



Unsupervised Learning: Clustering

Rishabh Iyer
University of Texas at Dallas

Based on the slides of Nick Rouzzi and Vibhav Gogate

Clustering

Clustering systems:

- Unsupervised learning

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

\uparrow \downarrow
Features

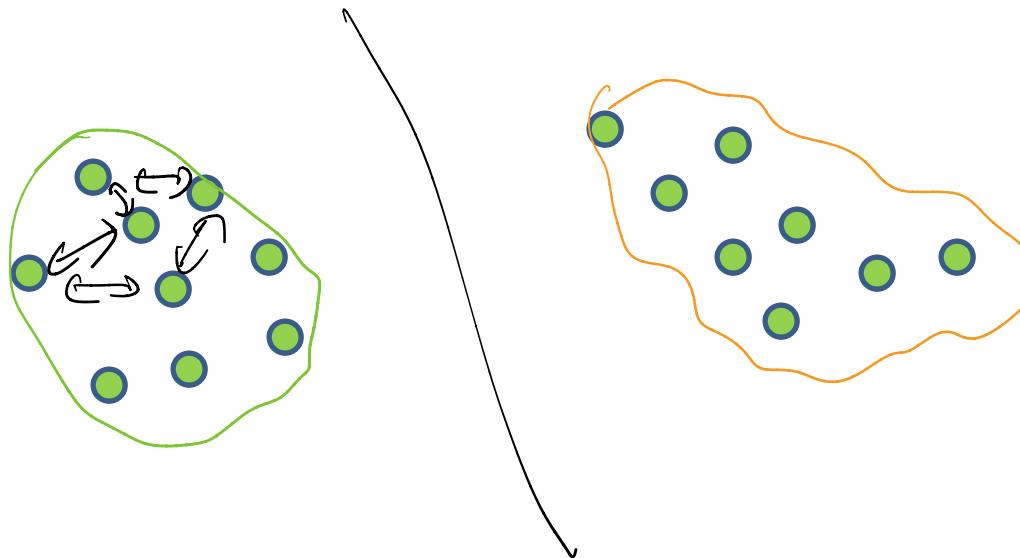
\mathcal{D} contains no labels

- 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



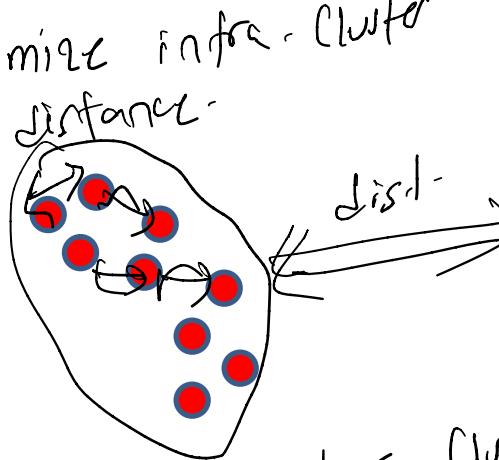
Clustering



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

③ Goals of clustering

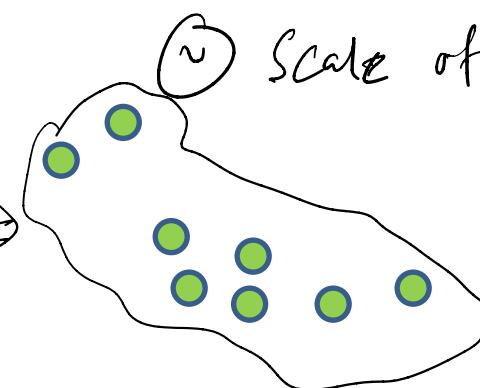
- minimize intra-cluster distance
- maximize inter-cluster distance



A hand-drawn diagram illustrating the goals of clustering. On the left, a cluster of red circles is enclosed within a large oval boundary. An arrow labeled "dist." points from the center of this cluster to the boundary, representing the intra-cluster distance. Another arrow points from the center to a group of blue circles outside the boundary, representing the inter-cluster distance. The text "- minimize intra-cluster distance" is written above the first arrow, and "- maximize inter-cluster distance" is written below the second arrow.

