Advanced Machine Learning : from Theory to Practice
Lecture 6
Graphs in Machine Learning

F. d'Alché-Buc and E. Le Pennec

Fall 2016

- Graphs as data
  - web, social networks, biological networks, wireless network, molecules, sensor network (IOT)...
  - Recommendation system, Link prediction, Activity prediction
- Data as graphs Today's course
  - data defined by a similarity or affinity matrix
  - use elements of graph theory to achieve clustering, semi-supervised learning, transductive learning

Application to :

- Clustering in unsupervised learning
- Semi-supervised and transductive learning

Unlabeled data

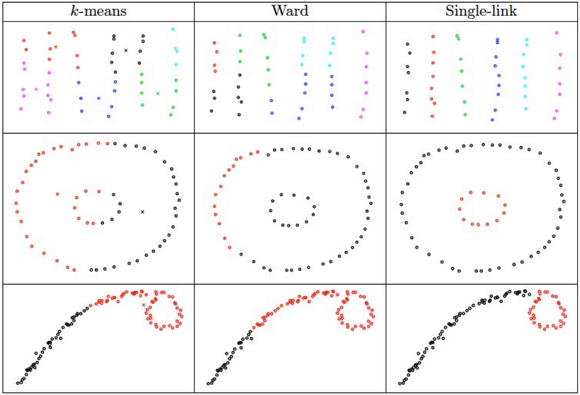
- Available data are unlabeled : documents, webpages, clients database ...
- Labeling data is expensive and requires some expertise

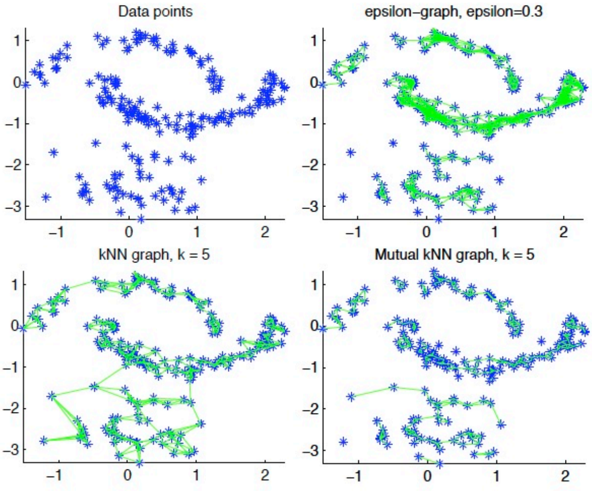
Learning from unlabeled data

- Modeling probability distribution $\rightarrow$ graphical models
- Dimension reduction $\rightarrow$ pre-processing for pattern recognition
- **Clustering** : group data into homogeneous clusters $\rightarrow$ organize your data, make easier access to them, pre and post processing, application in segmentation, document retrieval, bioinformatics ...

Clustering

Different clusterings





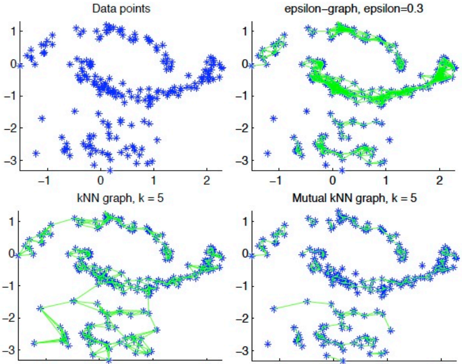
Credits :

Image : U. V .Luxburg.

- Data $x_1, \ldots, x_n$ with their similarity values $s_{ij} \geq 0$ or with their distance d_{ij} values
- Build a graph $G = (V, E)$
- V : set of vertices. A vertex v_i corresponds to data x_i
- E : set of edges. An edge links two nodes if x_i and x_j are close according to the ε-graph method or the k-nn method
- W : adjacency matrix $=$ binary symmetric matrix
- Definition : $w_{ij} = 1$ if there is an edge between node v_i and node v_j, 0 otherwise.

