Theory of Clustering

Issues with Clustering

There are several different methods for clustering

- Centroid based (k-means, k-medians, k-centers)
- Density based (DBSCAN, watershed, clustertrees)
- Hierarchical methods (linkage-trees)
- Similarity based (ncuts, spectral clustering)
- Bayesian/probabilistic methods (GMM, DPMM)

Despite having an abundance of methods, somehow it is still unsatisfactory…

For a new application we encounter, somehow none of these methods give what we want, and the practitioner is left with designing their own new clustering method

A Wholistic View of Clustering

Rather than designing yet another clustering algorithm (YACA^{TM}), can one list a set of conditions/principles which any reasonable clustering algorithm should satisfy?

- doing so provides a gold standard, and would help design a high-quality clustering algorithm.
- Since these conditions must apply to every clustering task, these need to be simple, intuitive and fundamental.

What would these fundamental principles/conditions be?

An Axiomatic View of Clustering

Given a set of points X and a notion of comparison/distance d, one can view clustering as a function f: $(X,d) \mapsto$ some partition of X

For f to be a reasonable clustering algorithm, it should satisfy the following very natural conditions…

- **Scale-Invariance.** $f(X,d) = f(X,\alpha d)$, for any $\alpha>0$

  changing the units doesn't change the clustering

- **Richness.** Different d's can yield different partitions. In fact, for all partitions P of X, there is a distance d, which can produce the partition. $\forall\, P\ \exists\, d,\ f(X,d) = P$

  The function f is flexible, and takes d into account… doesn't simply produce trivial partitions

- **Consistency.** If d produces a partition P, then any d' that *enhances* the partition, ie $d' \leq d$ for intracluster distances, and $d' \geq d$ for intercluster distances, then $f(X,d) = f(X,d')$

  Enhancing a clustering, should still yield that clustering

The Impossibility Result

Theorem. The three axioms (Scale-Invariance, Richness, and Consistency) are inconsistent!   That is, there is no function *f* that can simultaneously satisfy all three axioms.

This provides some indication on why practitioners are usually dissatisfied with a clustering algorithm…

The result is due to Kleinberg '15

The Proof

Theorem. The three axioms (Scale-Invariance, Richness, and Consistency) are inconsistent! That is, there is no function *f* that can simultaneously satisfy all three axioms. [Kleinberg '15]

Proof

Let *f* be a function that satisfies all three conditions and consider just three points $X = \{x_1, x_2, x_3\}$

By Richness, there exists *d* and *d'* such that

$$f(X,d) = \{ \{x_1\}, \{x_2\}, \{x_3\} \}, \qquad f(X,d') = \{ \{x_1, x_2\}, \{x_3\} \}$$

f(X,d) ≠ f(X,d')

Pick any $\alpha > 0$ sufficiently large such that $\alpha d' > d$.

Define $d'' := \alpha d'$, then

$$\underbrace{f(X,d) = f(X,d'')}_{\text{consistency}} = f(X,d') \quad \text{with } \overbrace{f(X,d'') = f(X,d')}^{\text{scale-invariance}}$$

∎