MF TUTORIAL PART 3B

Clustering -- Nonconvex

Sibylle Hess and Michiel Hochstenbach

THE MANY OBJECTIVES OF K-MEANS

The following objectives are equivalent to (KM):

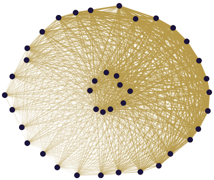
$$\min_{Y,X} \|D - YX^\top\|^2 \qquad \text{s.t. } X \in \mathbb{R}^{n\times k}, Y \in \mathbb{1}^{m\times k}$$

$$\max_{Y} \operatorname{tr}(Z^\top DD^\top Z) \qquad \text{s.t. } Z = Y(Y^\top Y)^{-1/2}, Y \in \mathbb{1}^{m\times k}$$

$$\max_{Y} \sum_{c=1}^{k} \frac{Y_{\cdot c}^\top DD^\top Y_{\cdot c}}{|Y_{\cdot c}|} \qquad \text{s.t. } Y \in \mathbb{1}^{m\times k}$$

The matrix $W = DD^\top$ has an interpretation as similarity matrix

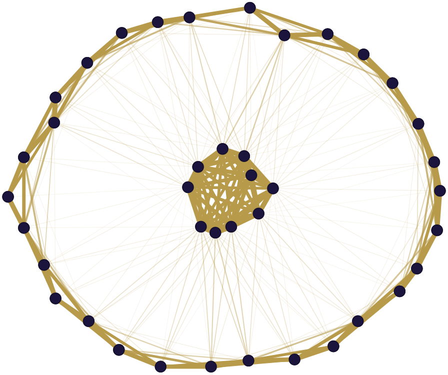
What if we use a **KERNEL MATRIX** $W = \phi(D)\phi(D)^\top$? Can we then cluster **NONCONVEX SHAPES**?



The trick in nonconvex clustering is to reflect **LOCAL SIMILARITIES** in the similarity matrix.

RBF KERNEL SIMILARITIES

$$W_{jl} = \exp\left(-\epsilon \|D_{j\cdot} - D_{l\cdot}\|^2\right)$$



The many objectives of kernel k-means

Given a kernel matrix K (p.s.d. and symmetric) and its symmetric decomposition $K = UU^\top$, the following objectives are **EQUIVALENT**:

$$\min_Y \|U - YX^\top\|^2 \qquad \text{s.t. } X \in \mathbb{R}^{n\times r}, Y \in \mathbb{1}^{m\times r}$$

$$\max_Y \operatorname{tr}(Z^\top K Z) \qquad \text{s.t. } Z = Y(Y^\top Y)^{-1/2}, Y \in \mathbb{1}^{m\times r}$$

$$\max_Y \sum_{s=1}^{r} \frac{Y_{\cdot s}^\top K Y_{\cdot s}}{|Y_{\cdot s}|} \qquad \text{s.t. } Y \in \mathbb{1}^{m\times r}$$

Hess 2019

Kernel k-means has an **INTERPRETATION OF A GRAPH CLUSTERING** method, where K is the weighted adjacency matrix of the graph.

Besides maximizing the similarities within clusters, another objective makes sense:

Minimizing the Cut

MINIMUM CUT OBJECTIVE

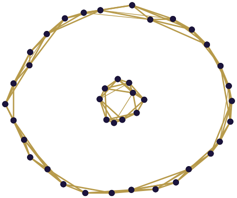
Given a symmetric weighted adjacency matrix $W \in \mathbb{R}_+^{m \times m}$, the **Minimum Cut Objective** is given by

$$\min_{Y \in \mathbb{1}^{m \times k}} \sum_{c=1}^{k} \frac{Y_{\cdot c}^\top W (\mathbf{1} - Y_{\cdot c})}{|Y_{\cdot c}|}.$$

The minimum cut introduces **Graph Laplacians**:

$$\begin{aligned} Y_{\cdot c}^\top W (\mathbf{1} - Y_{\cdot c}) &= Y_{\cdot c}^\top W \mathbf{1} - Y_{\cdot c}^\top W Y_{\cdot c} \\ &= Y_{\cdot c}^\top (I_W - W) Y_{\cdot c} \\ &= Y_{\cdot c}^\top L Y_{\cdot c} \end{aligned}$$

where $I_W = \mathrm{diag}(W\mathbf{1})$.



Hagen and Kahng 1992, Shi and Malik 2000

Nonconvex Clustering Objectives - Kernel k-means vs. Minimum Cut

kernel k-means Objective:

$$\max_{Y} \operatorname{tr}(Z^\top K Z) \qquad \text{s.t. } Z = Y(Y^\top Y)^{-1/2}, Y \in \mathbb{1}^{m \times k}$$

Minimum Cut Objective:

$$\min_{Y} \operatorname{tr}(Z^\top L Z) \qquad \text{s.t. } Z = Y(Y^\top Y)^{-1/2}, Y \in \mathbb{1}^{m \times k}$$

Popular Laplacians	
Difference Laplacian	$L_d = I_W - W$
Symmetric Normalized Laplacian	$L_s = I - I_W^{-1/2} W I_W^{-1/2}$
Random Walk Laplacian	$L_r = I - I_W^{-1} W$

Dhillon et al. 2004

Optimization

The more popular nonconvex clustering method is **SPECTRAL CLUSTERING**.

