Lecture 12

Unsupervised Learning: Clustering

Rishabh Iyer

University of Texas at Dallas

Based on the slides of Nick Rouzzi and Vibhav Gogate

Clustering

Clustering systems:

$D = \{ x^{(1)}, x^{(2)}, \ldots, x^{(m)} \}$

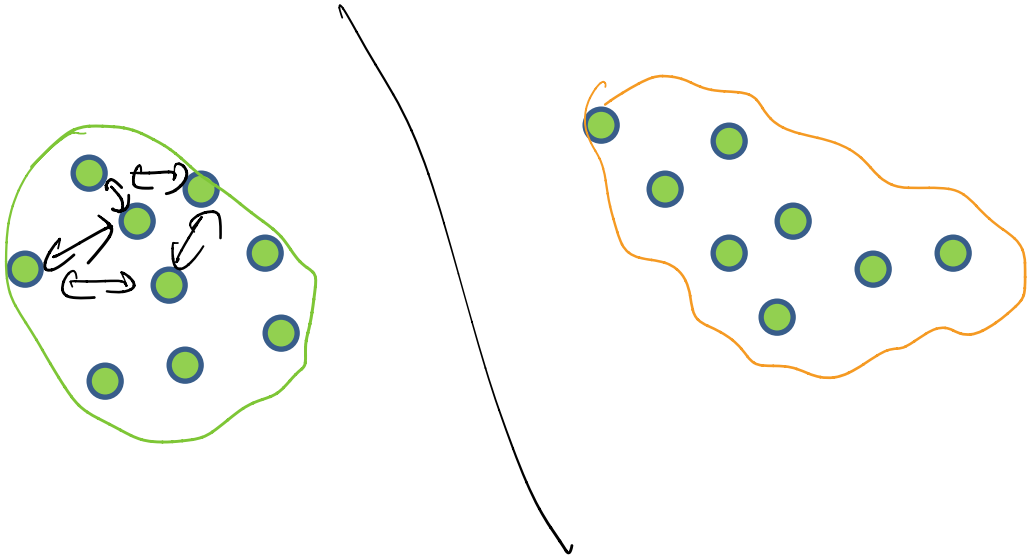
n Features

D contains no labels

- Unsupervised learning
- Requires data, but no labels
- Detect patterns, e.g., in
  - Group emails or search results
  - Customer shopping patterns
- Useful when don't know what you're looking for...
  - But often get gibberish

Clustering

- Want to group together parts of a dataset that are close together in some metric
  - Useful for finding the important parameters/features of a dataset

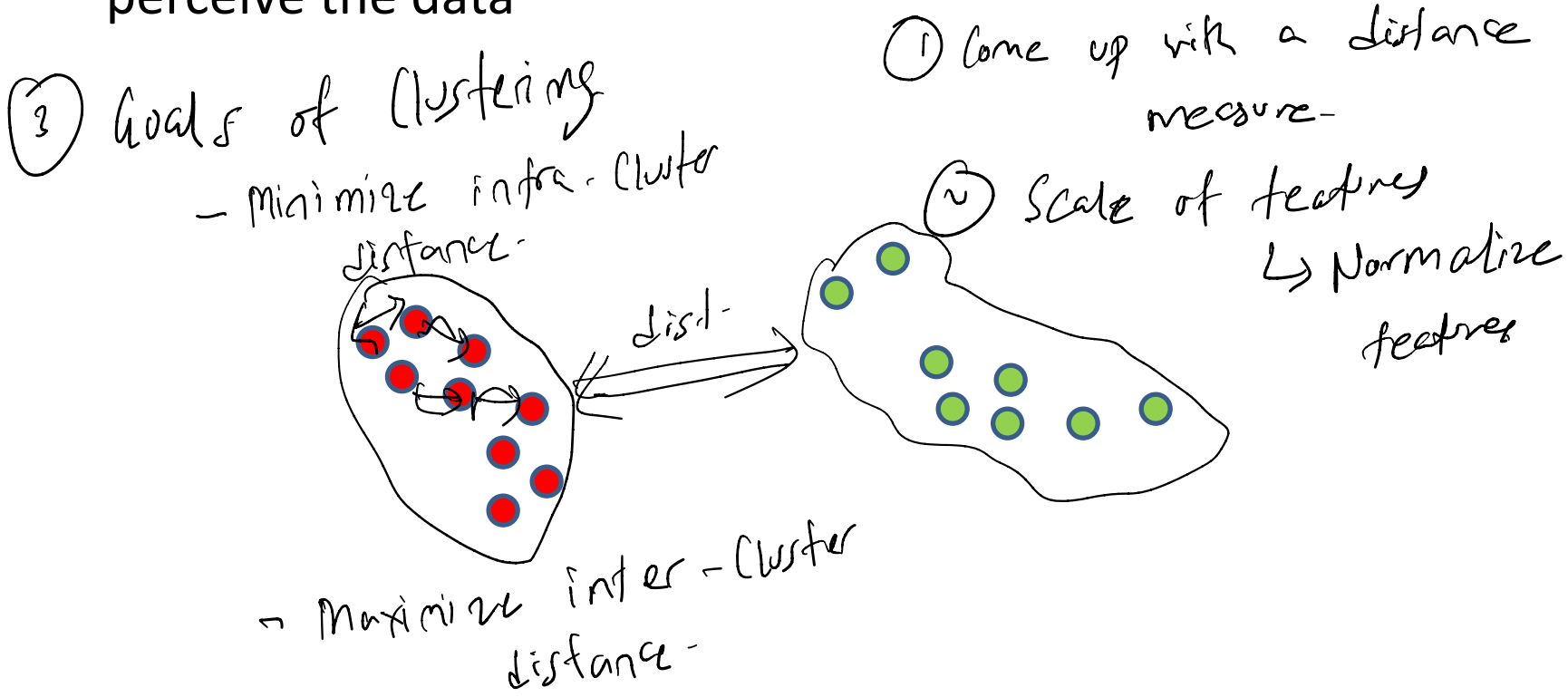


Clustering

- Want to group together parts of a dataset that are close together in some metric
  - Useful for finding the important parameters/features of a dataset

Clustering

- Intuitive notion of clustering is a somewhat ill-defined problem
  - Identification of clusters depends on the scale at which we perceive the data



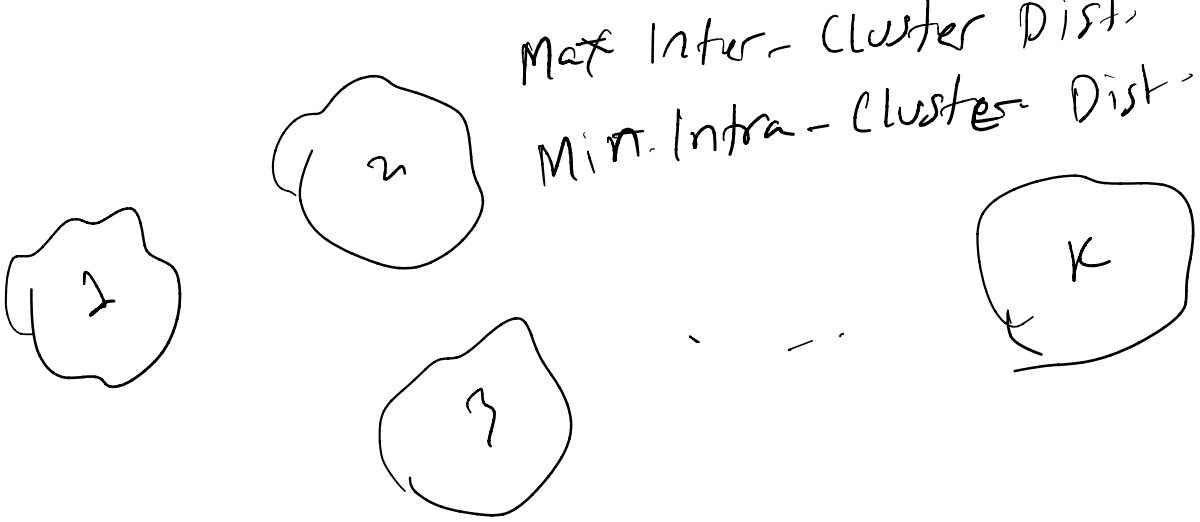
Clustering

- Intuitive notion of clustering is a somewhat ill-defined problem
  - Identification of clusters depends on the scale at which we perceive the data

