Introduction to Machine Learning (67577)
Lecture 12

Shai Shalev-Shwartz

School of CS and Engineering,
The Hebrew University of Jerusalem

Clustering

Clustering

- One of the most widely used techniques for exploratory data analysis
- Unsupervised learning: finding meaningful patterns in data

What is Clustering ?

- Intuitively: grouping a set of objects such that

What is Clustering ?

- Intuitively: grouping a set of objects such that
  - similar objects end up in the same group

What is Clustering ?

- Intuitively: grouping a set of objects such that
  - similar objects end up in the same group
  - dissimilar objects are separated into different groups

What is Clustering ?

- Intuitively: grouping a set of objects such that
  - similar objects end up in the same group
  - dissimilar objects are separated into different groups
- Imprecise, possibly ambiguous, definition

What is Clustering ?

- Intuitively: grouping a set of objects such that
  - similar objects end up in the same group
  - dissimilar objects are separated into different groups
- Imprecise, possibly ambiguous, definition
- Quite surprisingly, it is not at all clear how to come up with a more rigorous definition ...

Why is it hard to define what is clustering ?

- Our intuitive objective

Why is it hard to define what is clustering ?

- Our intuitive objective
  - similar objects end up in the same group

Why is it hard to define what is clustering ?

- Our intuitive objective
  - similar objects end up in the same group
  - dissimilar objects are separated into different groups

Why is it hard to define what is clustering ?

- Our intuitive objective
  - similar objects end up in the same group
  - dissimilar objects are separated into different groups
- Problem I: Two contradicting objectives: Similarity is not a transitive relation while class membership is transitive

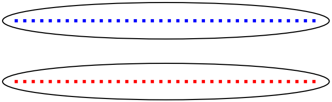
Why is it hard to define what is clustering ?

- Our intuitive objective
  - similar objects end up in the same group
  - dissimilar objects are separated into different groups
- Problem I: Two contradicting objectives: Similarity is not a transitive relation while class membership is transitive
- Problem II: Lack of ground truth

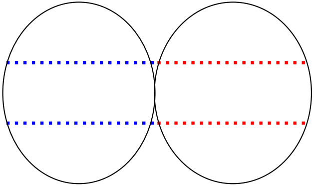
Why is it hard to define what is clustering ?

Why is it hard to define what is clustering ?

similar objects in same group



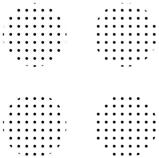
dissimilar objects are separated



Why is it hard to define what is clustering ?

Lack of ground truth:

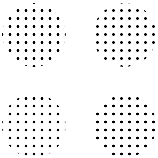
Cluster these points into **two** clusters.



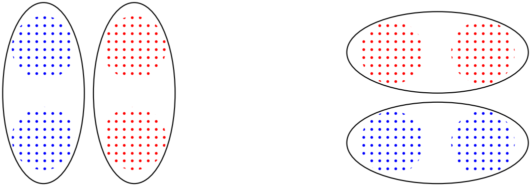
Why is it hard to define what is clustering ?

Lack of ground truth:

Cluster these points into **two** clusters.



We have two well justifiable solutions:



A clustering model

- Input: a set of elements and a distance $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$

A clustering model

- Input: a set of elements and a distance $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$
- Output: Partition of $\mathcal{X}$: $\mathcal{X} = \bigcup_{i=1}^{k} C_i$ with $C_i \cap C_j = \emptyset$

A clustering model

- Input: a set of elements and a distance $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$
- Output: Partition of $\mathcal{X}$: $\mathcal{X} = \bigcup_{i=1}^k C_i$ with $C_i \cap C_j = \emptyset$
- Remarks:

A clustering model

- Input: a set of elements and a distance $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$
- Output: Partition of $\mathcal{X}$: $\mathcal{X} = \bigcup_{i=1}^{k} C_i$ with $C_i \cap C_j = \emptyset$
- Remarks:
  - Sometimes the input also contains the number of desired clusters, k.

A clustering model

- Input: a set of elements and a distance $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$
- Output: Partition of $\mathcal{X}$: $\mathcal{X} = \bigcup_{i=1}^{k} C_i$ with $C_i \cap C_j = \emptyset$
- Remarks:
  - Sometimes the input also contains the number of desired clusters, k.
  - Sometimes, the output is a dendrogram (from Greek dendron = tree, gramma = drawing)

Outline

1. Linkage-based Clustering Algorithms
2. The k-means family

Linkage-based clustering

- Start from the trivial clustering that has each data point as a single-point cluster

Linkage-based clustering

- Start from the trivial clustering that has each data point as a single-point cluster
- Repeatedly merge the "closest" clusters of the previous clustering

Linkage-based clustering

- Start from the trivial clustering that has each data point as a single-point cluster
- Repeatedly merge the “closest” clusters of the previous clustering
- End when the result is the trivial clustering in which all of the domain points share one large cluster