Spectral clustering is typically presented like this:

The Spectral Relaxation of the Minimum Cut

$$\arg\min_{Y \in \mathbb{1}^{m \times k}} \operatorname{tr}(Z^\top L Z) = \arg\max_{Y \in \mathbb{1}^{m \times k}} \operatorname{tr}(Z^\top (-L) Z) \text{ s.t. } Z = Y(Y^\top Y)^{-1/2}$$

Ky-Fan Theorem ($\lambda_1 \geq \lambda_2 \geq \ldots$ are eigenvalues of $A \in \mathbb{R}^{m \times m}$)

$$\lambda_1 + \ldots + \lambda_k = \max_Z \operatorname{tr}(Z^\top A Z) \quad \text{s.t. } Z^\top Z = I, Z \in \mathbb{R}^{m \times k}$$

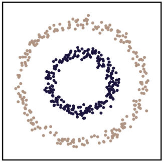
The optimizers Z^* of the relaxed problem are given by the eigenvectors of L, which are **DISCRETIZED** to crisp cluster assignments by k-**MEANS**.

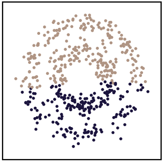
Von Luxburg 2007

Spectral Clustering

Given a dataset $D \in \mathbb{R}^{m \times n}$, number of clusters k

1. Choose a local similarity representation $W \in \mathbb{R}^{m \times m}$
2. Compute the truncated eigendecomposition of a Laplacian $L \approx V_k \Lambda_k {V_k}^\top$
3. Compute a k-means clustering on V_k





Pros:

- Fast

Cons:

- Heuristic
- Sensitive to similarity measure and noise

In this narrative, k-means is a discretization method which happens to work well in practice.

Actually, the application of **k-MEANS** is **WELL JUSTIFIED** by the objective.

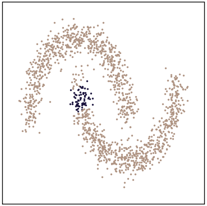
K-MEANS IN SPECTRAL CLUSTERING IS THEORETICALLY FOUNDED

The matrix $\lambda_{max}I - L$ is p.s.d. and has a symmetric decomposition
$\lambda_{max}I - L = V\Lambda V^\top = UU^\top$

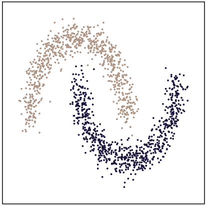
$$\begin{aligned}&\arg\max_{Y\in\mathbb{1}^{m\times k}} \mathrm{tr}(Z^\top(-L)Z)\\ &= \arg\max_{Y\in\mathbb{1}^{m\times k}} \mathrm{tr}(Z^\top(\lambda_{max}I - L)Z)\\ &= \arg\max_{Y\in\mathbb{1}^{m\times k}} \mathrm{tr}(Z^\top UU^\top Z) \text{ s.t. } Z = Y(Y^\top Y)^{-1/2}\end{aligned}$$

The minCut objective is theoretically equivalent to **k-MEANS** on U but that **DOESN'T WORK** in practice, gets stuck in local optima close to global optimum.

$\mathrm{tr}(Z^\top(-L)Z) = 2.039$



$\mathrm{tr}(Z^\top(-L)Z) = 2.093$



Hess et al. 2019

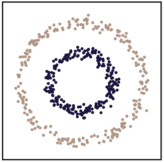
Can we not use the information given by the other eigenvectors than the first k?

YES, WITH A SMALL TWEAK..

SPECTACL

Given a dataset $D \in \mathbb{R}^{m \times n}$, number of clusters k

1. Choose similarity representation $W \in \mathbb{R}^{m \times m}$
2. Compute rank-$d > k$ eigendecomposition $W \approx V_d \Lambda_d V_d{}^\top$
3. Compute a k-means clustering on U, where $U_{je} = |V_{je} \Lambda_{ee}|$

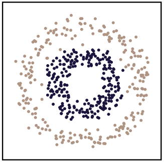


Pros:

- Fast
- More robust to noise
- U has interpretation as fuzzy cluster indicator matrix

Cons:

- Similarity measure must fit



Hess et al. 2019

Conclusions

DISCUSSION

- How to cluster robustly nonconvex shapes is still by and large an open problem
- One interesting research direction is to learn a suitable graph representation automatically

  Bojchevski et al. 2018, Kang et al. 2019
- The other approach is to learn a feature transformation ϕ onto a well-clusterable space ($\rightarrow$ Deep Clustering)

  Bianchi et al. 2020
- Simultaneous optimization of ϕ and the clustering is a young branch of research

  Boubekki et al. 2021

Some References Part IIIb

Von Luxburg 2007	Spectral Clustering Survey
Hess et al. 2019	SpectACl and connection of k-means to spectral clustering