Clustering

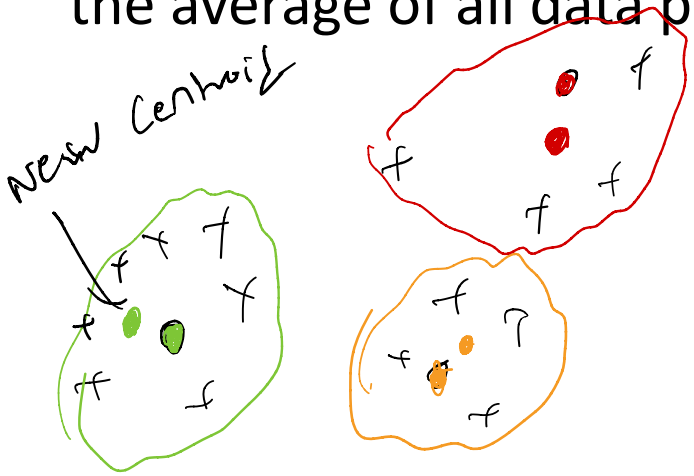
- Input: a collection of points $x^{(1)}, \ldots, x^{(m)} \in \mathbb{R}^n$, an integer k

- Output: A partitioning of the input points into k sets that minimizes some metric of closeness



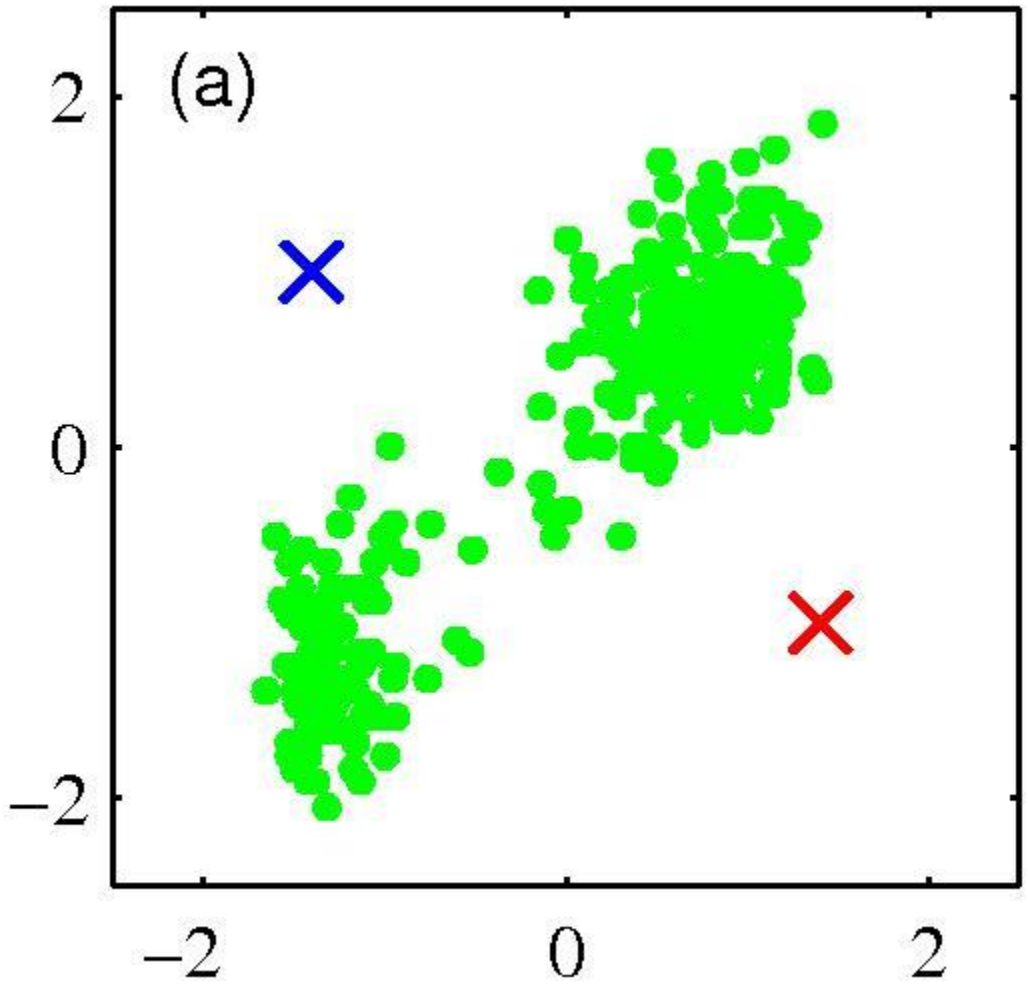
k-means Clustering

- Pick an initial set of k means (usually at random)
- Repeat until the clusters do not change:
  - Partition the data points, assigning each data point to a cluster based on the mean that is closest to it ← Assignment.
  - Update the cluster means so that the i^{th} mean is equal to the average of all data points assigned to cluster i



