Machine learning theory

Theory of clustering

Hamid Beigy

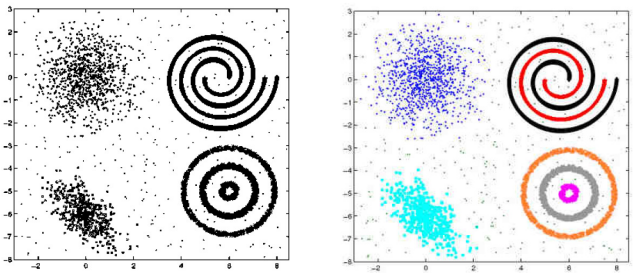
Sharif university of technology

June 6, 2022

Table of contents

Introduction

1. Clustering is the process of grouping a set of data objects into multiple groups or clusters so that objects within a cluster have high similarity, but are very dissimilar to objects in other clusters.
2. Dissimilarities and similarities are assessed based on the feature values describing the objects and often involve distance measures.
3. Clustering is usually an unsupervised learning problem.
4. Consider a dataset $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$, $x_i \in \mathbb{R}^n$.
5. Assume there are K clusters $C_1, \ldots, C_K$.
6. The goal is to group the examples into K homogeneous partitions.



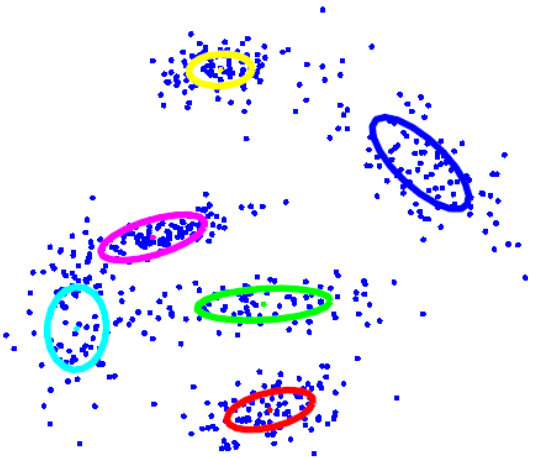
(a) Input data (b) Desired clustering

Picture courtesy: "Data Clustering: 50 Years Beyond K-Means", A.K. Jain (2008)

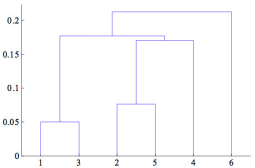
1. A good clustering is one that achieves:
   - High within-cluster similarity
   - Low inter-cluster similarity
2. Applications of clustering
   - Document/Image/Webpage Clustering
   - Image Segmentation
   - Clustering web-search results
   - Clustering (people) nodes in (social) networks/graphs
   - Pre-processing phase

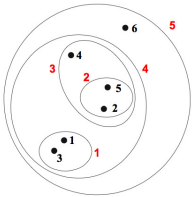
1. The clustering methods can be compared using the following aspects:
   - The partitioning criteria : In some methods, all the objects are partitioned so that no hierarchy exists among the clusters.
   - Separation of clusters : In some methods, data partitioned into mutually exclusive clusters while in some other methods, the clusters may not be exclusive, that is, a data object may belong to more than one cluster.
   - Similarity measure : Some methods determine the similarity between two objects by the distance between them; while in other methods, the similarity may be defined by connectivity based on density or contiguity.
   - Clustering space : Many clustering methods search for clusters within the entire data space. These methods are useful for low-dimensionality data sets. With high-dimensional data, however, there can be many irrelevant attributes, which can make similarity measurements unreliable. Consequently, clusters found in the full space are often meaningless. It's often better to instead search for clusters within different subspaces of the same data set.

Flat or Partitional clustering (Partitions are independent of each other)



Hierarchical clustering (Partitions can be visualized using a tree structure (a dendrogram))

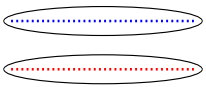




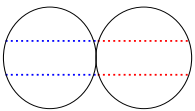
Possible to view partitions at different levels of granularities (i.e., can refine/coarsen clusters) using different K.

