Kernel k-means, Spectral Clustering and Normalized Cuts

- KDD’04

Author: Inderjit S. Dhillon, Yuqiang Guan, Brian Kulis

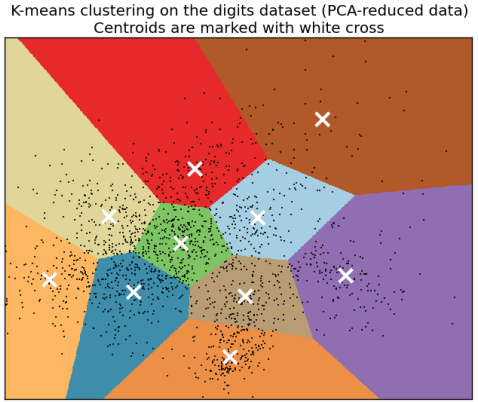
Outline:

1. **Background**
    a. k-means
    b. Weighted Kernel k-means
    c. Spectral Clustering

2. Contribution
    a. Connection
    b. Implication

k-means

1. Centroid-based clustering



2. Objective function

$$\text{minimize} \sum_{k=1}^{K} \sum_{C(i)=k} \left\| x_i - m_k \right\|^2$$

k-means

3. Algorithm
    a. Fix k centroids u_i, minimize objective function w.r.t X.

$$C(i) = \arg\min_{1 \le k \le K} \left\| x_i - m_k \right\|^2, \quad i = 1, \ldots, N$$

    b. Fix X, minimize objective function w.r.t u.

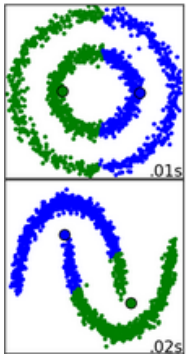
$$m_k = \frac{\sum_{i:C(i)=k} x_i}{N_k}, \; k = 1, \ldots, K.$$

k-means

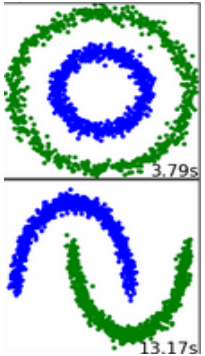
4. Drawback

It can’t separate clusters that are non-linearly separable in input space.

k-means



Ideal clustering



Weighted Kernel k-means

1. Objective function

$$\text{minimize} \quad \sum_{j=1}^{k} \sum_{\mathbf{a} \in \pi_j} w(\mathbf{a}) \|\phi(\mathbf{a}) - \mathbf{m}_j\|^2$$

$$\mathbf{m}_j = \frac{\sum_{\mathbf{b} \in \pi_j} w(\mathbf{b}) \phi(\mathbf{b})}{\sum_{\mathbf{b} \in \pi_j} w(\mathbf{b})}$$

Weighted Kernel k-means

2. Transformation

$$\|\phi(\mathbf{a}) - \mathbf{m}_j\|^2 = \left|\phi(\mathbf{a}) - \frac{\sum_{\mathbf{b}\in\pi_j} w(\mathbf{b})\phi(\mathbf{b})}{\sum_{\mathbf{b}\in\pi_j} w(\mathbf{b})}\right|^2 = \phi(\mathbf{a})\cdot\phi(\mathbf{a}) - \frac{2\sum_{\mathbf{b}\in\pi_j} w(\mathbf{b})\phi(\mathbf{a})\cdot\phi(\mathbf{b})}{\sum_{\mathbf{b}\in\pi_j} w(\mathbf{b})} + \frac{\sum_{\mathbf{b},\mathbf{c}\in\pi_j} w(\mathbf{b})w(\mathbf{c})\phi(\mathbf{b})\cdot\phi(\mathbf{c})}{(\sum_{\mathbf{b}\in\pi_j} w(\mathbf{b}))^2}.$$