Partitioning
- Every point is covered by exactly one group

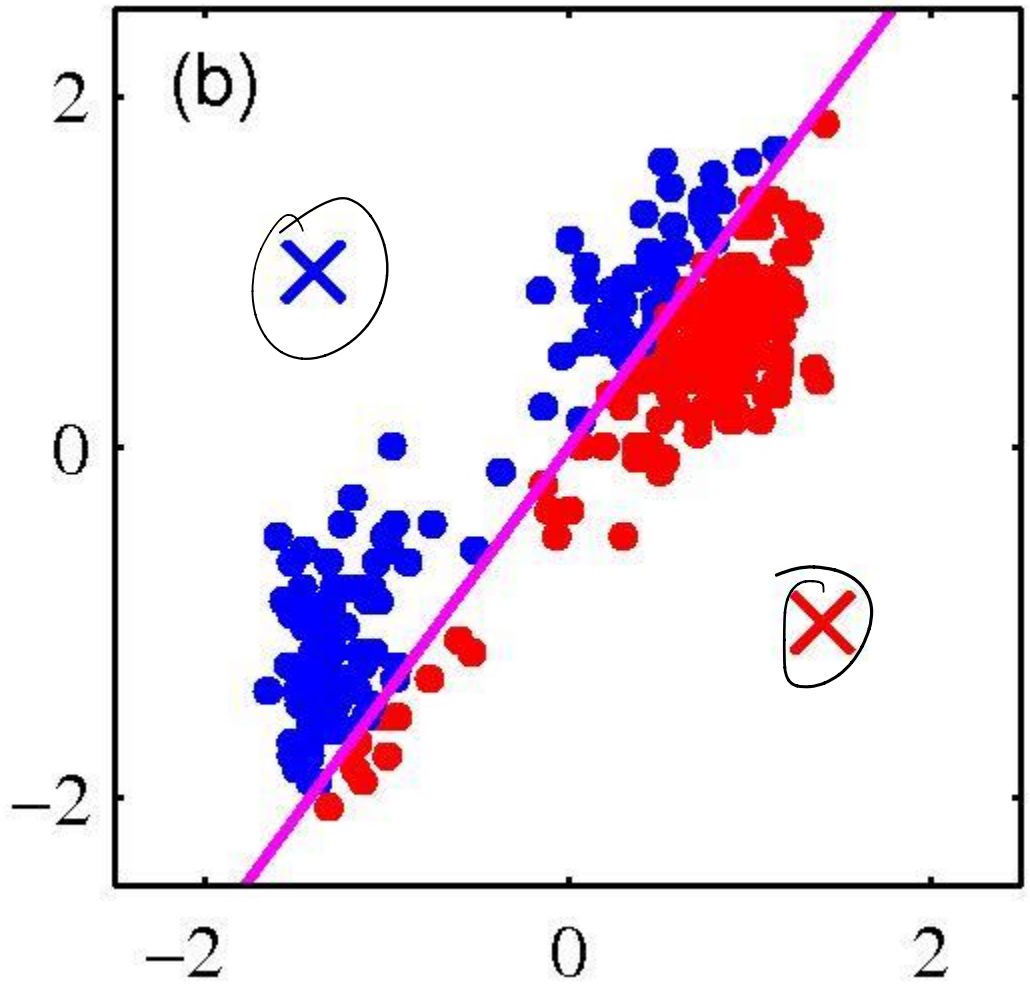
k-means clustering: Example



$K = 2$

Pick k random points as cluster centers (means)

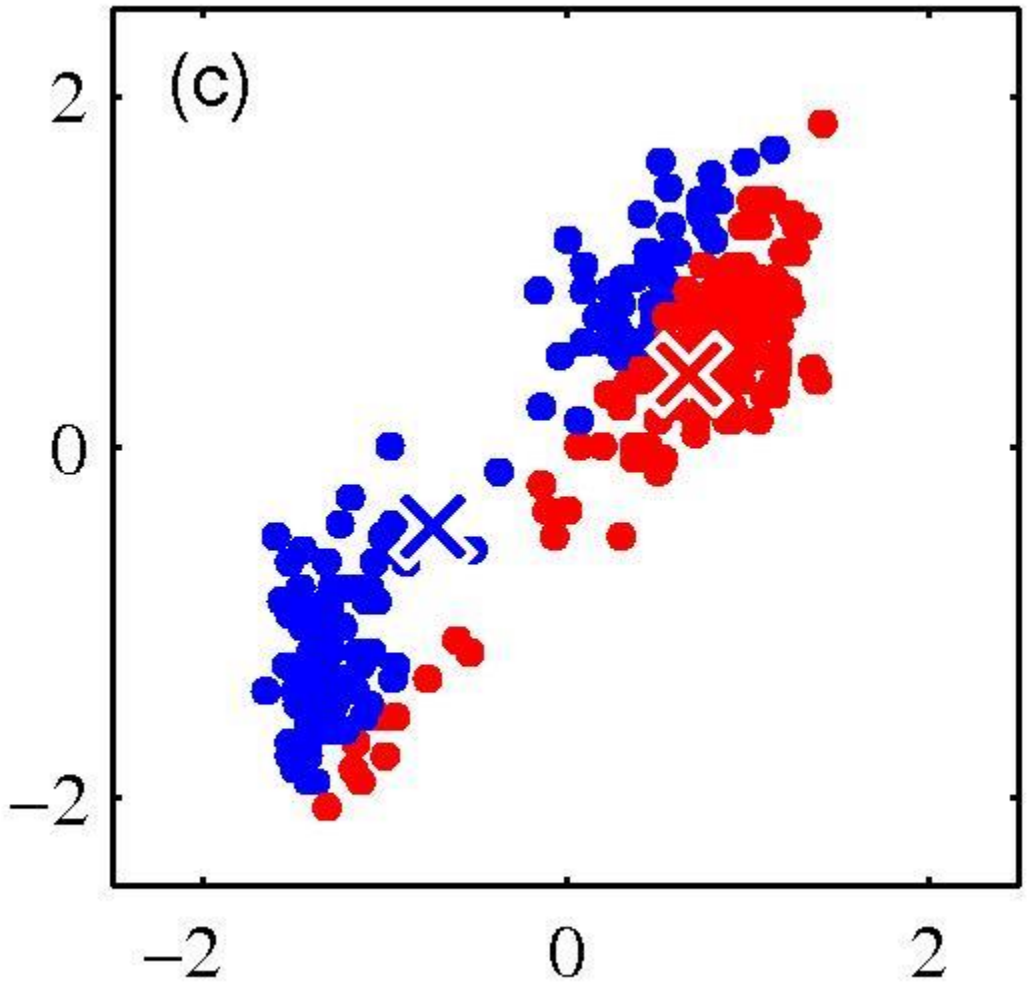
k-means clustering: Example



Iterative Step 1:

Assign data instances to closest cluster center

k-means clustering: Example



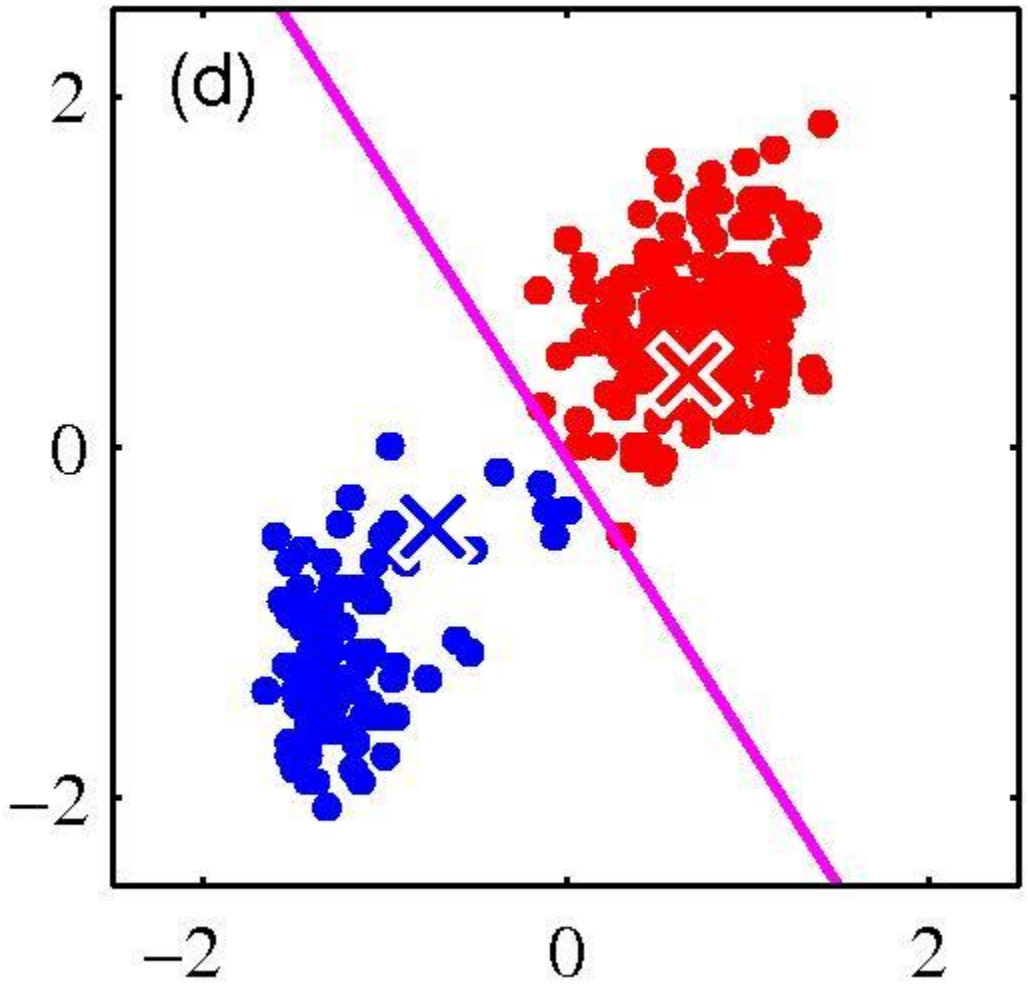
$x^{(1)}, \ldots, x^{(m)}$

$$\text{Mean} = \frac{\sum_{i=1}^{m} x^{(i)}}{M}$$

Iterative Step 2:

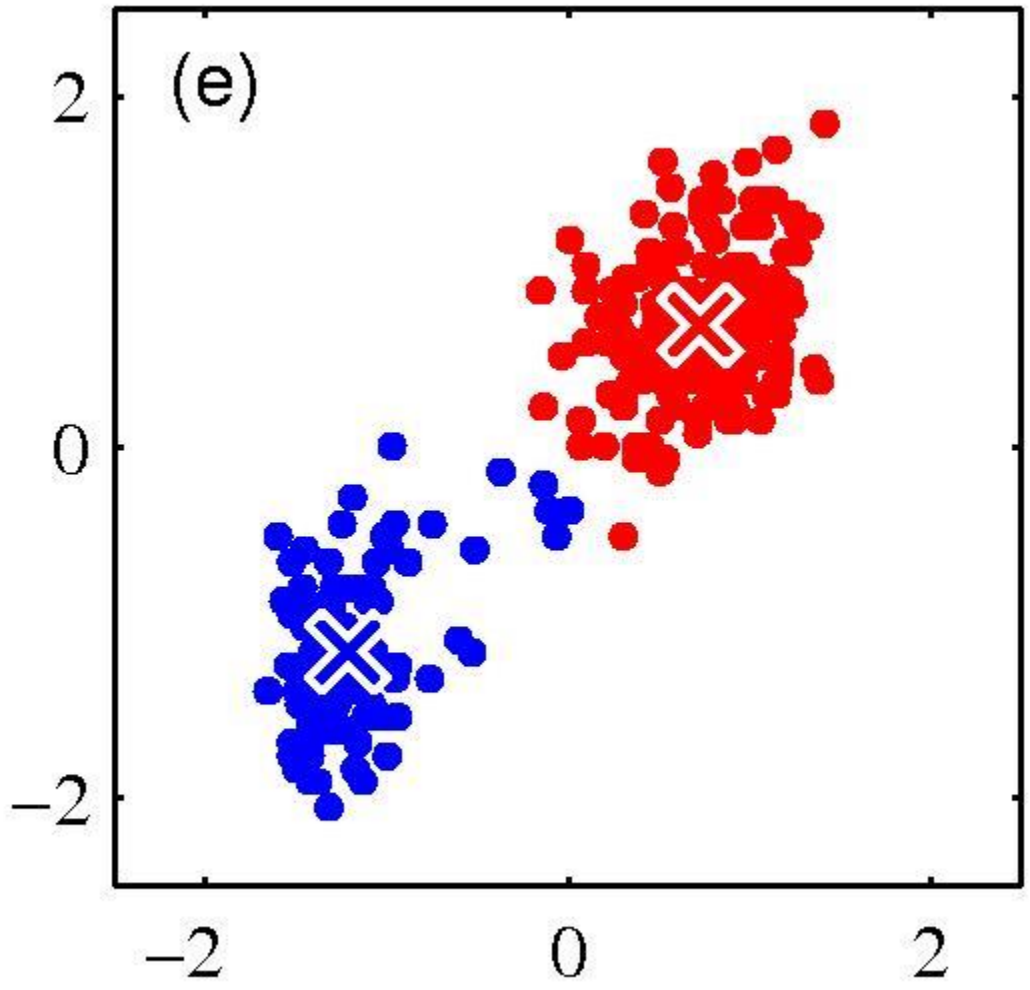
Change the cluster center to the average of the assigned points