① Come up with a distance measure-

~ Scale of features
↳ Normalize features



A hand-drawn diagram showing two sets of data points. On the left, a cluster of green circles is enclosed within a large, irregular boundary. An arrow labeled "dist." points from the center of this cluster to the boundary, representing the intra-cluster distance. The text "~ Scale of features" is written above the cluster, and "↳ Normalize features" is written below it. On the right, the same set of green circles is shown, but they are enclosed within a much smaller, more compact boundary, indicating that the features have been scaled or normalized.

Clustering



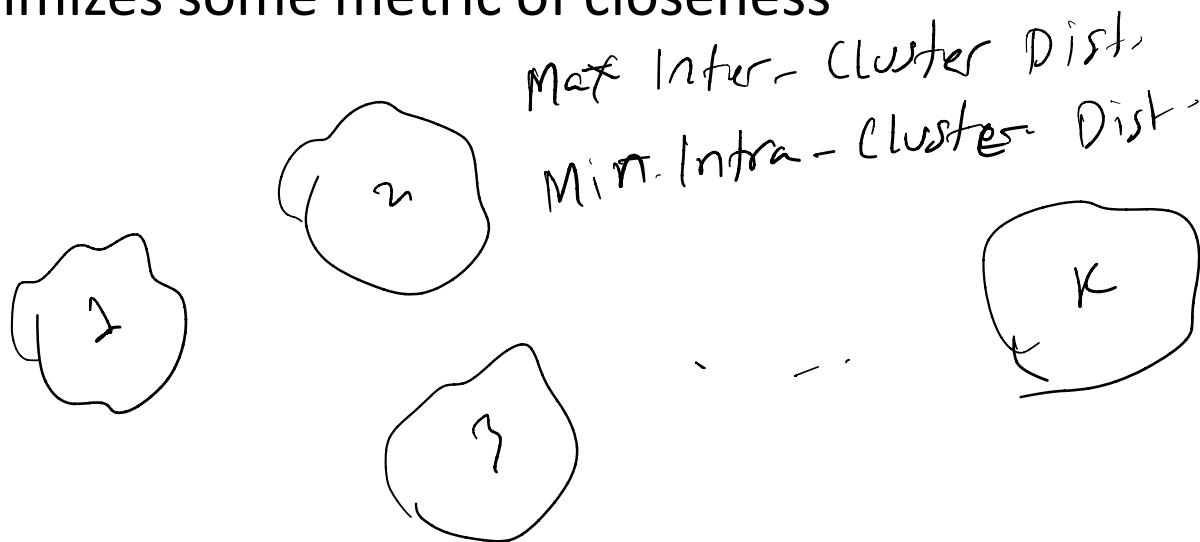
- Intuitive notion of clustering is a somewhat ill-defined problem
 - Identification of clusters depends on the scale at which we perceive the data



Clustering



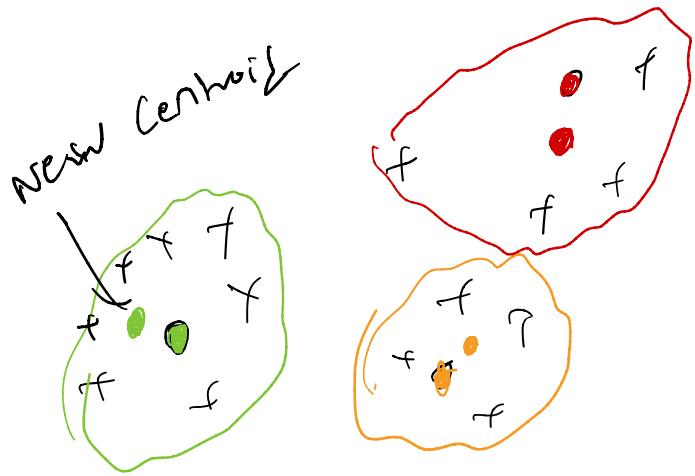
- Input: a collection of points $\underbrace{x^{(1)}, \dots, x^{(m)} \in \mathbb{R}^n}$, an integer k
- Output: A partitioning of the input points into k sets that minimizes some metric of closeness