Several ways to construct it :

- ε-graph : connect all points whose pairwise distance is at most ε (alt. whose pairwise similarity is at least ε
- k-nearest-neighbor-graph : connect v_i and v_j if x_i is among the k-nearest-neighbors of x_j OR x_i is among the k-nearest-neighbours of x_j

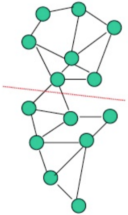


Notations : A and B are two disjoint subsets of the nodes set V that form a partition

- $cut(A, B) = \sum_{t \in A, u \in B} w_{t,u}$
- $vol(A) = \sum_{t \in A, u \in V} w_{t,u}$
- $|A| =$ nb of edges

Mincut problem

- $Cut(A, \bar{A}) = \sum_{i \in A, j \in \bar{A}} w_{ij}$
- Let $f_i \in \{-1, 1\}$ be the index class of x_i
- Clustering :=Find $(f_1, \ldots, f_n) \in \{-1, 1\}$ such that $Cut(A, \bar{A})$ is minimized.



For sake of simplicity : $B = \bar{A}$. Ratiocut :

$$\text{Ratiocut}(A, B) = \frac{cut(A, B)}{|A|} + \frac{cut(B, A)}{|B|}$$

Normalized cut

$$\text{Ncut}(A, B) = \frac{cut(A, B)}{vol(A)} + \frac{cut(B, A)}{vol(B)}$$

Some references :

- Courses/slides : Dan Spielman (Godel prize in 2015), Yale, ▸ Link
- Spectral Graph Theory, Fan R. K. Chung, Published by AMS , 1997, ▸ Link

Definitions

- W matrix : adjacency matrix
- Degree matrix D : $d_{ii} = \sum_j w_{ij}$, if $i \neq j$, $d_{ij} = 0$
- Unnormalized Graph Laplacian : $L = D - W$
- Normalized Graph Laplacians : $L_{sym} = D^{-1/2}(D - W)D^{-1/2}$, $L_{rw} = D^{-1}(D - W)$.

Eigenvalue/eigenvectors

1. L is a symmetric and positive semi-definite matrix
2. Vector 1_n is a eigenvector of L with eigenvalue 0.

Proof :

1.

$$\begin{aligned}
f^T L f &= f^T (D - W) f \\
&= f^T D f - f^T W f \\
&= \sum_i d_i f_i^2 - \sum_{ij} w_{ij} f_i f_j \\
&= \frac{1}{2} \left(\sum_i d_i f_i^2 - 2 \sum_{ij} w_{ij} f_i f_j + \sum_j d_j f_j^2 \right) \\
&= \frac{1}{2} \sum_{i,j} w_{ij} (f_i - f_j)^2
\end{aligned}$$

2. We notice that : $(D - W) 1_n = 0$.

Connected components

Proposition

- The multiplicity of the smallest eigenvalue (0) of L is the number of connected components in the graph

$$L = \begin{pmatrix} L_1 & & & \\ & L_2 & & \\ & & \ddots & \\ & & & L_k \end{pmatrix}$$

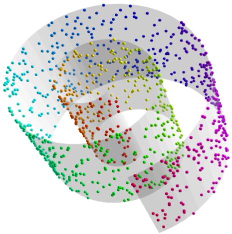
The normalized Laplacians satisfy :

1. For every $f \in \mathbb{R}^n$, $f^T L_{sym} f = \frac{1}{2}\sum_{i,j=1}^{n} w_{ij}(\frac{f_i}{\sqrt{d_i}} - \frac{f_j}{\sqrt{d_j}})^2$.$\lambda$ is an eigenvalue of L_{rw} with eigenvector u iff λ is an eigenvalue of L_{sym} with eigenvector : $v = D^{1/2}u$.

