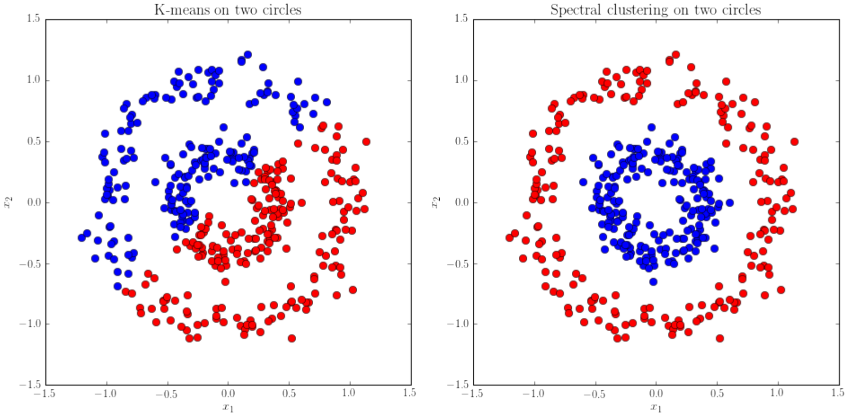
SC4/SM8 Advanced Topics in Statistical Machine Learning
Spectral Clustering

Dino Sejdinovic
Department of Statistics
Oxford

Slides and other materials available at:
http://www.stats.ox.ac.uk/~sejdinov/atsml/

Nonlinear cluster structures



K-means algorithm will often fail when applied to data with elongated or non-convex cluster structures.

Clustering and Graph Cuts

- Construct a weighted undirected **similarity** graph $G = (\{1, \dots, n\}, \mathbf{W})$, where vertices correspond to data items and $\mathbf{W}$ is the matrix of edge weights corresponding to pairwise item similarities.
- Partition the graph vertices into $C_1, C_2, \dots, C_K$ to minimize the **graph cut**.
- The unnormalized **graph cut** across clusters is given by

$$\text{cut}\,(C_1, \dots, C_K) = \sum_{k=1}^{K} \text{cut}(C_k, \bar{C}_k),$$

  where $\bar{C}_k$ is the complement of C_k and $\text{cut}(A, B) = \sum_{i \in A, j \in B} w_{ij}$ is the sum of the weights separating vertex subset A from the vertex subset B, where A and B are disjoint.
- Typically results with singleton clusters, so one needs to balance the cuts by the cluster sizes in the partition. One approach is to consider the notion of “ratio cut"

$$\text{ratio-cut}\,(C_1, \dots, C_K) = \sum_{k=1}^{K} \frac{\text{cut}(C_k, \bar{C}_k)}{|C_k|}.$$

Graph Laplacian

The **(unnormalized) Laplacian** of a graph $G = (\{1, \ldots, n\}, \mathbf{W})$ is an $n \times n$ matrix given by

$$\mathbf{L} = \mathbf{D} - \mathbf{W},$$

where $\mathbf{D}$ is a diagonal matrix with $\mathbf{D}_{ii} = \mathrm{deg}(i)$, and $\mathrm{deg}(i)$ denotes the **degree** of vertex i defined as

$$\mathrm{deg}(i) = \sum_{j=1}^{n} w_{ij}.$$

- Laplacian always has the column vector $\mathbf{1}$ as an eigenvector with eigenvalue 0 (since all rows sum to zero)
- (**exercise**) Laplacian is a positive semi-definite matrix so all the eigenvalues are non-negative.

Laplacian and Ratio Cuts

Lemma

For a given partition $C_1, C_2, \dots, C_K$ *define the column vectors* $h_k \in \mathbb{R}^n$ *as*

$$h_{k,i} = \frac{1}{\sqrt{|C_k|}} \mathbf{1}_{\{i \in C_k\}}.$$

Then

$$\textit{ratio-cut}\,(C_1, \dots, C_K) = \sum_{k=1}^{K} h_k^\top \mathbf{L} h_k. \tag{1}$$

To minimize the ratio cut, search for orthonormal vectors h_k with entries either 0 or $1/\sqrt{|C_k|}$ which minimize the RHS in (1).
Equivalent to integer programming so computationally hard.