k-means Clustering

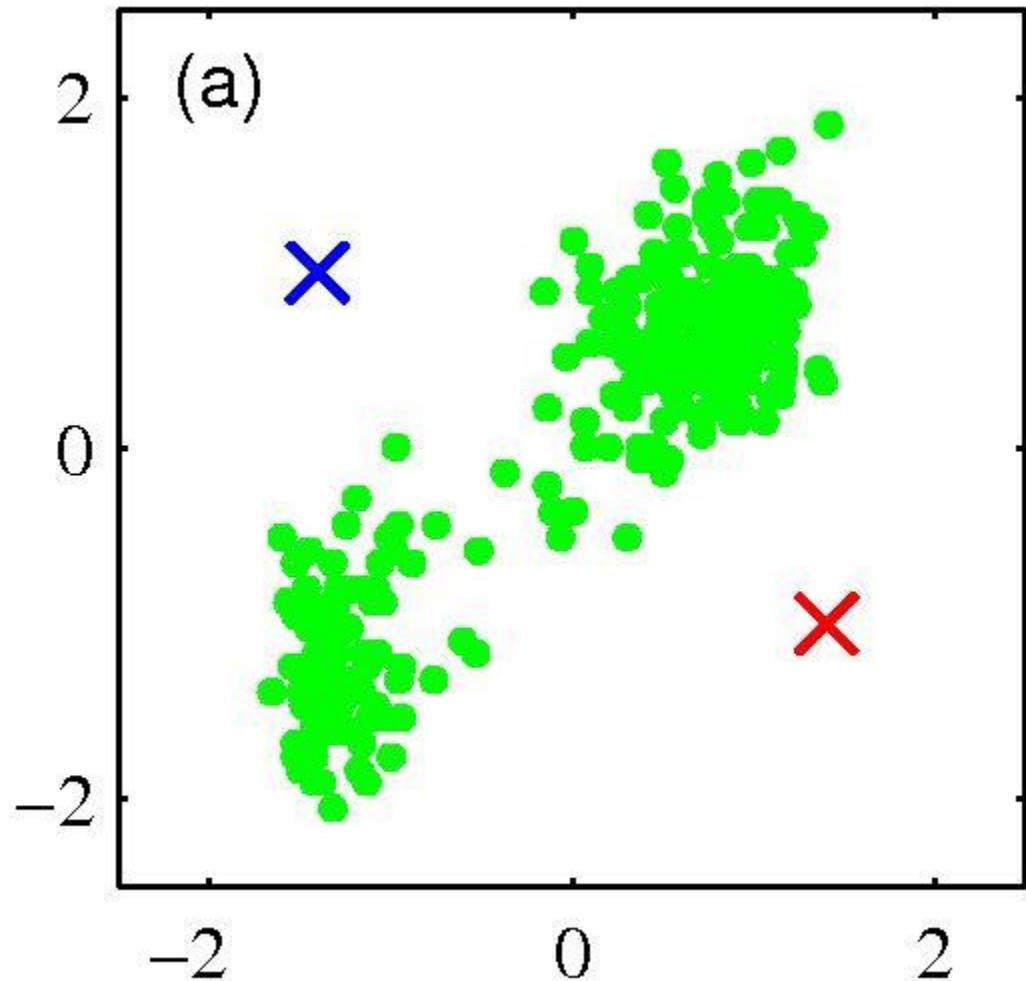


- Pick an initial set of k means (usually at random)
- Repeat until the clusters do not change:
 - Partition the data points, assigning each data point to a cluster based on the mean that is closest to it
 - Update the cluster means so that the i^{th} mean is equal to the average of all data points assigned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



