Clustering

Peng Lingwei

August 20, 2019

Contents

22 Clustering

The basic problem of clustering is without rigorous definition, because:

1. Similarity and dissimilar are not transitive relations: If a is similar to b, and b is similar to c, we can't get a is similar to c.
2. Unsupervised learning problem has no clear success evaluation procedure for clustering, even on the basis of full knowledge of the underlying data distribution.

Hence, different clustering algorithms will output very different clusterings.

Definition 1. ***(A clustering model).***

1. ***Input****:*
   - *a set of elements* $S \in \mathcal{X}^m$*;*
   - *a distance function* $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ *(satisfies* $d(x_1, x_2) = d(x_2, x_1), d(x, x) = 0, d(x_1, x_2) \leq d(x_1, x_3) + d(x_2, x_3)$*);*
   - *clusters number* k*.*
2. ***Output****:*
   - *Hard cluster:* $C = (C_1, \ldots, C_k)$*, where* $\cup_{i=1}^{k} C_i = \mathcal{X}$*, and* $i \neq j \Rightarrow C_i \cap C_j = \emptyset$*;*
   - *Soft cluster:* $\forall x \in \mathcal{X}, P(x) = (p_1(x), \ldots, p_k(x))$*, where* $p_i(x) = \mathbb{P}[x \in C_i]$*.*

22.1 LINKAGE-BASED CLUSTERING ALGORITHMS

These kind of algorithms start from the trivial clustering that has each data point as a single-point cluster. Then, repeatedly merge the "closest" clusters of the previous clustering.

The distance between domain subsets:

1. Single Linkage clustering: $D(A, B) = \min \{d(x, y) : x \in A, y \in B\}$;
2. Average Linkage clustering: $D(A, B) = \frac{1}{|A||B|} \sum_{x \in A, y \in B} d(x, y)$;
3. Max Linkage clustering: $D(A, B) = \max \{d(x, y) : x \in A, y \in B\}$.

Stopping criteria:

1. Fixed number of cluster;
2. Distance upper bound, fix some $r \in \mathbb{R}_+$.

22.2 k-MEANS AND OTHER COST MINIMIZATION CLUSTERINGS

1. The centroid of C_i is defined to be: $\mu_i(C_i) = \arg\min_{\mu \in \mathcal{X}} \sum_{s \in C_i} d(s, \mu)^2$;
2. $G_{k-means}((S,d),(C_1,\ldots,C_k)) = \min_{\mu_1,\ldots,\mu_k \in \mathcal{X}} \sum_{i=1}^{k} \sum_{x \in C_i} d(x, \mu_i)^2$;
3. $G_{k-medoid}((S,d),(C_1,\ldots,C_k)) = \min_{\mu_1,\ldots,\mu_k \in S} \sum_{i=1}^{k} \sum_{x \in C_i} d(x, \mu_i)^2$;
4. $G_{k-median}((S,d),(C_1,\ldots,C_k)) = \min_{\mu_1,\ldots,\mu_k \in S} \sum_{i=1}^{k} \sum_{x \in C_i} d(x, \mu_i)$;
5. The sum of in-cluster distances (not center based): $G_{k-SOD}((S,d),(C_1,\ldots,C_k)) = \sum_{i=1}^{k} \sum_{x,y \in C_i} d(x,y)$;

Algorithm 1 k-Means

Require: $S \in \mathcal{X}^m$; number of cluster k.
Ensure: Randomly choose initial centroids $\mu_1, \ldots, \mu_k$
  while not convergence **do**
    $\forall i \in [k], C_i = \{x \in \mathcal{S} : i = \arg\min_j \|x - \mu_j\|\}$
    $\forall i \in [k]$, update $\mu_i = \frac{1}{|C_i|} \sum_{x \in C_i} x$
  end while.

Lemma 1. *K-means algorithm does not increase the k-means objective function.*

Proof.

$$\arg\min_{\mu \in \mathbb{R}^n} \sum_{x \in C_i} \|x - \mu\|^2 = \frac{1}{|C_i|} \sum_{x \in C_i} x$$

$$G_{k-means}(C_1^{(t-1)}, \ldots, C_k^{(t-1)}) = \sum_{i=1}^{k} \sum_{x \in C_i^{(t-1)}} \|x - \mu_i^{(t-1)}\|^2$$

$$G_{k-means}(C_1^{(t)}, \ldots, C_k^{(t)}) \leq \sum_{i=1}^{k} \sum_{x \in C_i^{(t)}} \|x - \mu_i^{(t-1)}\|^2 \leq \sum_{i=1}^{k} \sum_{x \in C_i^{(t-1)}} \|x - \mu_i^{(t-1)}\|^2$$

□

The k-means might converge to a point which is not even a local minimum.