1. Why is it hard to define what is clustering?

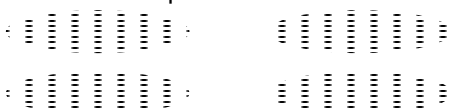
Similar objects in same group



Dissimilar objects are separated



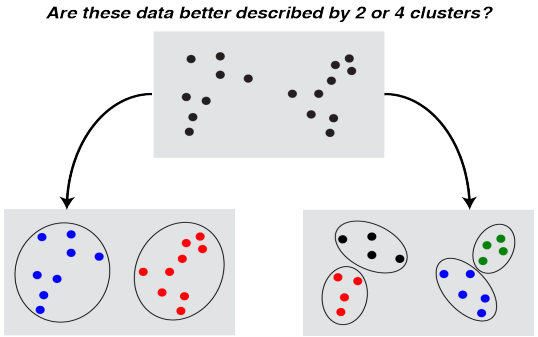
2. Lack of ground truth: Cluster these points into two clusters.



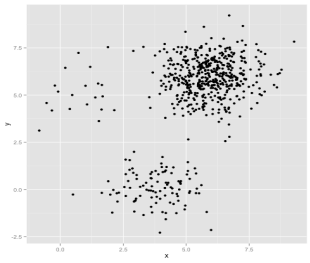
3. We have two well justifiable solutions.

Why is it hard to define what is clustering?

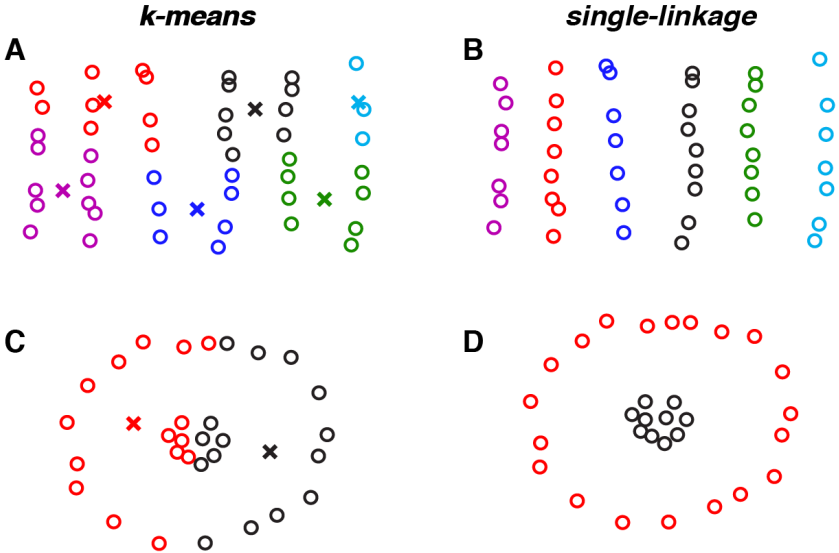
1. It is difficult to determine the number of clusters in a dataset (Williams 2015).



2. It is difficult to cluster outliers.

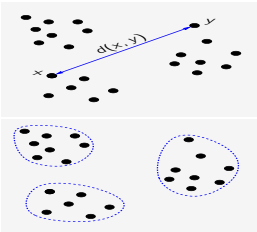


It is difficult to cluster non-spherical, overlapping data (Williams 2015).



Distance based clustering

1. Let $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m\}$ be the dataset.
2. Let $d : \mathbf{X} \times \mathbf{X} \mapsto \mathbb{R}$ be the distance function.
   - $d(\mathbf{x}, \mathbf{y}) \geq 0$ for all $\mathbf{x}, \mathbf{y} \in \mathbf{X}$.
   - $d(\mathbf{x}, \mathbf{y}) = 0$ if and only if $\mathbf{x} = \mathbf{y}$.
   - $d(\mathbf{x}, \mathbf{y}) = d(\mathbf{y}, \mathbf{x})$.
3. A clustering $\mathcal{C}$ is a partition of $\mathbf{X}$.