K = $\phi(\mathbf{a})\cdot\phi(\mathbf{b})$

Weighted Kernel k-means

3. Algorithm

WEIGHTED_KERNEL_KMEANS(K, k, w, $C_1, ..., C_k$)
Input: K: kernel matrix, k: number of clusters, w: weights for each point
Output: $C_1, ..., C_k$: partitioning of the points

1. Initialize the k clusters: $C_1^{(0)}, ..., C_k^{(0)}$.
2. Set $t = 0$.
3. For each point $\mathbf{a}$, find its new cluster index as

$$j^*(\mathbf{a}) = \text{argmin}_j \|\phi(\mathbf{a}) - \mathbf{m}_j\|^2, \text{ using (2).}$$

4. Compute the updated clusters as

$$C_j^{t+1} = \{\mathbf{a} \ : \ j^*(\mathbf{a}) = j\}.$$

5. If not converged, set $t = t + 1$ and go to Step 3; Otherwise, stop.

Spectral Clustering, Normalized Cut

W: similarity matrix.

W(i,j) = 0 when i, j are not connected.

$$\text{links}(\mathbb{A}, \mathbb{B}) = \sum_{i \in \mathbb{A}, j \in \mathbb{B}} W(i,j) \qquad \text{degree}(\mathbb{A}) = \text{links}(\mathbb{A}, \mathbb{V})$$

$$\text{linkratio}(\mathbb{A}, \mathbb{B}) = \frac{\text{links}(\mathbb{A}, \mathbb{B})}{\text{degree}(\mathbb{A})}$$

Spectral Clustering, Normalized Cut

Objective function:

Minimize $\mathrm{kncuts}(\Gamma_{\mathbb{V}}^{K}) = \frac{1}{K}\sum_{l=1}^{K}\mathrm{linkratio}(\mathbb{V}_l, \mathbb{V}\setminus\mathbb{V}_l)$

Let X be the n*k indicator matrix. $D = \mathrm{Diag}(W1_N)$

Minimize $\frac{1}{K}\sum_{1}^{K}\frac{X_l^T(D-W)X_l}{X_l^T D X_l}$

Spectral Clustering, Normalized Cut

Objective function:

$$\begin{aligned} \text{maximize} \quad & \varepsilon(X) = \frac{1}{K} \sum_{l=1}^{K} \frac{X_l^T W X_l}{X_l^T D X_l} \\ \text{subject to} \quad & X \in \{0,1\}^{N \times K} \\ & X\, 1_K = 1_N. \end{aligned}$$

Spectral Clustering, Normalized Cut

$$Z_l = D^{1/2} X_l^T$$

Objective function:

$$Maximize \frac{1}{K} \sum_{1}^{K} Z_l^T D^{-1/2} W D^{-1/2} Z_l$$

$$Z^T Z = I$$

Outline:

1. Background
   a. k-means
   b. Weighted Kernel k-means
   c. Spectral Clustering

2. **Contribution**
   a. Connection
   b. Implication

Connection

1. By choosing the weights in particular way, the weighted kernel k-means objective function is identical to the spectral clustering normalized cut.

$$\mathbf{m}_j = \frac{\sum_{\mathbf{b}\in\pi_j} w(\mathbf{b})\phi(\mathbf{b})}{\sum_{\mathbf{b}\in\pi_j} w(\mathbf{b})} \longrightarrow \mathbf{m}_j = \Phi_j \frac{W_j \mathbf{e}}{s_j}$$

$$\Phi = [\phi(\mathbf{a}_1, \phi(\mathbf{a}_2), \ldots, \phi(\mathbf{a}_n)] \qquad s_j = \sum_{\mathbf{a}\in\pi_j} w(\mathbf{a})$$

W is diagonal matrix of all weights.