22.3 SPECTRAL CLUSTERING

Similarity graph: every two vertices are connected by an edge whose weight is their similarity $W_{i,j} = s(x_i, x_j)$. Such as: $W_{i,j} = \exp(-d(x_i, x_j)^2/\sigma^2)$.

The question becomes cutting the minimum weights edges.

22.3.1 Graph Cut

Find the mincut:

1. $Cut(C_1, \ldots, C_k) = \sum_{i=1}^{k} \sum_{r \in C_i, s \notin C_i} W_{r,s}$;
2. $RatioCut(C_1, \ldots, C_k) = \sum_{i=1}^{k} \frac{1}{|C_i|} \sum_{r \in C_i, s \notin C_i} W_{r,s}$. (Balance)

Definition 2. ***(Unnormalized Graph Laplacian).***

$$L = D - W, \quad D_{ij} = \sum_{j=1}^{m} W_{i,j}$$

Diagonal matrix D is called the degree matrix.

Lemma 2. *Let* $H_{i,j} = \frac{1}{\sqrt{|C_j|}} 1_{[i \in C_j]}$*,where* $H \in \mathbb{R}^{m,k}$*, then*

$$RatioCut(C_1, \ldots, C_k) = trace(H^T L H).$$

Proof.

$$\vec{v}^T L \vec{v} = \sum_{i=1}^{m} \sum_{j=1}^{m} L_{i,j} v_i v_j = \frac{1}{2} \left(\sum_{r} D_{r,r} v_r^2 - 2 \sum_{r,s} v_r v_s W_{r,s} + \sum_{s} D_{s,s} v_s^2 \right) = \frac{1}{2} \sum_{r,s} W_{r,s} (v_r - v_s)^2$$

$$trace(H^T L H) = \sum_{i=1}^{k} h_i^T L h_i = \sum_{i=1}^{k} \frac{1}{2} \sum_{r,s} W_{r,s} (h_r - h_s)^2 = \sum_{i=1}^{k} \frac{1}{|C_i|} \sum_{r \in C_i, s \notin C_i} W_{r,s}$$

□

Algorithm 2 Unnormalized Spectral Clustering

Require: $W \in \mathbb{R}^{m,m}$; Number of clusters k.
Compute Laplacian L.
Let $U \in \mathbb{R}^{m,k}$ be the matrix whose columns are the eigenvectors of L corresponding to the k smallest eigenvalues.
Let $v_1, \ldots, v_m$ be the rows of U.
Cluster the points $v_1, \ldots, v_m$ using k-means.
return Clusters $C_1, \ldots, C_K$ of k-means algorithm.

22.4 INFORMATION BOTTLENECK

22.5 A HIGH LEVEL VIEW OF CLUSTERING

Kleinberg tried to solve the question what is clustering. He think a clustering function F should have three properties:

1. Scale Invariance (SI): $F(\mathcal{X}, d) = F(\mathcal{X}, \alpha d)$;
2. Richness (Ri): $\forall S \in \mathcal{X}^m, C = (C_1, \ldots, C_k), \exists d, F(S, d) = C$;
3. Consistency (Co): $\forall d, d'$, if x, y belong to the same cluster in $F(S, d) \Rightarrow d'(x, y) \leq d(x, y)$, and if x, y belong to different clusters in $F(S, d) \Rightarrow d'(x, y) \geq d(x, y)$, then $F(S, d) = F(S, d')$

Theorem 1. *There exists no function, F, that satisfies all the three properties: Scale Invvariance, Richness, and Consistency.*

Proof. Assume that F does satisfy all three properties.

We choost S with at least 3 points. By Richness, d_1, d_2 satisfy $F(S, d_1) \neq F(S, d_2)$ and $F(S, d_1) = \{\{x\} : x \in S\}$.

Then, $\exists \alpha > 0$, $\forall x, y \in S, \alpha d_2(x, y) \geq d_1(x, y)$. With consistency property and the special structure of $F(S, d_1)$, we have $F(S, d_2) = F(S, \alpha d_2) = F(S, d_1)$. □

1. Center-based fails the consistency property.
2. If we fix k, and there exists F satisfying $k - Richness$, Scale Invariance and Consistency.
3. One can come up with many other different properties of clustering functions by prior knowledge.
4. There is no "ideal" clustering function, as the No-Free-Lunch theorem in classification problems.