Linkage-based clustering

- Start from the trivial clustering that has each data point as a single-point cluster
- Repeatedly merge the "closest" clusters of the previous clustering
- End when the result is the trivial clustering in which all of the domain points share one large cluster

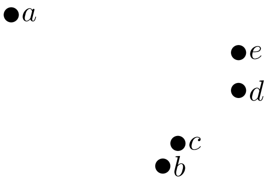
Linkage-based clustering

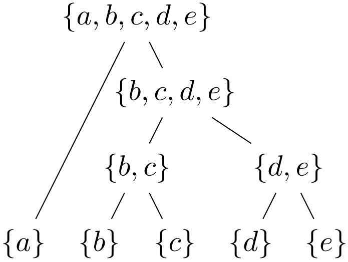
- Start from the trivial clustering that has each data point as a single-point cluster
- Repeatedly merge the "closest" clusters of the previous clustering
- End when the result is the trivial clustering in which all of the domain points share one large cluster

Different linkage methods differ in how they extend the distance function d from points to clusters:

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

The output of linkage clustering is a Dendrogram





Outline

Cost Minimization Clustering

- Define a function, G, that takes as input $(\mathcal{X}, d)$ and a proposed clustering $C = (C_1, \ldots, C_k)$, and returns a quality (positive scalar)
- Return the clustering C that minimizes $G((\mathcal{X}, d), C)$

The k-means objective

$$G_{\mathrm{k-means}}((\mathcal{X}, d), (C_1, \ldots, C_k)) = \min_{\mu_1, \ldots \mu_k \in \mathcal{X}'} \sum_{i=1}^{k} \sum_{x \in C_i} d(x, \mu_i)^2$$

The k-means objective

$$G_{\mathrm{k-means}}((\mathcal{X}, d), (C_1, \ldots, C_k)) = \min_{\mu_1, \ldots \mu_k \in \mathcal{X}'} \sum_{i=1}^{k} \sum_{x \in C_i} d(x, \mu_i)^2$$

- $\mathcal{X} \subset \mathcal{X}'$ (e.g., data points are in $\mathbb{R}^d$)

The k-means objective

$$G_{\mathrm{k-means}}((\mathcal{X}, d), (C_1, \ldots, C_k)) = \min_{\mu_1, \ldots \mu_k \in \mathcal{X}'} \sum_{i=1}^{k} \sum_{x \in C_i} d(x, \mu_i)^2$$

- $\mathcal{X} \subset \mathcal{X}'$ (e.g., data points are in $\mathbb{R}^d$)
- If we define the centroid of C_i as

$$\mu_i(C_i) = \operatorname*{argmin}_{\mu \in \mathcal{X}'} \sum_{x \in C_i} d(x, \mu)^2 \ .$$

Then, the k-means objective becomes

$$G_{\mathrm{k-means}}((\mathcal{X}, d), (C_1, \ldots, C_k)) = \sum_{i=1}^{k} \sum_{x \in C_i} d(x, \mu_i(C_i))^2 \ .$$

Other objectives from the k-means family

k-Medoids:

$$G_{\mathrm{K-medoid}}((\mathcal{X}, d), (C_1, \ldots, C_k)) = \min_{\mu_1, \ldots \mu_k \in \mathcal{X}} \sum_{i=1}^{k} \sum_{x \in C_i} d(x, \mu_i)^2 \ .$$

k-median:

$$G_{\mathrm{K-median}}((\mathcal{X}, d), (C_1, \ldots, C_k)) = \min_{\mu_1, \ldots \mu_k \in \mathcal{X}} \sum_{i=1}^{k} \sum_{x \in C_i} d(x, \mu_i) \ .$$

How to solve the k-means optimization problem?

NP hard ...

How to solve the k-means optimization problem?

NP hard ...

A good practical heuristic is Lloyd's algorithm

How to solve the k-means optimization problem?

NP hard ...

A good practical heuristic is Lloyd's algorithm

k-`means`

input: $\mathcal{X} \subset \mathbb{R}^n$; Number of clusters k

initialize: Randomly choose initial centroids $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k$

repeat until convergence

$\forall i \in [k]$ set $C_i = \{\mathbf{x} \in \mathcal{X} : i = \operatorname{argmin}_j \|\mathbf{x} - \boldsymbol{\mu}_j\|\}$

(break ties in some arbitrary manner)

$\forall i \in [k]$ update $\boldsymbol{\mu}_i = \frac{1}{|C_i|} \sum_{\mathbf{x} \in C_i} \mathbf{x}$

Summary

- Clustering is a very intuitive task, but there's no good rigorous defintion
- Linkage based family and k-means family
- There are many other clustering methods: spectral clustering, information bottleneck, ...