k-means clustering: Example

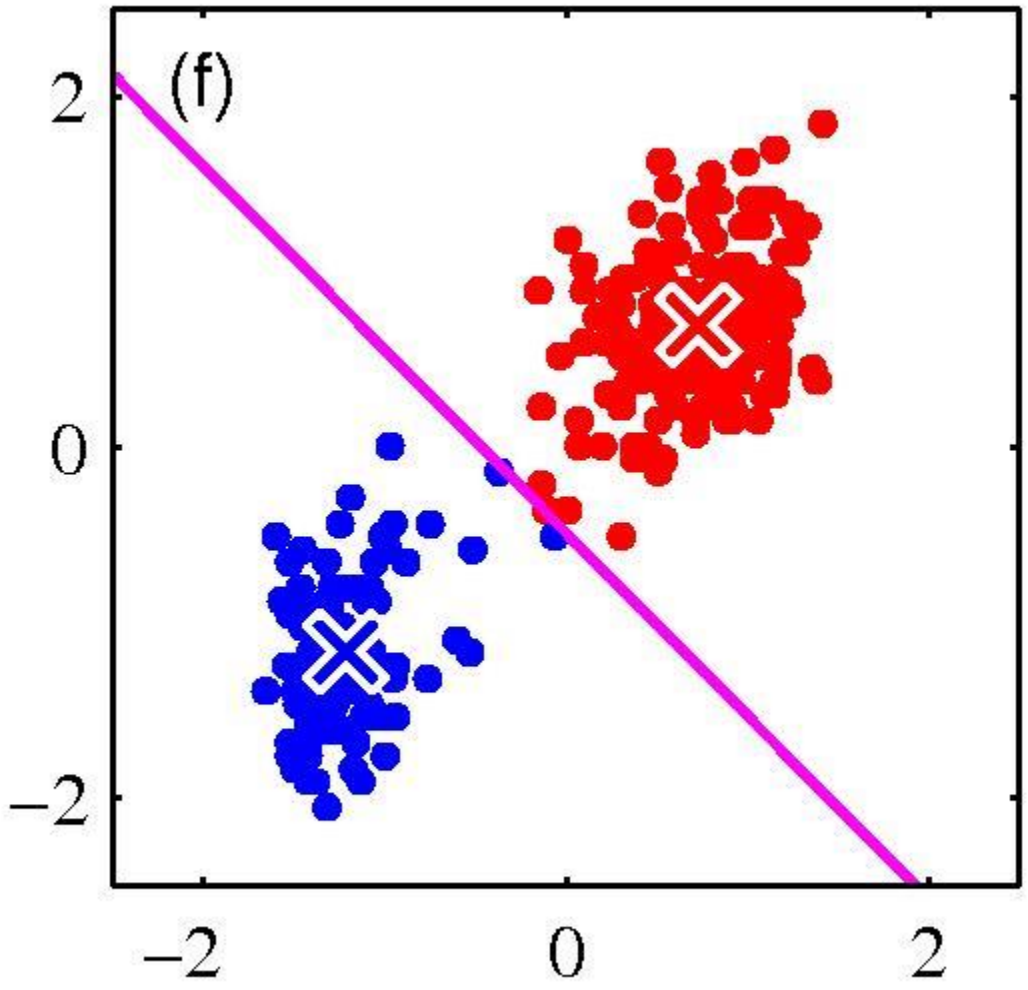


Repeat until convergence

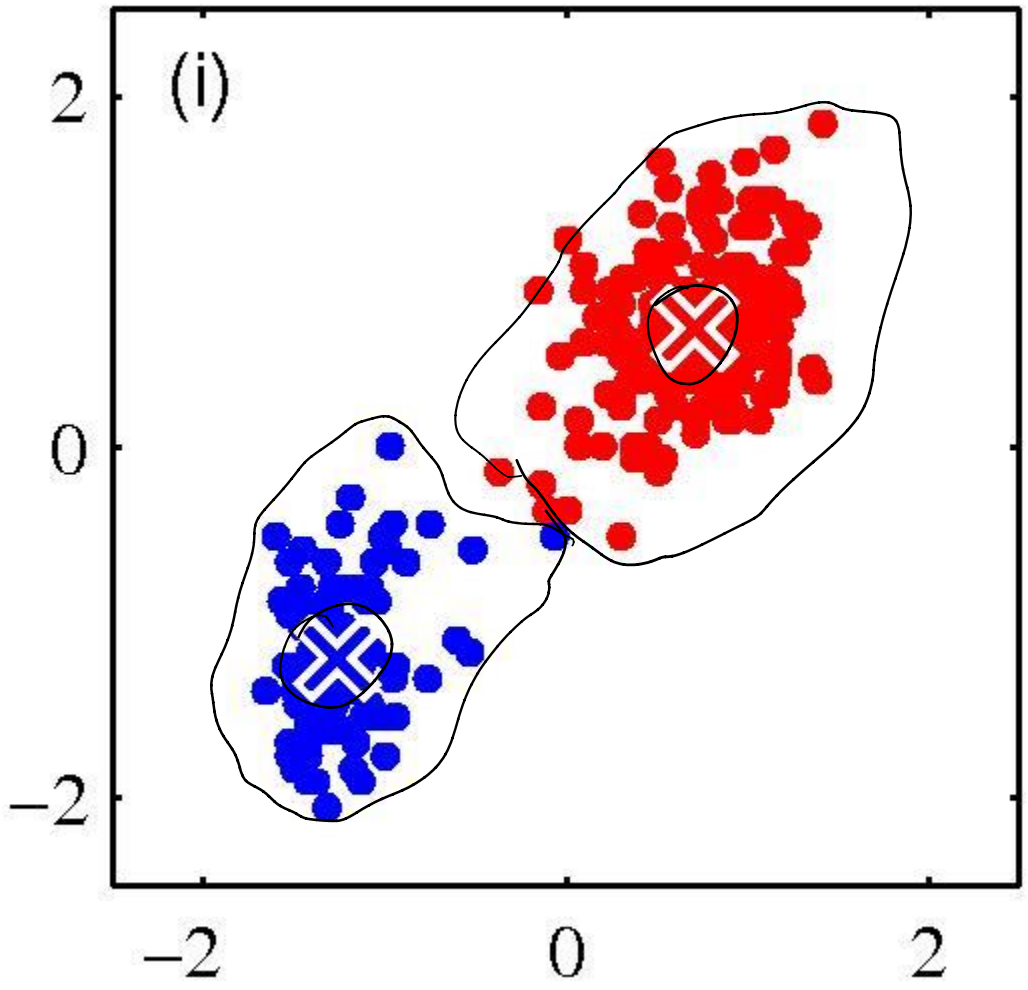
k-means clustering: Example



k-means clustering: Example



k-means clustering: Example



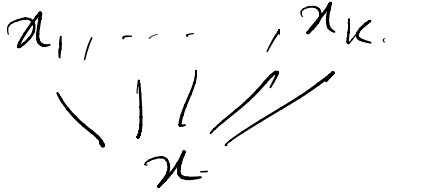
Stopping Criterion

$\forall i = 1:K,$

$\|\mu_i^{t+1} - \mu_i^t\|_2 \leq \varepsilon$

$C_i : x_j, j \in C_i$

$$\mu_i = \frac{\sum_{j \in C_i} x_j}{|C_i|}$$



$$\arg\min_{x \in \mathbb{R}^n} \underbrace{\sum_{i=1}^{k} \|x_i - x\|_2^2}_{g(x)}$$

$$\nabla g(x) = 2 \sum_{i=1}^{k} (x - x_i) = 0$$

$$\Rightarrow \quad k x = \sum_{i=1}^{k} x_i$$

$$\Rightarrow \quad x = \sum_{i=1}^{k} x_i / k$$

$C = \{x_1, x_2, x_3\}$

$$\mu = \underset{x}{\operatorname{argmin}} \sum_{i=1}^{3} \|x - x_i\|^2$$

$$= \underset{x}{\operatorname{argmin}} \; \|x - x_1\|^2 + \|x - x_2\|^2 + \|x - x_3\|^2$$

$$= \underset{x}{\operatorname{argmin}} \left[3\|x\|^2 - 2\left(\langle x, x_1\rangle + \langle x, x_2\rangle + \langle x, x_3\rangle\right)\right]$$

$$= \underset{x}{\operatorname{argmin}} \; 3\|x\|^2 - 2\langle x, x_1 + x_2 + x_3\rangle$$

$$6x - 2(x_1 + x_2 + x_3)$$

$$\Rightarrow \boxed{x = \frac{x_1 + x_2 + x_3}{3}}$$

$C : \{x_1, x_2, \ldots, x_l\}$

$$\mu = \frac{x_1 + x_2 + \ldots + x_l}{l} = \frac{\sum_{i=1}^{l} x_i}{l}$$

k-Means for Segmentation

$k = 2$

Goal of segmentation is to partition an image into regions, each of which has reasonably homogenous visual appearance

Original

k-Means for Segmentation

$k = 2$

$k = 3$

Original

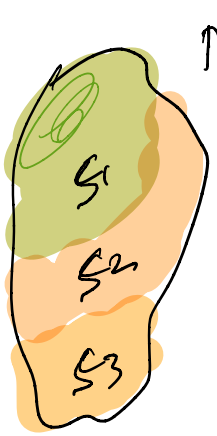
k-Means for Segmentation

$k = 2$	$k = 3$	$k = 10$	Original
			
			

k-means Clustering as Optimization

- Minimize the distance of each input point to the mean of the cluster/partition that contains it

$$\min_{S_1,\ldots,S_k} \sum_{i=1}^{k} \sum_{j \in S_i} \left\| x^{(j)} - \mu_i \right\|^2$$



where

- $S_i \subseteq \{1, \ldots, M\}$ is the i^{th} cluster
- $S_i \cap S_j = \emptyset$ for $i \neq j$, $\cup_i S_i = \{1, \ldots, M\}$
- μ_i is the centroid of the i^{th} cluster

$S_1 \cup S_2 \cup S_3 = [1 : M]$

$S_i \cap S_j = \emptyset$

m points
2 clusters

$$\|x_1 - \mu_1\|_2^2 + \|x_2 - \mu_1\|_2^2 + \|x_3 - \mu_2\|_2^2 + \|x_4 - \mu_2\|_2^2$$

$$\mu_1 = \frac{x_1 + x_4}{2}$$

$$\mu_2 = \frac{x_3 + x_4}{2}$$

Clusterings

$$= 2^m - 2$$

$$= \binom{m}{1} + \binom{m}{2} + \cdots \binom{m}{m/2}$$

k-means Clustering as Optimization

- Minimize the distance of each input point to the mean of the cluster/partition that contains it

$$\min_{S_1,\dots,S_k} \sum_{i=1}^{k} \sum_{j \in S_i} \left\| x^{(j)} - \mu_i \right\|^2$$

where

- $S_i \subseteq \{1, \dots, M\}$ is the i^{th} cluster
- $S_i \cap S_j = \emptyset$ for $i \neq j$, $\cup_i S_i = \{1, \dots, n\}$
- μ_i is the centroid of the i^{th} cluster