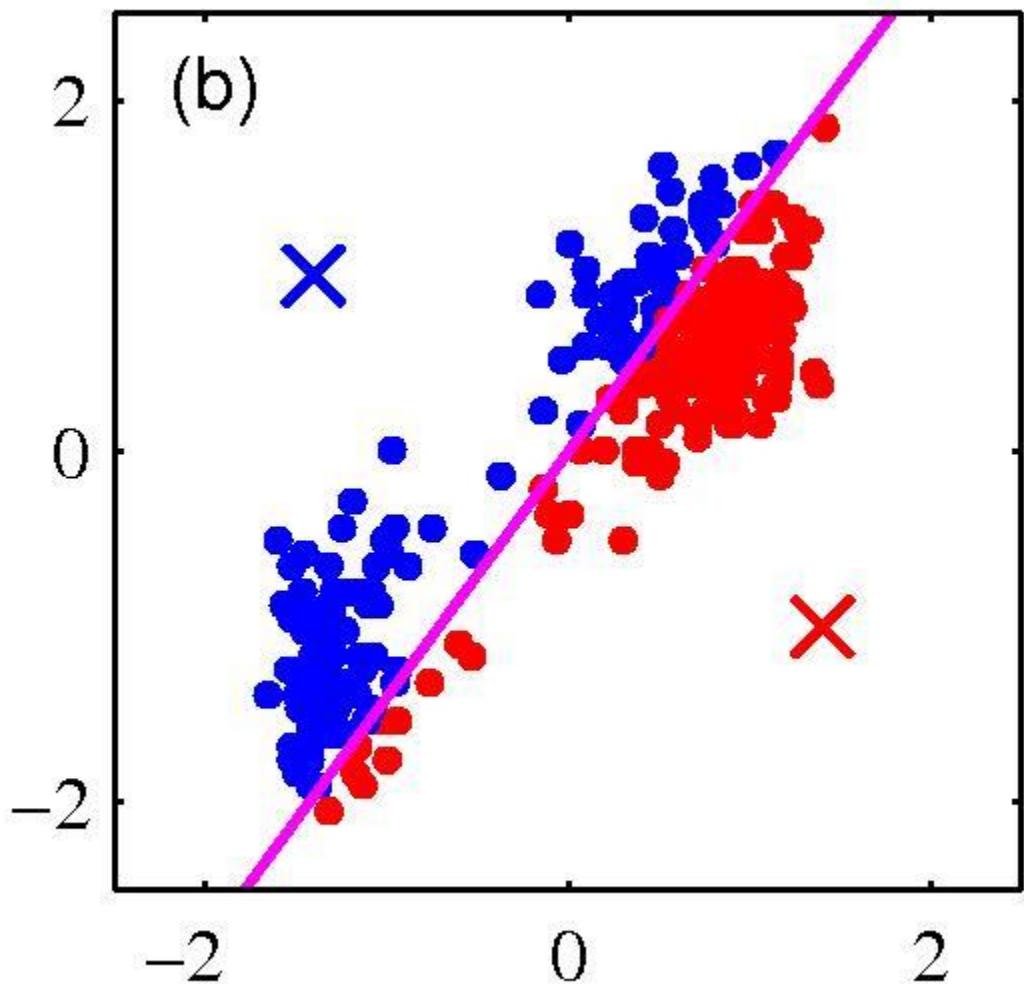
Partitioning
- Every point is covered by exactly one group

