COL870: Clustering Algorithms

Ragesh Jaiswal, CSE, IIT Delhi

An Axiomatic Framework for Clustering

Axiomatic Framework for Clustering

- Given a dataset, how do we define natural clusters in the dataset?
- One way to resolve the above question is to define a clustering function that takes a dataset as input and output a partition/clustering of the dataset and then trying to figure out the specific properties that this function should satisfy.
- In other words, in what ways does a clustering function differ from an arbitrary function that takes the dataset as input and outputs a partitioning of the dataset?
- So, we would like to define some basic axioms/properties that clustering functions should satisfy.
- We will study a specific axiomatic framework, known as *Klienberg's framework* and see how no function could simultaneously satisfy a set of axioms (*an impossibility result*) that appear to be natural properties for clustering.

- We will use S to denote the dataset.
- Definition: Any function $D : S \times S \rightarrow \mathbb{R}$ is a distance function if it satisfies the following two properties:
  - $\forall x, y \in S, D(x, y) \geq 0$ with equality iff $x = y$.
  - $\forall x, y \in S, D(x, y) = D(y, x)$.
- We will consider clustering functions that take only S and D as input. So, we consider functions f takes an $|S| \times |S|$ matrix denoting distances and outputs a partitioning Γ of the dataset. For example:

$$f\begin{pmatrix} 0 & 10 & 10 \\ 10 & 0 & 1 \\ 10 & 1 & 0 \end{pmatrix} = (\{1\}, \{2, 3\})$$

- We will use S to denote the dataset.
- Definition: Any function $D : S \times S \rightarrow \mathbb{R}$ is a distance function if it satisfies the following two properties:
    - $\forall x, y \in S, D(x, y) \geq 0$ with equality iff $x = y$.
    - $\forall x, y \in S, D(x, y) = D(y, x)$.
- We will consider clustering functions that take only S and D as input. So, we consider functions f takes an $|S| \times |S|$ matrix denoting distances and outputs a partitioning Γ of the dataset.
- Following are three reasonable axioms that any reasonable clustering function should satisfy:
    1. Scale Invariance: For any $\alpha > 0$ and any D, $f(\alpha \cdot D) = f(D)$.
    2. Richness: $Range(f) = \{\text{all possible partitions of } [n]\}$.
    3. Consistency: If $f(D) = \Gamma$ and D' is a "Γ-enhancing" transformation of D, then $f(D') = \Gamma$.
        - D' is a Γ-enhancing transformation of D if

$$\begin{aligned} D'(i,j) &\leq D(i,j) && \text{for } i,j \text{ in the same cluster of } \Gamma \\ D'(i,j) &\geq D(i,j) && \text{for } i,j \text{ in different cluster of } \Gamma \end{aligned}$$

- We will use S to denote the dataset.
- Definition: Any function $D : S \times S \rightarrow \mathbb{R}$ is a distance function if it satisfies the following two properties:
  - $\forall x, y \in S, D(x, y) \geq 0$ with equality iff $x = y$.
  - $\forall x, y \in S, D(x, y) = D(y, x)$.
- We will consider clustering functions that take only S and D as input. So, we consider functions f takes an $|S| \times |S|$ matrix denoting distances and outputs a partitioning Γ of the dataset.
- Following are three reasonable axioms that any reasonable clustering function should satisfy:
  1. Scale Invariance: For any $\alpha > 0$ and any D, $f(\alpha \cdot D) = f(D)$.
  2. Richness: $Range(f) = \{\text{all possible partitions of } [n]\}$.
  3. Consistency: If $f(D) = \Gamma$ and D' is a "Γ-enhancing" transformation of D, then $f(D') = \Gamma$.
     - D' is a Γ-enhancing transformation of D if

$$\begin{aligned} D'(i,j) &\leq D(i,j) \quad \text{for } i,j \text{ in the same cluster of } \Gamma \\ D'(i,j) &\geq D(i,j) \quad \text{for } i,j \text{ in different cluster of } \Gamma \end{aligned}$$

Theorem

There is no clustering function that satisfies all three axioms.

End