Introduction to Machine Learning

Lecture 4: Clustering

Alexis Zubiolo
alexis.zubiolo@gmail.com

Data Science Team Lead @ Adcash

November 17, 2016

What is clustering?

Clustering algorithms aim at **grouping unlabeled objects**.

What is clustering?

Clustering algorithms aim at **grouping unlabeled objects**.

Goal: Find clusters such that objects in the same cluster are more similar to each other than to objects in other clusters.

What is clustering?

Clustering algorithms aim at **grouping unlabeled objects**.

Goal: Find clusters such that objects in the same cluster are more similar to each other than to objects in other clusters.

Main challenges:

- What does *similar* mean?
- Given a similarity definition, how do we define clusters?
- How many clusters?

k-means

k-means

Split the set of points into k classes.

k-means

We look for a partition $S = \{S_1, S_2, \ldots, S_k\}$ minimizing the within-cluster sum of squares.

$$\arg\min_S \sum_{i=1}^{k} \sum_{x \in S_i} \|x - \mu_i\|_2^2$$

where

$$\mu_i = \frac{1}{|S_i|} \sum_{x \in S_i} x$$

is the mean of points in S_i.

k-means

Remark: The k-means solution depends on the initial position of the μ_is centroids.
(see animation by Andrey Shabalin)

k-means

Remark: The k-means solution depends on the initial position of the μ_is centroids.
(see animation by Andrey Shabalin)

2 related questions:

1. How to choose the initial μ_is?
2. How to have more stable results?

k-means

Remark: The k-means solution depends on the initial position of the μ_is centroids.
(see animation by Andrey Shabalin)

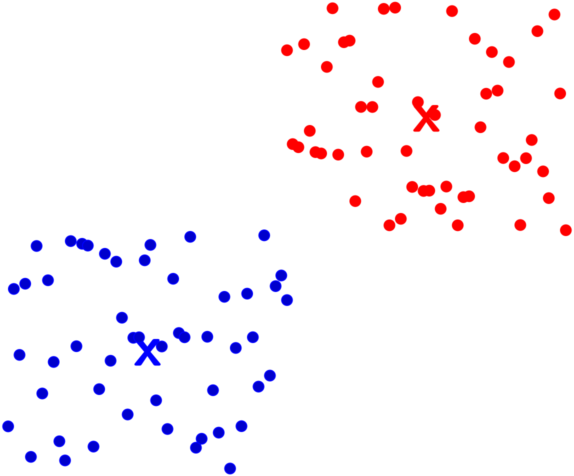
2 related questions:

1. How to choose the initial μ_is?
2. How to have more stable results?

Unfortunately, no miracle strategy for Q1. A common strategy:

- Several k-means with random initializations
- Majority vote

$k = 2$



Speeding up k-means

Each k-means iteration iterates over all the points in the dataset. This can be computationally expensive, especially if

- There are many points
- The point density is big

What to do to speed up the process?

Speeding up k-means

Each k-means iteration iterates over all the points in the dataset. This can be computationally expensive, especially if

- There are many points
- The point density is big

What to do to speed up the process?

Alternative: **Mini-batch** k-means. At each iteration

- Choose a subset of points
- Apply a k-means iteration

Number of clusters

In some applications, you know how many clusters you want. In this case, k is **easy to set**.

In other applications, we don't know the optimal number of classes we want. Ideally, we would like k to be selected automatically.

There is always some ambiguity in selection the *optimal* number of clusters. This is normal: When doing unsupervised learning, there is necessarily some inherent subjectivity in the labeling process!

Number of clusters

That being said, it is possible to define some criterias to determine whether k_1 is a better number of clusters than k_2. We can use the sum of squared errors to the centroids:

$$\mathrm{SSE}(k) = \sum_{i=1}^{k} \sum_{x \in S_i} \|x - \mu_i\|_2^2$$

And apply the **Elbow method**.

Number of clusters

That being said, it is possible to define some criterias to determine whether k_1 is a better number of clusters than k_2. We can use the sum of squared errors to the centroids:

$$\mathrm{SSE}(k) = \sum_{i=1}^{k} \sum_{x \in S_i} \|x - \mu_i\|_2^2$$

And apply the **Elbow method**.

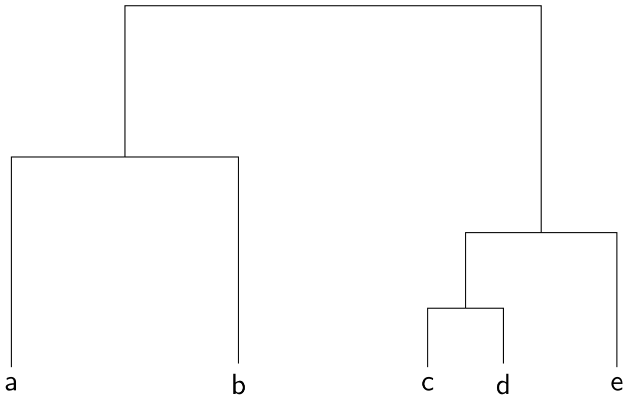
Note that this is not a miracle solution.