Exactly minimizing this function is NP-hard (even for $k = 2$)

k-means Clustering

- The k-means clustering algorithm performs a block coordinate descent on the objective function

alternating min

$$f(S_i, \mu_i) = \sum_{i=1}^{k} \sum_{j \in S_i} \left\| x^{(j)} - \mu_i \right\|^2$$

- This is not a convex function: could get stuck in local minima

$$f(S_1, \ldots S_k, \mu_1, \ldots \mu_k) = \sum_{i=1}^{k} \sum_{j \in S_i} \left\| x^{(j)} - \mu_i \right\|^2$$

Alternating Minimization

$$\min_{x,y} f(x,y)$$

k-means

$x = S_1, \ldots, S_k$

$y = M_1 \ldots M_k$

Initialize x^0, y^0

Iterate: $t = 1 : T$

$$x^{t+1} = \min_{x} f(x, y^t)$$ ← assignment

$$y^{t+1} = \min_{y} f(x^{t+1}, y)$$ ← means

until $\left| f(x^{t+1}, y^{t+1}) - f(x^t, y^t) \right| < \varepsilon$

Thm

if each step in Alt-Min can be solved exactly

$$f(x^{t+1}, y^{t+1}) \leq f(x^t, y^t)$$

Alt-Min for k-means

$$f(S_1, \ldots, S_k, \mu_1, \ldots, \mu_k) = \sum_{i=1}^{k} \sum_{j \in S_i} \|x^{(j)} - \mu_i\|^2$$

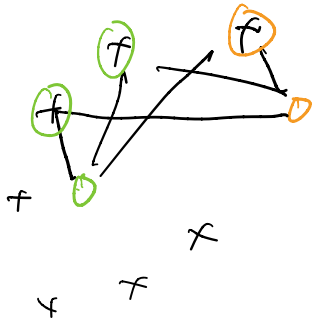
$$\min_{\mu_1 \ldots \mu_k} \sum_{i=1}^{k} \sum_{j \in S_i} \|x^{(j)} - \mu_i\|_2^2$$

$$\min_{\mu_i} \sum_{j \in S_i} \|x^{(j)} - \mu_i\|^2, \quad \forall\ i = 1:k.$$

↑ computing / centroids

$$\min_{S_1 \ldots S_k} \sum_{i=1}^{k} \sum_{j \in S_i} \|x^{(j)} - \mu_i\|^2 \leftarrow \text{Assignment}$$

$S_i : \{ j : \|x_j - \mu_i\| \leq \|x_j - \mu_\ell\|$ for all other centroids $\ell \}$



k-Means as Optimization

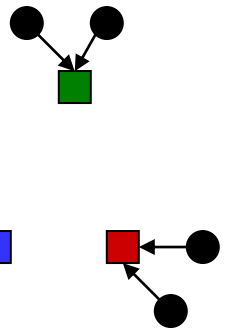
- Consider the k-means objective function

$$\phi(x, S, \mu) = \sum_{i=1}^{k} \sum_{j \in S_i} \left\| x^{(j)} - \mu_i \right\|^2$$

points

cluster assignments

cluster means



- Two stages each iteration
  - Update cluster assignments: fix means μ, change assignments S
  - Update means: fix assignments S, change means μ

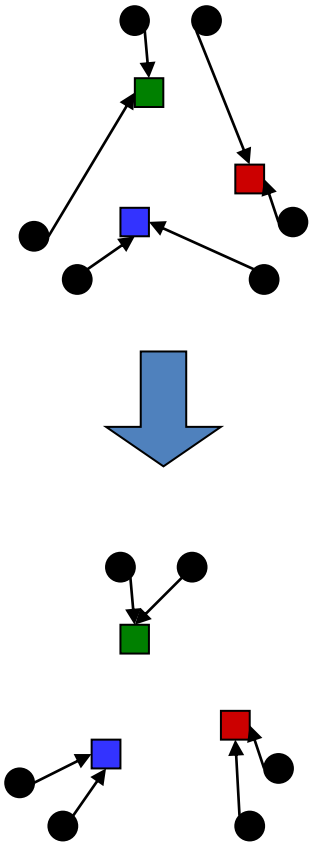
Phase I: Update Assignments

- For each point, re-assign to closest mean, $x^{(j)} \in S_i$ if

$$j \in \arg\min_i \left\| x^{(j)} - \mu_i \right\|^2$$

- Can only decrease ϕ as the sum of the distances of all points to their respective means must decrease

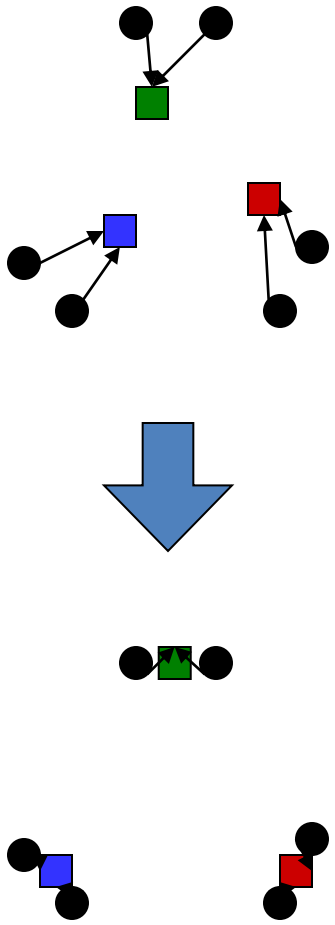
$$\phi(x, S, \mu) = \sum_{i=1}^{k} \sum_{j \in S_i} \left\| x^{(j)} - \mu_i \right\|^2$$



Phase II: Update Means

- Move each mean to the average of its assigned points

$$\mu_i = \sum_{j \in S_i} \frac{x^{(j)}}{|S_i|}$$



- Also can only decrease total distance…
  - Why?

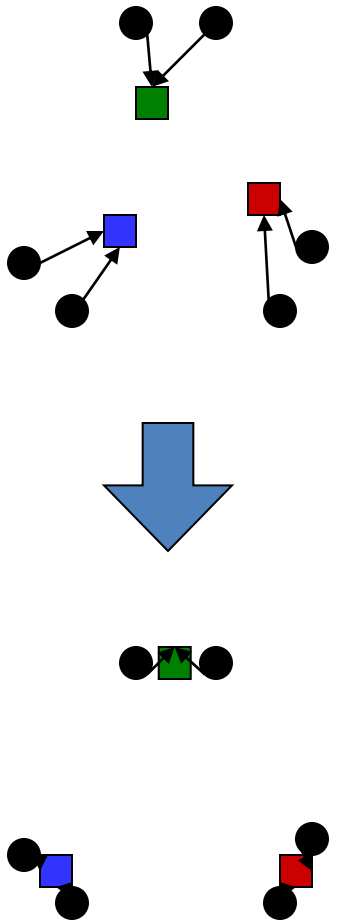
$$\mu_i = \min_{\mu} \sum_{j \in S_i} \|\mu - x^{(j)}\|^2$$

$$= \sum_{j \in S_i} \frac{x^{(j)}}{|S_i|}$$

Phase II: Update Means

- Move each mean to the average of its assigned points

$$\mu_i = \sum_{j \in S_i} \frac{x^{(j)}}{|S_i|}$$



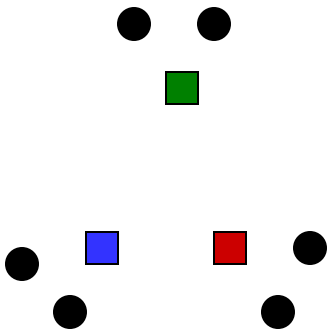
- Also can only decrease total distance…
  - The point y with minimum squared Euclidean distance to a set of points is their mean

Initialization

- K-means is sensitive to initialization
  - It does matter what you pick!
  - What can go wrong?

