Approximation**

These are all approximations. How bad can they be?

**Example:** cockroach graphs

No efficient approximation exist. Other relaxations possible.

https://www.cs.cmu.edu/~glmiller/Publications/Papers/GuMi95.pdf
Spectral Clustering: 1D Example

**Elbow rule/EigenGap heuristic** for number of clusters
Spectral Clustering: Understanding

**Compactness vs. Connectivity**

For which kind of data we can use one vs. the other?

Any disadvantages of spectral clustering?
Spectral Clustering: 1D Example - Histogram

http://www.informatik.uni-hamburg.de/ML/contents/people/luxburg/publications/Luxburg07_tutorial.pdf
Spectral Clustering: 1D Example - Eigenvectors

![Diagram showing eigenvectors and eigenvalues for spectral clustering.](image)
Spectral Clustering: Bibliography


Manifold Learning: Recap

**Problem:** Definition reduction/manifold learning

Given \( \{x_i\}_{i=1}^{N} \) from \( \mathbb{R}^d \) find \( \{y_i\}_{i=1}^{N} \) in \( \mathbb{R}^m \), where \( m \ll d \).

▶ What do we know about the **dimensionality reduction**
  ▶ representation/visualization (2D or 3D)
  ▶ an old example: globe to a map
  ▶ often assuming \( \mathcal{M} \subset \mathbb{R}^d \)
  ▶ feature extraction
  ▶ linear vs. nonlinear dimensionality reduction

▶ What do we know about linear vs. nonlinear methods?
  ▶ linear: ICA, PCA, SVD, ...
  ▶ nonlinear often preserve only **local** distances
Manifold Learning: Linear vs. Non-linear
Manifold Learning: Preserving (just) local distances

\[ d(y_i, y_j) = d(x_i, x_j) \quad \text{only if} \quad d(x_i, x_j) \quad \text{is small} \]

\[ \min \sum_{ij} w_{ij} ||y_i - y_j||^2 \]

Looks familiar?
Manifold Learning: Laplacian Eigenmaps

**Step 1:** Solve generalized eigenproblem:

\[ Lf = \lambda Df \]

**Step 2:** Assign \( m \) new coordinates:

\[ x_i \mapsto (f_2(i), \ldots, f_{m+1}(i)) \]

**Note 1:** we need to get \( m + 1 \) smallest eigenvectors

**Note 2:** \( f_1 \) is useless

http://web.cse.ohio-state.edu/~mbelkin/papers/LEM_NC_03.pdf
Manifold Learning: Laplacian Eigenmaps to 1D

Laplacian Eigenmaps 1D objective

\[
\min_{f} f^{T}L f \quad \text{s.t.} \quad f_i \in \mathbb{R}, \quad f^{T}D1 = 0, \quad f^{T}Df = 1
\]

The meaning of the constraints is similar as for spectral clustering:

\[f^{T}Df = 1\] is for scaling

\[f^{T}D1 = 0\] is to not get \(v_1\)

What is the solution?
Manifold Learning: Example

Semi-supervised learning: How is it possible?

This is how children learn! hypothesis
Semi-supervised learning (SSL)

SSL problem: definition

Given \( \{x_i\}_{i=1}^N \) from \( \mathbb{R}^d \) and \( \{y_i\}_{i=1}^{n_l} \), with \( n_l \ll N \), find \( \{y_i\}_{i=n_l+1}^n \) (transductive) or find \( f \) predicting \( y \) well beyond that (inductive).

Some facts about SSL

- assumes that the unlabeled data is useful
- works with data geometry assumptions
  - cluster assumption — low-density separation
  - manifold assumption
  - smoothness assumptions, generative models, ...
- now it helps now, now it does not (sic)
  - provable cases when it helps
- inductive or transductive/out-of-sample extension

SSL: Self-Training

(a) Iteration 1

(b) Iteration 25

(c) Iteration 74

(d) Final labeling of all instances
Input: $\mathcal{L} = \{x_i, y_i\}_{i=1}^{n_l}$ and $\mathcal{U} = \{x_i\}_{i=n_l+1}^{N}$

Repeat:
- train $f$ using $\mathcal{L}$
- apply $f$ to (some) $\mathcal{U}$ and add them to $\mathcal{L}$

What are the properties of self-training?
- it's a wrapper method
- heavily depends on the internal classifier
- some theory exists for specific classifiers
- nobody uses it anymore
- errors propagate (unless the clusters are well separated)
SSL: Self-Training: Bad Case

![Graphs showing height vs. weight with outliers in different scenarios](image-url)
Michal Valko
michal.valko@inria.fr
ENS Paris-Saclay, MVA 2016/2017
SequeL team, Inria Lille — Nord Europe
https://team.inria.fr/sequel/