3. λ is an eigenvalue of L_{rw} with eigenvector u iff λ and u solve the generalized eigen problem : $Lu = \lambda Du$.
4. 0 is an eigenvalue of L_{rw} with the constant vector 1_n. 0 is an eigenvalue of L_{sym} with eigenvector $D^{1/2}1$.

A function $f : V \to \mathbb{R}$.
Smoothness of the graph function :

$$\|f\|_L^2 = f^T L f = \frac{1}{2} \sum_{i,j} w_{ij}(f_i - f_j)^2$$



Manifold $\mathcal{M}$: topological space that locally resembles Euclidean space near each point.
More generally, measure of the smoothness of a function on a manifold :

$$\|f\|_{\mathcal{M}}^2 = \int_{\mathcal{M}} ||\nabla_{\mathcal{M}} f(x)\|^2 p(x)dx$$

- $f_i, i = 1, \ldots, n$: membership of data i to clusters
- $f_i = 1$ if $x_i \in A$, otherwise -1 (in B)

Balanced Mincut problem

Find $f \in \{-1, 1\}^n$ that minimizes $J(f) = \sum_{i \in A, j \in B} w_{ij}$ such that $|A| = |B|$

Notice that $|A| = |B| \iff \sum_{i=1}^{n} f_i = 0$ (as many 1's than -1's).
$\sum_{i=1}^{n} f_i = 0 \iff f \perp 1_n$.

$$\begin{aligned} J(f) &= \sum_{i \in A, j \in B}^{n} w_{ij} = \frac{1}{8} \sum_{i,j} w_{ij}(f_i - f_j)^2 \\ &= \frac{1}{8} \sum_{i,j} w_{ij}(f_i^2 + f_j^2 - 2 f_i f_j) \\ &= \frac{1}{4} f^T (D - W) f \end{aligned}$$

Constraints :

- Avoiding trivial solution : $f \perp 1_n$
- Controlling the complexity of f (ℓ_2 regularization) : $\sum_i f_i^2 = n$

Now $f \in \mathbb{R}^n$

$\min_{f \in \mathbb{R}^n} f^T L f$
subject to : $f \perp 1$, $||f|| = \sqrt{n}$

First Order Optimality Conditions to solve the optimization problem :

- Equality constraint of the form : $g(x) = b$, insert $+\lambda(b - g(x))$ into the Lagrangian function
- Build the Lagrangian : $\mathcal{L}(f, \lambda) = f^T L f + \lambda(n - ||f||^2)$
- at the minimum, we have : $\frac{\partial L(f,\lambda)}{\partial f} = 2Lf - 2\lambda f = 0$

If we solve this eigenvector problem and take the second eigenvector, $\hat{f}$, we get $\hat{f} \perp 1$, 1 being the first eigenvector. To get final integer values : threshold the values of f to get discrete values1 and -1 OR use 2-means (better).

Algorithm

- Solve the previous relaxed problem $\rightarrow$ take the k first eigenvectors (note that you can omit $1_n = v_1$ in balanced min cut)
- Represent your data in the new space spanned by these k vectors : form the matrix V with the v_k's as column vectors
- Each row of V represents an individual
- Apply k-means in the k-dimensional space

- Relaxation of Ratiocut
- Relaxation of Mincut

$$\begin{aligned} Ratiocut(A, B) &= \frac{cut(A, B)}{|A|} + \frac{cut(B, A)}{|B|} \\ &= cut(A, B)(\frac{1}{|A|} + \frac{1}{|B|}) \end{aligned}$$

Define (1) :

if $v_i \in A,\ f_i = \sqrt{\frac{|B|}{|A|}}$.

