Machine Learning for Data Science (CS4786)
Lecture 12

Clustering

Course Webpage :
http://www.cs.cornell.edu/Courses/cs4786/2016sp/

Clustering

1. K-means clustering: Find cluster assignments and centroids to minimize,

$$\sum_{k=1}^{K} \sum_{x_t \in C_k} \|x_t - r_k\|_2^2$$

k depends on Data
但实际上k = n时值最小

2. Single link clustering: Find cluster assignments to maximize,

$$\min_{x_t, x_s : c(x_t) \neq c(x_s)} d(x_t, x_s)$$

k = 1 值最大

3. Spectral clustering: Find cluster assignments to minimize,

$$\text{NCut} = \sum_{j=1}^{K} \frac{\text{Cut}(C_j)}{\#\text{Edges}(C_j)}$$

k = 1 时值最小

(Can convert distances between points into weighted graph).

How do we find K?

- Find K that optimizes objective?
  - What is optimal K for K-means, single link, for Spectral clustering (normalized or unnormalized)?
- Is there a good strategy to pick K?
  - Single link, stop merging when
    - Between cluster distance becomes smaller than threshold θ
    - Between cluster distance becomes smaller than $\alpha \cdot \max_{\mathbf{x}_t, \mathbf{x}_2} d(\mathbf{x}_s, \mathbf{x}_t)$
  - Spectral clustering, pick K by looking for jump (gap) in eigen value
  - K-means add penalty for larger K (Eg. penalize by number of clusters)
- Is there a universal way to select K?

Properties of clustering?

- Given n points $\mathcal{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ and distance function $d : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}^+$, produce clustering $C(d)$ (or c_d).

C: clustering assignment
C(d): given distance function d, the clustering assignment under d

- Desirable properties:

1. Scale invariance: for any d and any $\alpha > 0$, let $d_\alpha(\mathbf{x}_s, \mathbf{x}_t) = \alpha \cdot d(\mathbf{x}_t, \mathbf{x}_s)$ for all pairs of points in $\mathcal{X}$. We would like,

$$C(d) = C(d_\alpha)$$

2. Consistency: Let d be any distance function. Let d' be such that:

same clustering, distance not goes larger $\forall \mathbf{x}_t, \mathbf{x}_s \text{ s.t. } c_d(\mathbf{x}_t) = c_d(\mathbf{x}_s), \quad d'(\mathbf{x}_t, \mathbf{x}_s) \le d(\mathbf{x}_t, \mathbf{x}_s)$

different clustering, distance not goes smaller $\forall \mathbf{x}_t, \mathbf{x}_s \text{ s.t. } c_d(\mathbf{x}_t) \neq c_d(\mathbf{x}_s), \quad d'(\mathbf{x}_t, \mathbf{x}_s) \ge d(\mathbf{x}_t, \mathbf{x}_s)$

==> C(d') = C(d)

only 2 of these 3 will hold for any algorithm

3. Richness/Universality:

{1} {2} {3} {4}
and
{1}, {2}, {3,4}
should all be possible

$$\text{Range}(C) = \text{Set of all partitions}$$

Properties of clustering?

Fixed K:

	K-means,	single link ,	spectrum
1	yes	yes	depend on normalized or not
2	yes	yes	depends
3	no	no	depends

Stop when dist > theta

	single link
1	no
2	yes
3	yes

- Which properties do the various algorithms satisfy?
  - Fixed K: Kmeans, Single link
  - Thresholded single-link

Can any clustering algorithm have all 3 desired properties?

Impossibility

Theorem

Any clustering algorithm that has scale invariance and consistency does not have richness.

IMPOSSIBILITY

- If algorithm has scale invariance and consistency, then algorithm cannot produce refinements of clusters.

  Eg. $\{1,2\},\{3\},\{4,5,6\}$ is a refinement of partition $\{1,2,3\},\{4,5,6\}$

- Why is refinement a problem for "scale invariance + consistency"?