Laplacian and Ratio Cuts

Lemma

For a given partition $C_1, C_2, \dots, C_K$ *define the column vectors* $h_k \in \mathbb{R}^n$ *as*

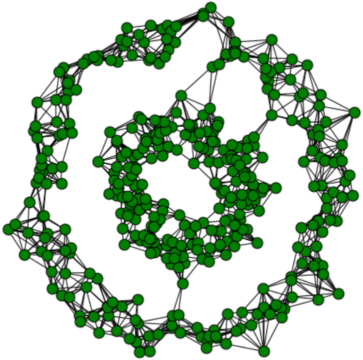
$$h_{k,i} = \frac{1}{\sqrt{|C_k|}} \mathbf{1}_{\{i \in C_k\}}.$$

Then

$$\textit{ratio-cut}\,(C_1, \dots, C_K) = \sum_{k=1}^{K} h_k^\top \mathbf{L} h_k. \tag{1}$$

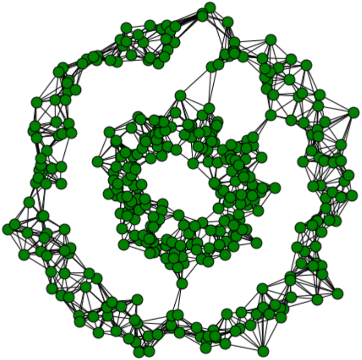
Relaxation: Search for **any collection of orthonormal vectors** h_k in $\mathbb{R}^n$ that minimize RHS in (1) – which corresponds to the eigendecomposition of the Laplacian.

Laplacian and Connected Components



If the original graph is disconnected, in addition to $\mathbf{1}$, there would be other 0-eigenvectors of $\mathbf{L}$, corresponding to the indicators of the connected components of the graph (**Murphy** – Theorem 25.4.1).

Laplacian and Connected Components



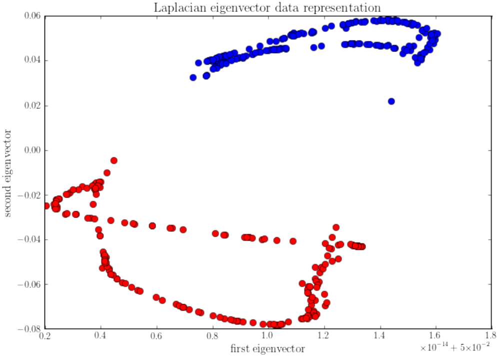
Spectral clustering treats the constructed graph as a “small perturbation" of a disconnected graph.

Eigenvectors as dimensionality reduction

Spectral Clustering. Eigendecompose $\mathbf{L}$ and take the K eigenvectors corresponding to the K smallest eigenvalues – this gives a new "data matrix"

$$\mathbf{Z} = [u_1, \ldots, u_K] \in \mathbb{R}^{n \times K}$$

on which we can apply a more conventional clustering algorithm, such as K-means.



Further reading

- von Luxburg: Tutorial on Spectral Clustering
- Clustering on scikit-learn

Laplacian matrices for Manifold Regularization

- Manifold regularization [Belkin et al, 2006] is useful in semisupervised learning. Assuming we have a labelled set of examples $\{(x_i, y_i)_{i=1}^n\}$ and an unlabelled set of inputs $\{x_{n+i}\}_{i=1}^u$, we form an $(n+u) \times (n+u)$ Laplacian matrix $\mathbf{L}$ and consider the ERM with an additional (**intrinsic**) regularizer

$$\mathbf{f}^\top \mathbf{L}\mathbf{f} = \frac{1}{2}\sum_{i=1}^{n+u}\sum_{j=1}^{n+u} w_{ij}(f(x_i) - f(x_j))^2$$

  for the vector $\mathbf{f} = [f(x_1), \ldots, f(x_{n+u})]^\top$ of function values on all inputs

$$\min_{f\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} L(y_i, f(x_i)) + \lambda\|f\|_{\mathcal{H}}^2 + \lambda_M \mathbf{f}^\top \mathbf{L}\mathbf{f}$$

- The additional regularizer penalizes large differences between function values at the neighbouring vertices.
- If $\mathcal{H} = \mathcal{H}_k$ is an RKHS for a kernel k, representer theorem still applies, but with the solution spanned using **all** inputs:

$$f_\star = \sum_{i=1}^{n+u} \alpha_i k(x_i, \cdot).$$