Definition (Clustering function)

A clustering function is a function f which given a data set $\mathbf{X}$ and a distance function d on $\mathbf{X}$ it returns a partition $\mathcal{C}$ of $\mathbf{X}$.

$$f : (\mathbf{X}, d) \mapsto \mathcal{C}$$

Definition (Clustering quality function)

A clustering quality function is any function Q which given a data set $\mathbf{X}$, a partitioning $\mathcal{C}$ of $\mathbf{X}$ and a distance function d it returns a real number.

$$Q : (\mathbf{X}, d, \mathcal{C}) \mapsto \mathbb{R}$$

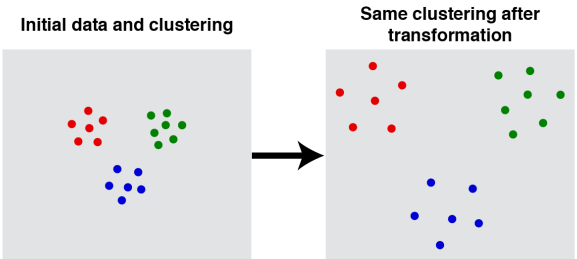
Why axioms? (Kleinberg 2002)

1. There is no unique definition of clustering.
2. Can we formalize our intuition of good objective functions?
3. Are existing objective functions good?
4. Can we design better objective functions?
5. Instead of designing clustering algorithm, we 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.

1. If d is a distance function, we write $\alpha \times d$ to denote the distance function in which the distance between i and j is $\alpha \times d(i,j)$.

Definition (Scale invariance)

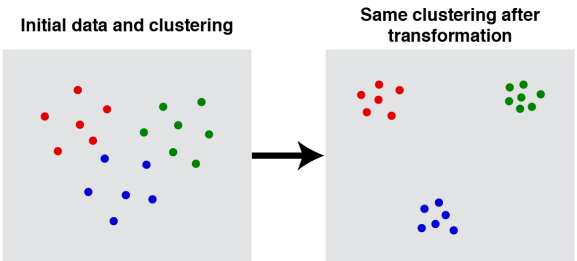
For any distance function d and any $\alpha > 0$, we have $f(d) = f(\alpha \times d)$.



This means that an ideal clustering function does not change its result when the data are scaled equally in all directions.

Definition (Consistency)

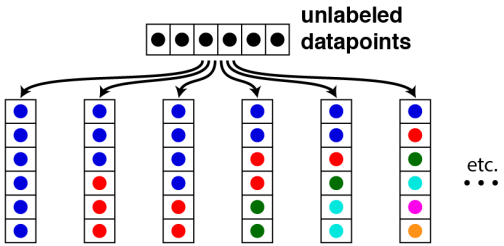
Let d and d' be two distance functions. The clustering function produces a partition of points for the first distance function, d. If, for every pair (i,j) belonging to the same cluster, $d(i,j) \geq d'(i,j)$, and for every pair belonging to different clusters, $d(i,j) \leq d'(i,j)$, then the clustering result shouldn't change: $f(d) = f(d')$.



This means that if we stretch the data so that the distances between clusters increases and/or the distances within clusters decreases, then the clustering shouldn't change.

Definition (Richness)

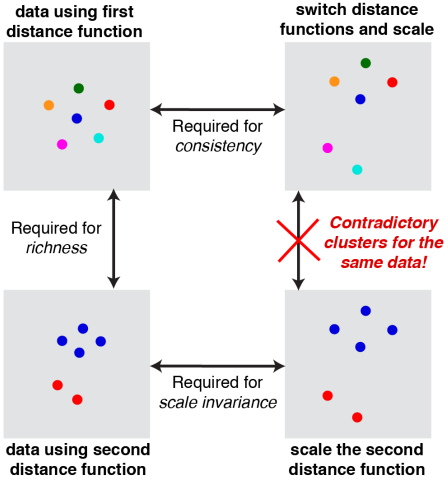
Let the size of the dataset be m and Range(f) is equal to the set of all partitions of $\mathbf{X}$. For a clustering function, f, richness implies that Range(f) is equal to all possible partitions of a set of length m.