Initialization

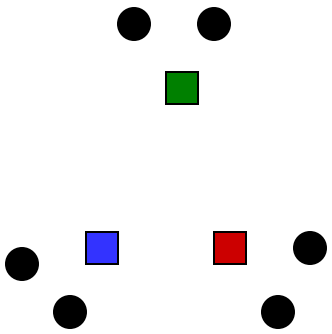
- K-means is sensitive to initialization
  - It does matter what you pick!
  - What can go wrong?

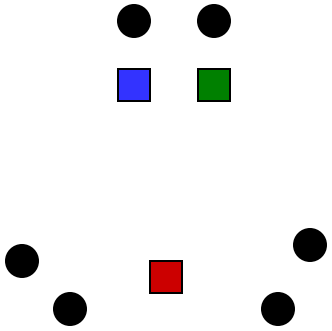


Initialization

- K-means is sensitive to initialization
  - It does matter what you pick!
  - What can go wrong?
    - Various schemes to help alleviate this problem: initialization heuristics

[K-means++ - Theoretical]





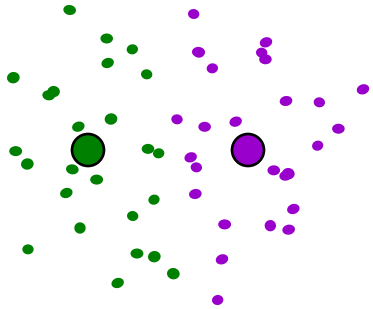
k-means Clustering (selecting k)

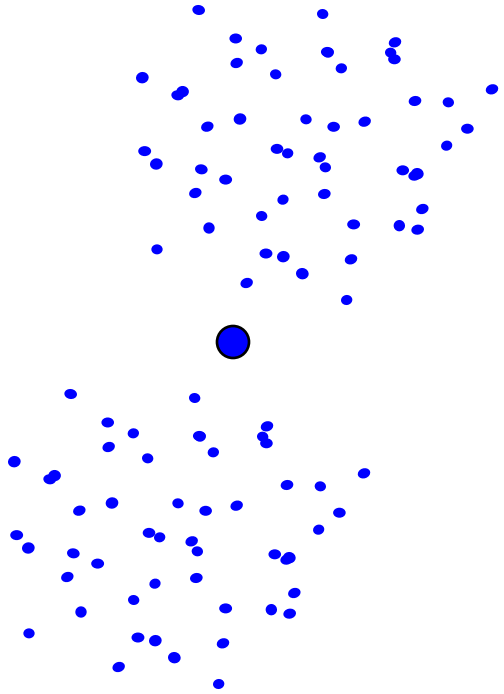
- Not clear how to figure out the "best" k in advance
- Want to choose k to pick out the interesting clusters, but not to overfit the data points
  - Large k doesn't necessarily pick out interesting clusters
  - Small k can result in large clusters than can be broken down further

Recent work

Local Optima ⇒ Bad Initialization

- Assignment
- centroid





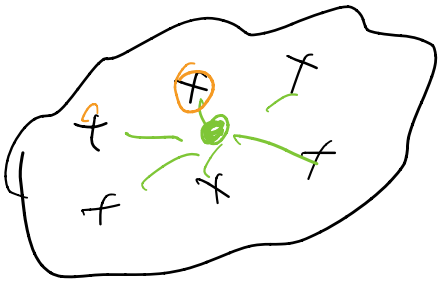
k-Means Summary

$\varepsilon \rightarrow O\left(\frac{kMn}{\varepsilon}\right)$

- Guaranteed to converge
  - But not to a global optimum
- Choice of k and initialization can greatly affect the outcome
- Runtime: $O(kMn)$ per iteration (Bound)
- Popular because it is fast, though there are other clustering methods that may be more suitable depending on your data

K-mediods Clustering and Extensions

- Very similar to k-means, except that the centroid is one of the examples from the cluster
- The update means step therefore involves finding the point within the cluster which has the minimum average distance to the rest



Centroid : $\min_{n \in S_i} \sum_{j \in S_i} \|x^{(j)} - n\|^2$

No-closed form.

- Extensions of k-means: Other distance measures, e.g. the Bregman divergence

$d(x, y) = \|x - y\|^2$

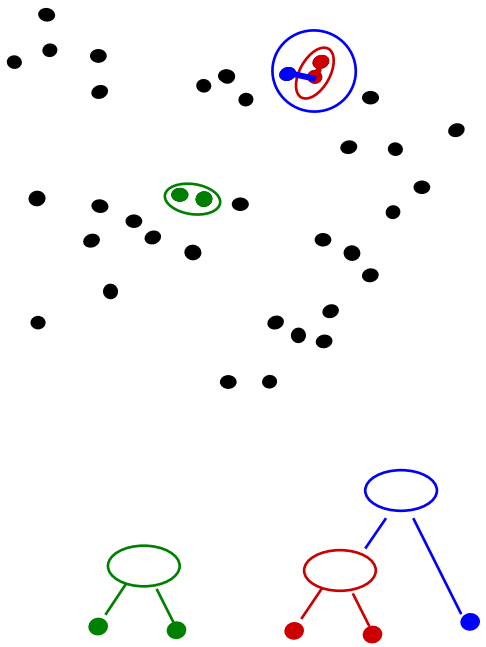
$O[|S_i|^2]$

$x_1 \quad x_2 \; x_3$

$$\text{arg}\min \left\{ \begin{array}{l} \|x_1 - x_2\|^2 + \|x_1 - x_3\|^2 \\ \|x_2 - x_1\|^2 + \|x_2 - x_3\|^2 \\ \|x_3 - x_1\|^2 + \|x_3 - x_2\|^2 \end{array} \right\}$$

Hierarchical Clustering

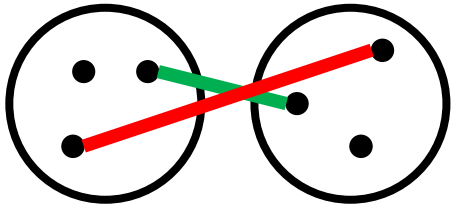
- Agglomerative clustering
  - Incrementally build larger clusters out of smaller clusters
- Algorithm:
  - Maintain a set of clusters
  - Initially, each instance in its own cluster
  - Repeat:
    - Pick the two closest clusters
    - Merge them into a new cluster
    - Stop when there is only one cluster left



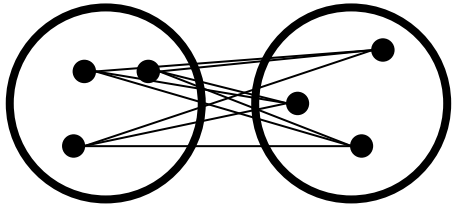
- Produces not one clustering, but a family of clusterings represented by a dendrogram

Agglomerative Clustering

- How should we define "closest" for clusters with multiple elements?