k -means clustering: Example



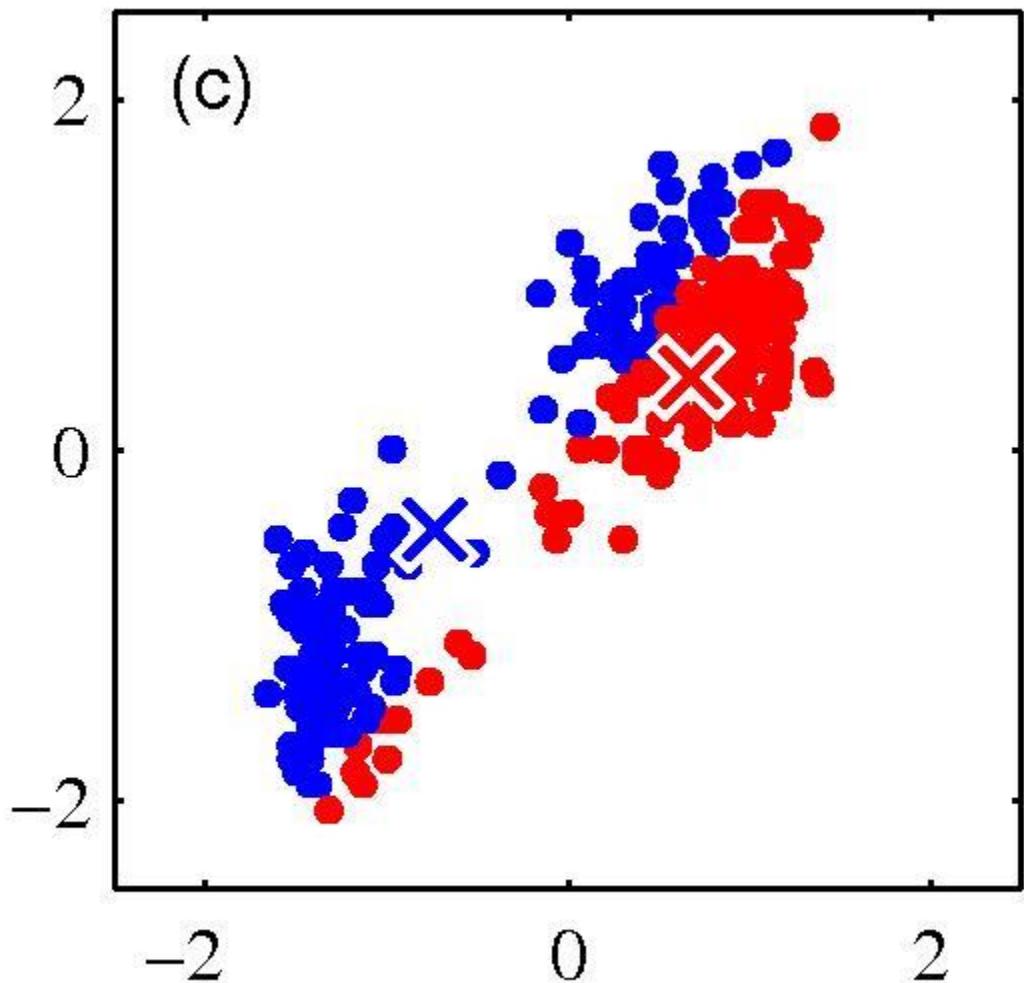
Pick k random points
as cluster centers
(means)

k -means clustering: Example



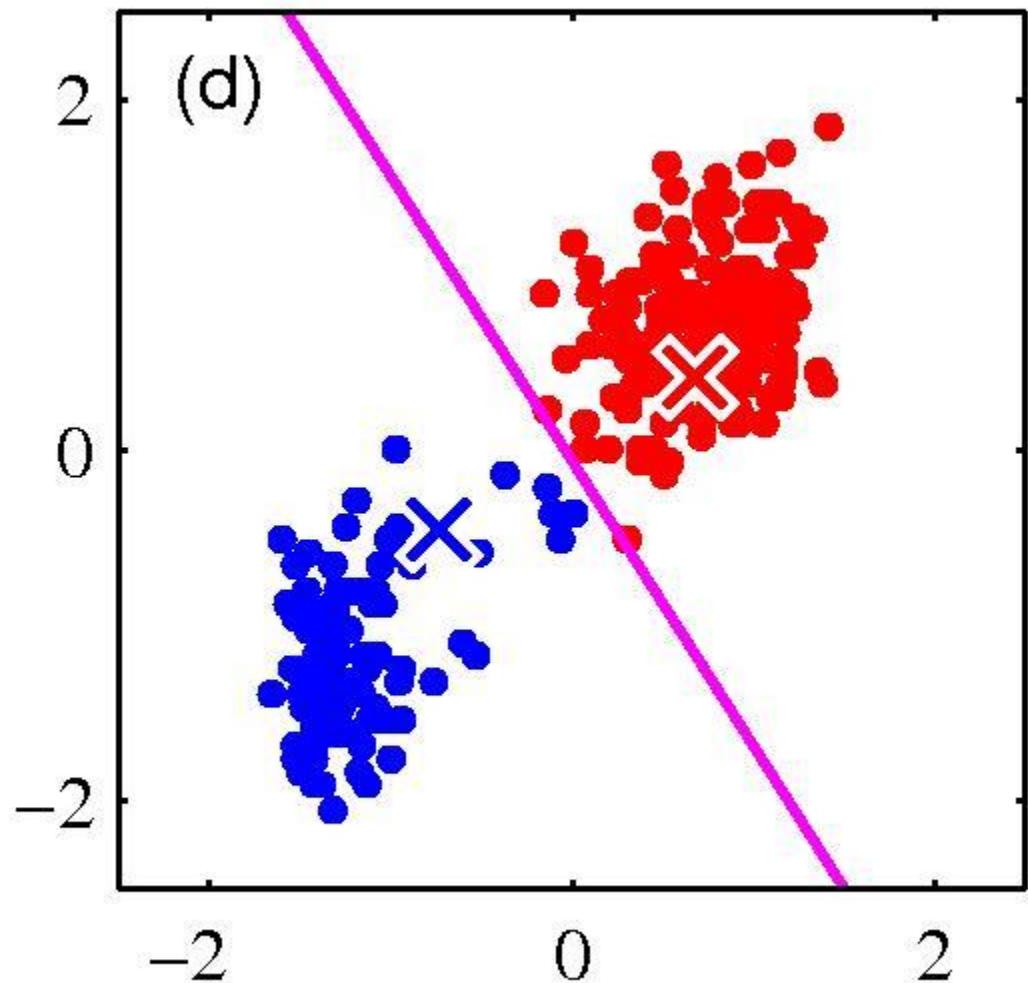
Iterative Step 1:
Assign data instances
to closest cluster
center