This means that an ideal clustering function would be flexible enough to produce all possible partition/clusterings of this set. This means that it automatically determines both the number and clusters in the dataset.

Theorem (Kleinberg's impossibility theorem)

For each $m \geq 2$, there is no clustering function f that satisfies Scale-Invariance, Richness, and Consistency.



1. Kleinberg's results was focusing on clustering functions.
2. Ackerman and Ben-David studied that the clustering quality measures as the object to be axiomatized (Ackerman and Ben-David 2008).

Definition (Quality functions)

The clustering quality function $Q : \mathbf{X} \times \mathcal{C} \mapsto \mathbb{R}_+$ maps a distance function and a clustering into a non-negative real number, $(d, c) \mapsto r$.

Definition (Scale invariance)

Q is scale invariant if for every clustering $\mathcal{C}$ of $(\mathbf{X}, d)$ and every $\alpha > 0$, $Q(\mathbf{X}, d, \mathcal{C}) = Q(\mathbf{X}, \alpha d, \mathcal{C})$.

Definition (Richness)

Q is rich if for any $\mathcal{C}^*$ of $\mathbf{X}$, there exists some d over $\mathbf{X}$ such that $\mathcal{C}^* = \underset{\mathcal{C}}{\operatorname{argmax}}\ Q(\mathbf{X}, d, \mathcal{C})$.

Definition (Consistency)

Q is consistent if for any $\mathcal{C}$ of $\mathbf{X}$, if $d_{\mathcal{C}}$ corresponds to d where intra (extra) cluster distances are decreased (increased) then $Q(\mathbf{X}, d, \mathcal{C}) = Q(\mathbf{X}, d_{\mathcal{C}}, \mathcal{C})$.

Theorem (Consistency of new axioms)

Consistency, scale invariance, and richness for clustering quality measures form a consistent set of requirements.

Summary

Summary

1. Kleinberg's work on axioms for clustering function is framed in terms of distance functions.
2. Kleinberg's impossibility result is for clustering functions.
3. Quality functions are more flexible and allow for axiomatization of data clustering (Ackerman 2012; Ackerman and Ben-David 2008, 2016; Ackerman, Ben-David, and Loker 2010a,b).
4. Graphs are flexible for clustering and needs to be axiomatized (Laarhoven and Marchiori 2014).

Readings

1. Chapter 22 of Shai Shalev-Shwartz and Shai Ben-David (2014). *Understanding machine learning: From theory to algorithms*. Cambridge University Press.

- Ackerman, Margareta (2012). "Towards Theoretical Foundations of Clustering". PhD thesis. University of Waterloo, Ontario, Canada.
- Ackerman, Margareta and Shai Ben-David (2008). "Measures of Clustering Quality: A Working Set of Axioms for Clustering". In: *Proceedings of the 24th Annual Conference on Neural Information Processing Systems*.
- – (2016). "A Characterization of Linkage-Based Hierarchical Clustering". In: *Journal of Machine Learning Research* 17.231, pp. 1–17.
- Ackerman, Margareta, Shai Ben-David, and David Loker (2010a). "Characterization of Linkage-based Clustering". In: *Proceedings of the 23rd Conference on Learning Theory*, pp. 270–281.
- – (2010b). "Towards Property-Based Classification of Clustering Paradigms". In: *Proceedings of the 24th Annual Conference on Neural Information Processing Systems*, pp. 10–18.
- Kleinberg, Jon (2002). "An Impossibility Theorem for Clustering". In: *Proceedings of Conference on Neural Information Processing Systems*, pp. 446–453.
- Laarhoven, Twan van and Elena Marchiori (2014). "Axioms for Graph Clustering Quality Functions". In: *Journal of Machine Learning Research* 15.6, pp. 193–215.


Shalev-Shwartz, Shai and Shai Ben-David (2014). *Understanding machine learning: From theory to algorithms*. Cambridge University Press.


Williams, Alex (2015). *What is clustering and why is it hard?* URL: http://alexhwilliams.info/itsneuronalblog/2015/09/11/clustering1/.

Questions?