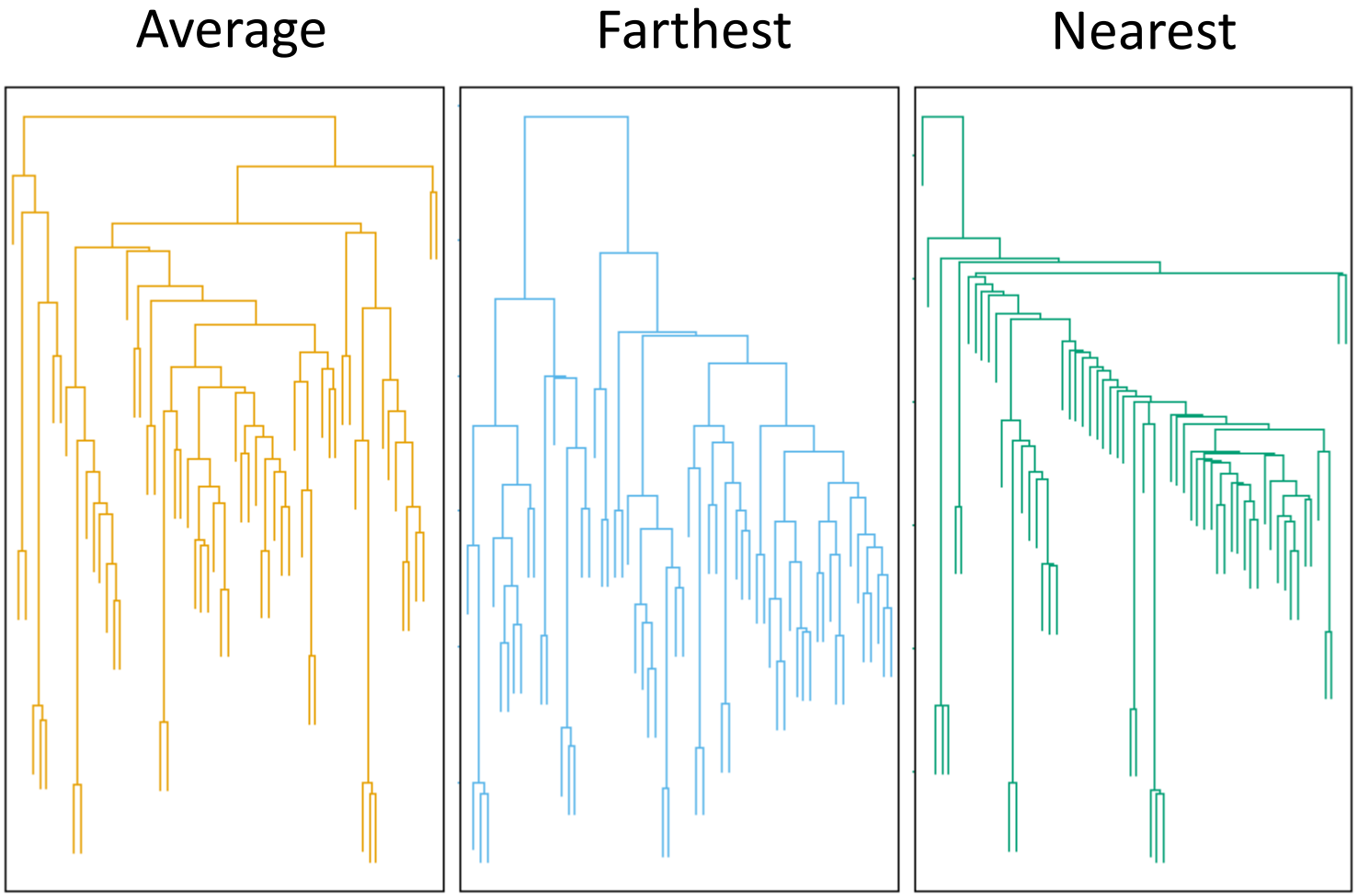
Closest / farthest pair



Average of all pairs

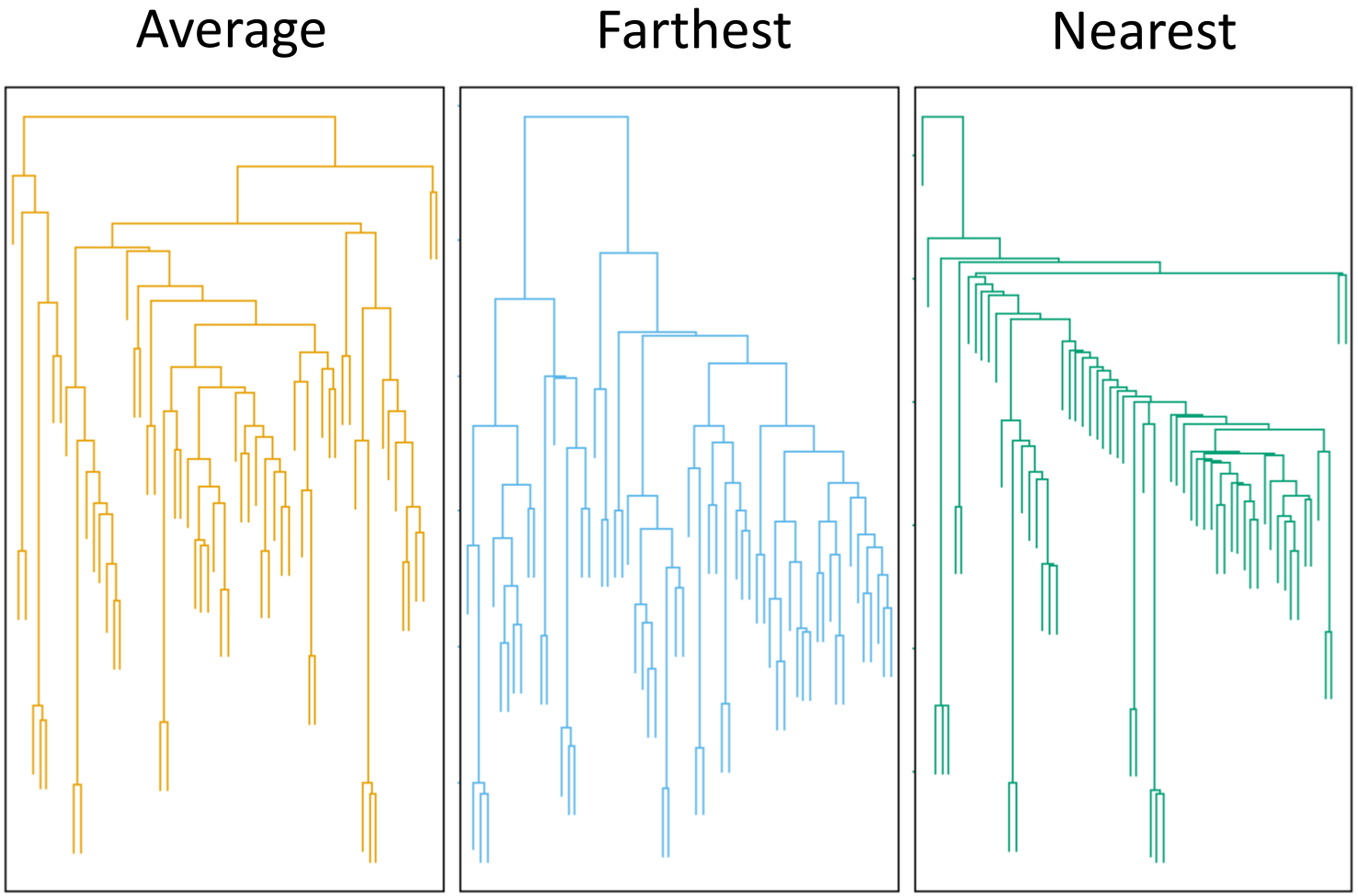
- Many more choices, each produces a different clustering...

Clustering Behavior



Mouse tumor data from [Hastie]

Clustering Behavior



Mouse tumor data from [Hastie]