k -means clustering: Example



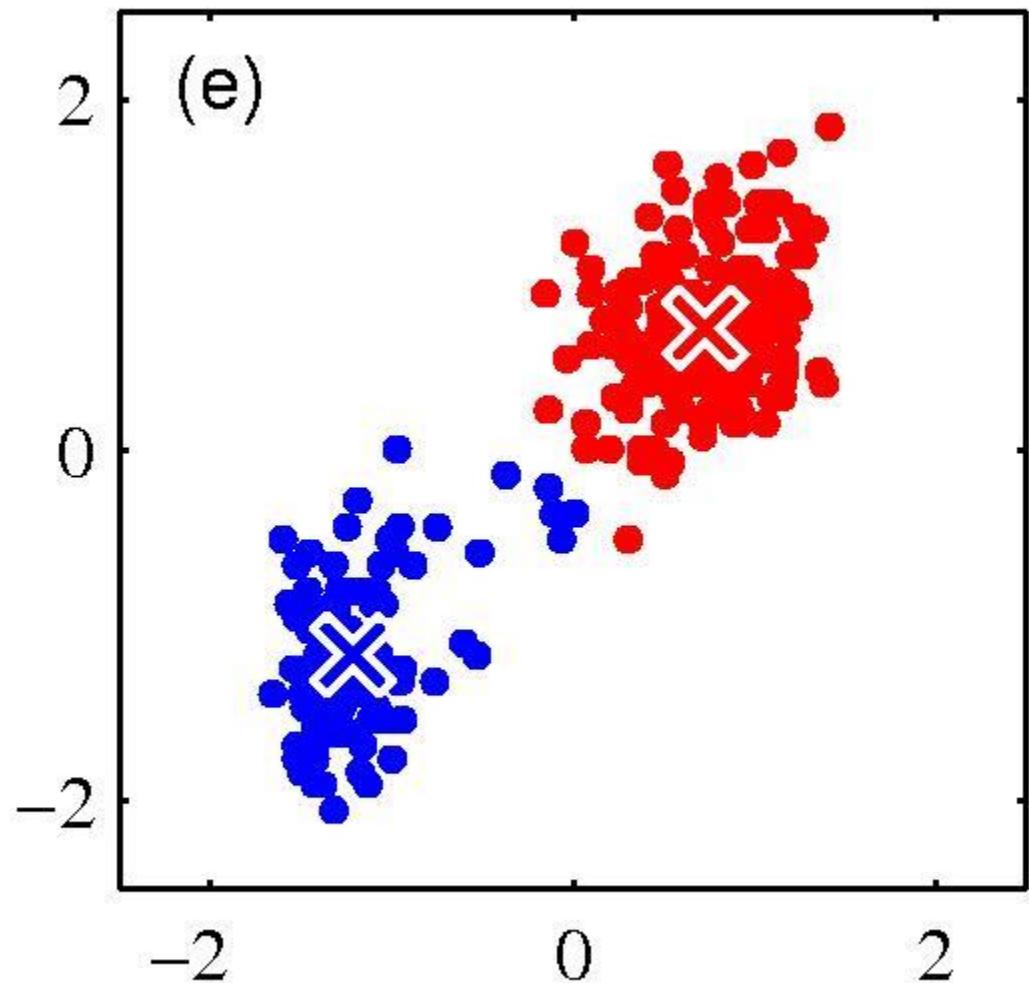
Iterative Step 2:
Change the cluster
center to the average of
the assigned points