if $v_i \in B\ ,\ f_i = -\frac{\sqrt{|A|}}{\sqrt{|B|}}$

$$
\begin{aligned}
f^T L f &= \frac{1}{2}\sum_{i,j} w_{ij}(f_i - f_j)^2 \\
&= \frac{1}{2}\sum_{i\in A, j\in B} w_{ij}\left(\sqrt{\frac{|B|}{|A|}} + \sqrt{\frac{|A|}{|B|}}\right)^2 + \frac{1}{2}\sum_{i\in B, j\in A}\left(-\sqrt{\frac{|A|}{|B|}} - \sqrt{\frac{|B|}{|A|}}\right)^2 \\
&= cut(A,B)\left(\frac{|B|}{|A|} + \frac{|A|}{|B|} + 2\right) \\
&= cut(A,B)\left(\frac{|A|+|B|}{|A|} + \frac{|A|+|B|}{|B|}\right) \\
&= |V|\, ratiocut(A,B)
\end{aligned}
$$

We have also :

- f as defined for Ratiocut satisfies : $\sum_i f_i = 0$
- $\|f\|^2 = n$

Altogether :

Approximating Ratiocut

$\min_f f^T L f$, s.t. $f \perp 1$, $||f||^2 = n$

- Normalized cut (avoid isolated subset) :
  $Ncut(A, B) = \frac{cut(A,B)}{vol(A)} + \frac{cut(B,A)}{vol(B)}$
- $f_i = \sqrt{\frac{vol(B)}{vol(A)}}, if\, v_i \in A, \sqrt{\frac{vol(A)}{vol(B)}}, if\, v_i \in B.$
- Notice that :
  - $vol(V) = f^T Df.$
  - $(Df)^T 1 = 0$
  - $f^T Lf = vol(V) Ncut(A, B)$

$\min_{f \in \mathbb{R}^n} \frac{f^T L f}{f^T D f}$
subject to : $f^T D 1_n = 0$

$\min_{f \in \mathbb{R}^n} \frac{f^T L f}{f^T D f}$
subject to : $f^T D 1_n = 0$

Solve the generalized eigenvalue problem :
$(D - W)f = \lambda D f$ which can be re-written as
$D^{-\frac{1}{2}}(D - W)D^{-\frac{1}{2}} z = \lambda z$
with $z = D^{-\frac{1}{2}} f$.

The problem boils down to find second eigenvector of L_{sym}.

- Importance of the initial graph : several ways to construct it (k-neighbors)
- Able to extract clusters on a manifold
- Consistency (U. Von Luxburg)
- Stability
- Model selection : eigengap
- High complexity in time

Spectral clustering
Eigengap heuristic

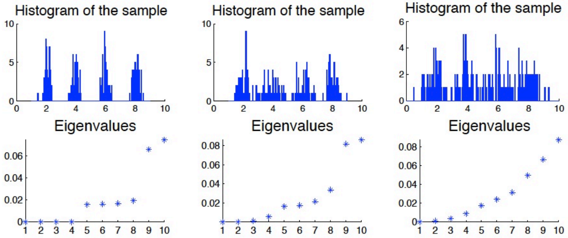
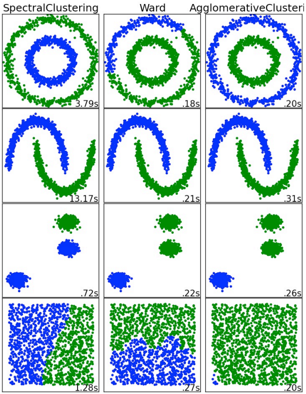


Figure 4: Three data sets, and the smallest 10 eigenvalues of L_{rw}.

- Source Tutorial U. Von Luxburg



- Figure from scikitkearn :

Exercise ;

- Elaborate ideas to scale up spectral clustering

References

- Weiss, Segmentation using eigenvectors, Int. Conf. Computer visions, 1999.
- Shi, Malik, IEEE PAMI, 2000 and Ng, Jordan, Spectral Clustering, 2001.
- Large scale : Fast Approximate Spectral Clustering, Yan, Luang, Jordan.