Connection

$$\mathcal{D}(\{\pi_j\}_{j=1}^k) = \sum_{j=1}^k d(\pi_j)$$

$$\begin{aligned} d(\pi_j) &= \sum_{\mathbf{a}\in\pi_j} w(\mathbf{a})\|\phi(\mathbf{a}) - \mathbf{m}_j\|^2 \\ &= \sum_{\mathbf{a}\in\pi_j} w(\mathbf{a})\|\phi(\mathbf{a}) - \Phi_j \frac{W_j\mathbf{e}}{s_j}\|^2 \\ &= \|(\Phi_j - \Phi_j \frac{W_j\mathbf{e}\mathbf{e}^T}{s_j})W_j^{1/2}\|_F^2 \\ &= \|(\Phi_j W_j^{1/2}(I - \frac{W_j^{1/2}\mathbf{e}\mathbf{e}^T W_j^{1/2}}{s_j})\|_F^2 \end{aligned}$$

Connection

1. $\text{trace}(AA^T) = \text{trace}(A^TA) = \|A\|_F^2$
2. $I - \frac{W_j^{1/2}\mathbf{e}\mathbf{e}^TW_j^{1/2}}{s_j} = P$ since $s_j = \mathbf{e}^TW_j\mathbf{e}$ we know $P^2 = P$

$$
\begin{aligned}
d(\pi_j) &= \text{trace}\ \Phi_j W_j^{1/2} \left| I - \frac{W_j^{1/2}\mathbf{e}\mathbf{e}^TW_j^{1/2}}{s_j} \right|^2 W_j^{1/2}\Phi_j^T \\
&= \text{trace}\ \Phi_j W_j^{1/2} \ I - \frac{W_j^{1/2}\mathbf{e}\mathbf{e}^TW_j^{1/2}}{s_j} \ W_j^{1/2}\Phi_j^T \\
&= \text{trace}(W_j^{1/2}\Phi_j^T\Phi_j W_j^{1/2}) - \frac{\mathbf{e}^TW_j}{\sqrt{s_j}}\Phi_j^T\Phi_j\frac{W_j\mathbf{e}}{\sqrt{s_j}}.
\end{aligned}
$$

Connection

$$\mathcal{D}(\{\pi_j\}_{j=1}^k) = \mathrm{trace}(W^{1/2}\Phi^T\Phi W^{1/2}) - \mathrm{trace}(Y^T W^{1/2}\Phi^T\Phi W^{1/2} Y)$$

$$Y = \begin{bmatrix} \frac{W_1^{1/2}\mathbf{e}}{\sqrt{s_1}} & & & \\ & \frac{W_2^{1/2}\mathbf{e}}{\sqrt{s_2}} & & \\ & & \cdots & \\ & & & \frac{W_k^{1/2}\mathbf{e}}{\sqrt{s_k}} \end{bmatrix}.$$

First term is constant, K = $\Phi^T\Phi$

Connection

So, the problem:

Minimize $\mathcal{D}(\{\pi_j\}_{j=1}^k)$

becomes:

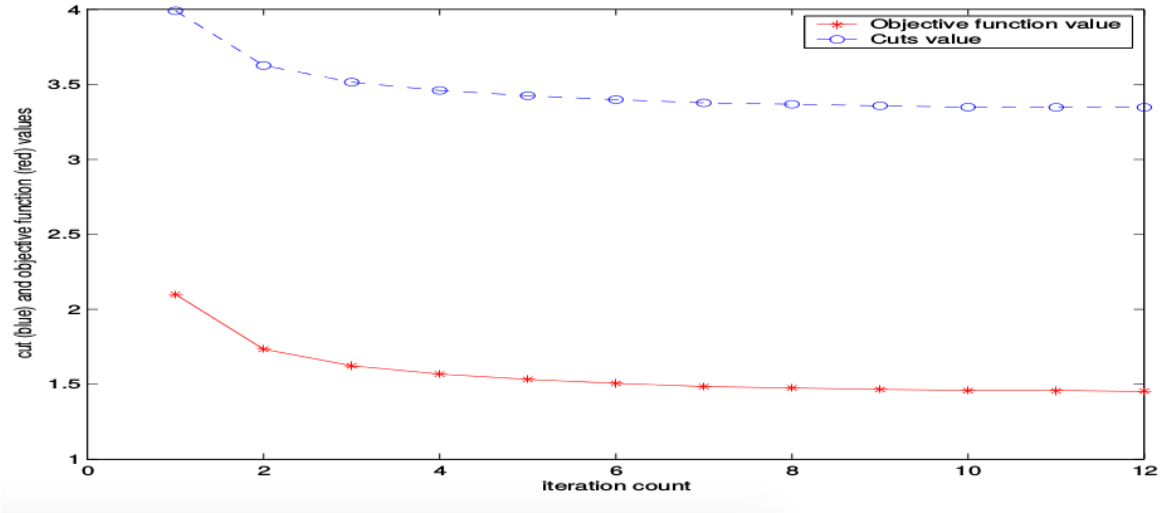
Maximize $\mathrm{trace}(Y^T W^{1/2} K W^{1/2} Y)$