Hierarchical clustering

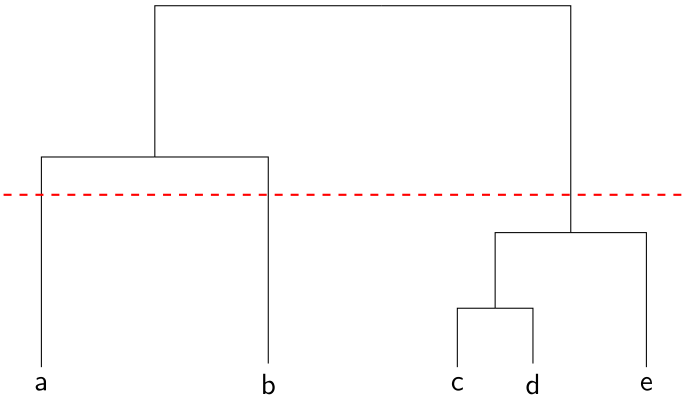
Hierarchical clustering

But then, which partition do we use?

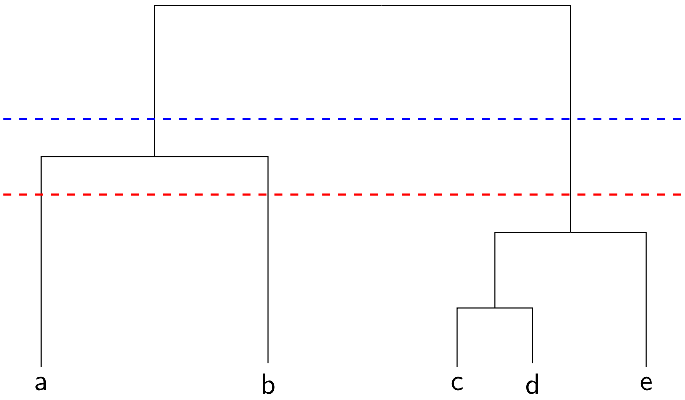
Hierarchical clustering: Dendograms



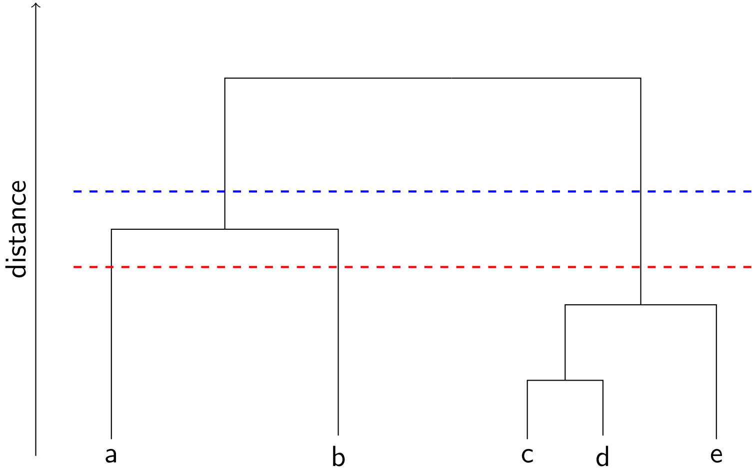
Hierarchical clustering: Dendograms



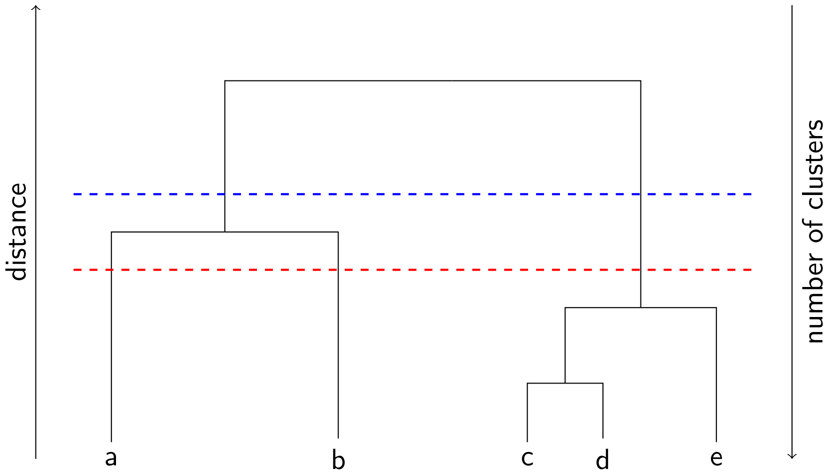
Hierarchical clustering: Dendograms



Hierarchical clustering: Dendograms



Hierarchical clustering: Dendograms



Hierarchical clustering, pros & cons

Conclusion

Clustering methods create groups of unlabeled points.

Conclusion

Clustering methods create groups of unlabeled points.

They rely on:

- A similarity measure (*e.g.* the Euclidian norm)
- A few parameters, *e.g.*
  - The number of clusters for k-means
  - The dendogram cut for hierarchical clustering

Conclusion

Clustering methods create groups of unlabeled points.

They rely on:

- A similarity measure (*e.g.* the Euclidian norm)
- A few parameters, *e.g.*
  - The number of clusters for k-means
  - The dendogram cut for hierarchical clustering

Clusters can be represented by a dendogram.

Thank you! Questions