s.t. $Y^T Y = I$

Y is an n*k orthonormal matrix, so the optimal Y is the top k eigenvectors of $W^{1/2} K W^{1/2}$

Implication

1. Compute normalized cuts using weighted kernel k-means.



Implication

2. Techniques in kernel k-means could be used to compute normalized cuts.

   a. Local search.
   b. Pruning procedure.

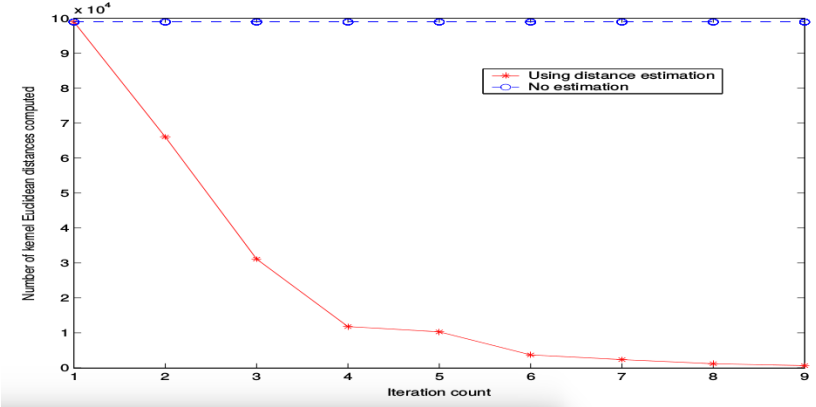
$$\|x - m_j^n\| \geq \|x - m_j^0\| - \|m_j^n - m_j^0\|$$

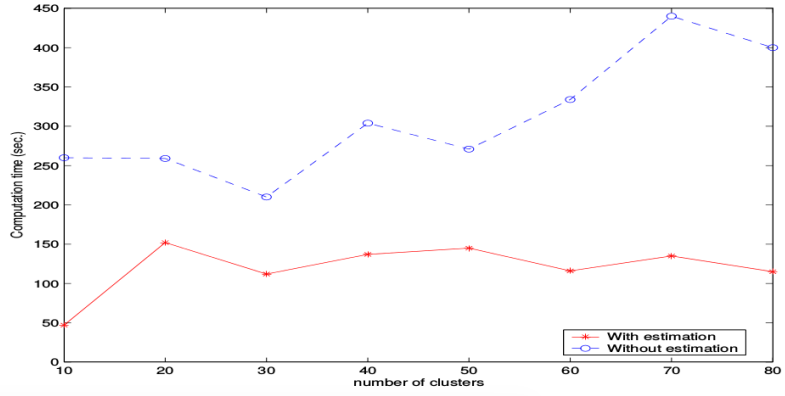
Use right side to compute the lower bound as estimation.

If estimation is smaller than the distance from x to its cluster center, compute the distance from x to m_j.

Implication

Pruning procedure.





Implication

3. Compute kernel k-means using eigenvectors.
   a. run spectral clustering to get an initial partition.
   b. run kernel k-means on this partition.

	initial	final	NMI
random	.0213	.0062	.666
spectral	.0081	.0059	.698

Summary

1. Connection between weighted kernel k-means and spectral clustering.

2. Implications of this connection.

Thanks