k -means clustering: Example

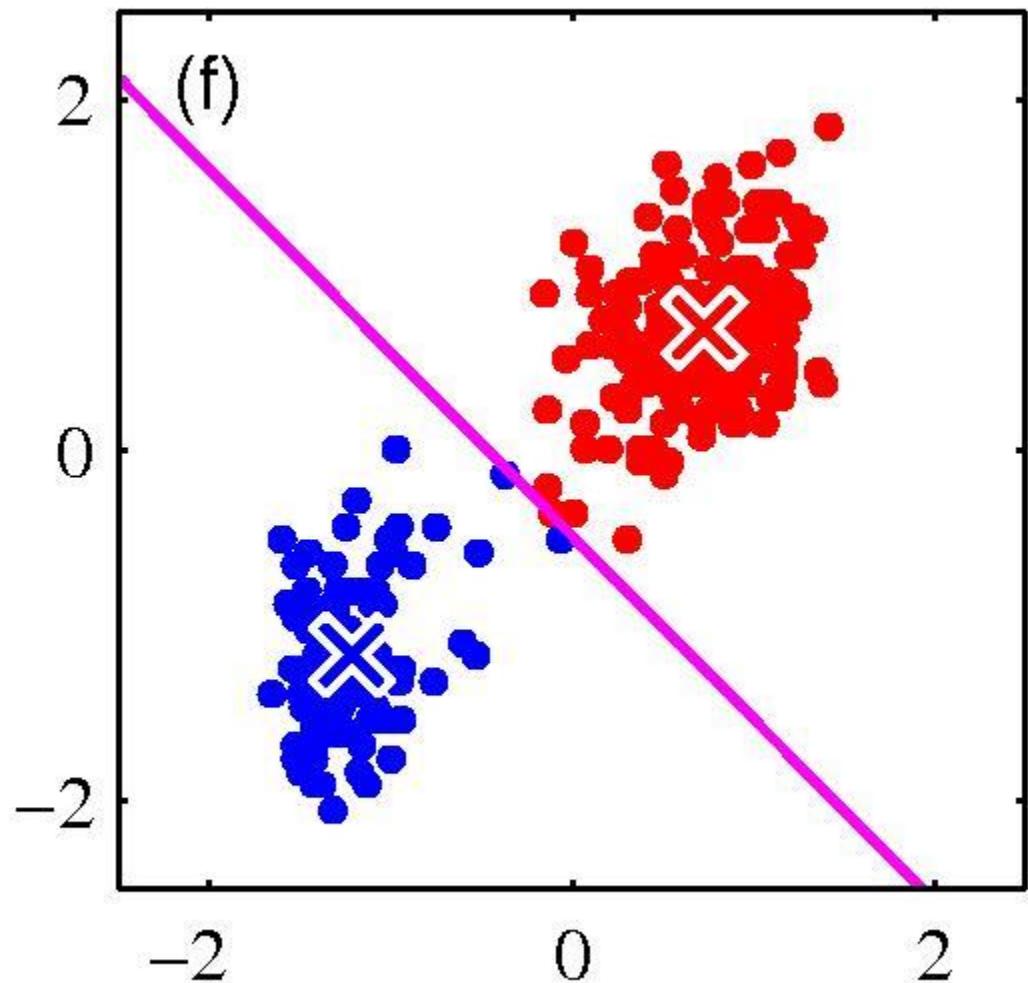


Repeat until
convergence

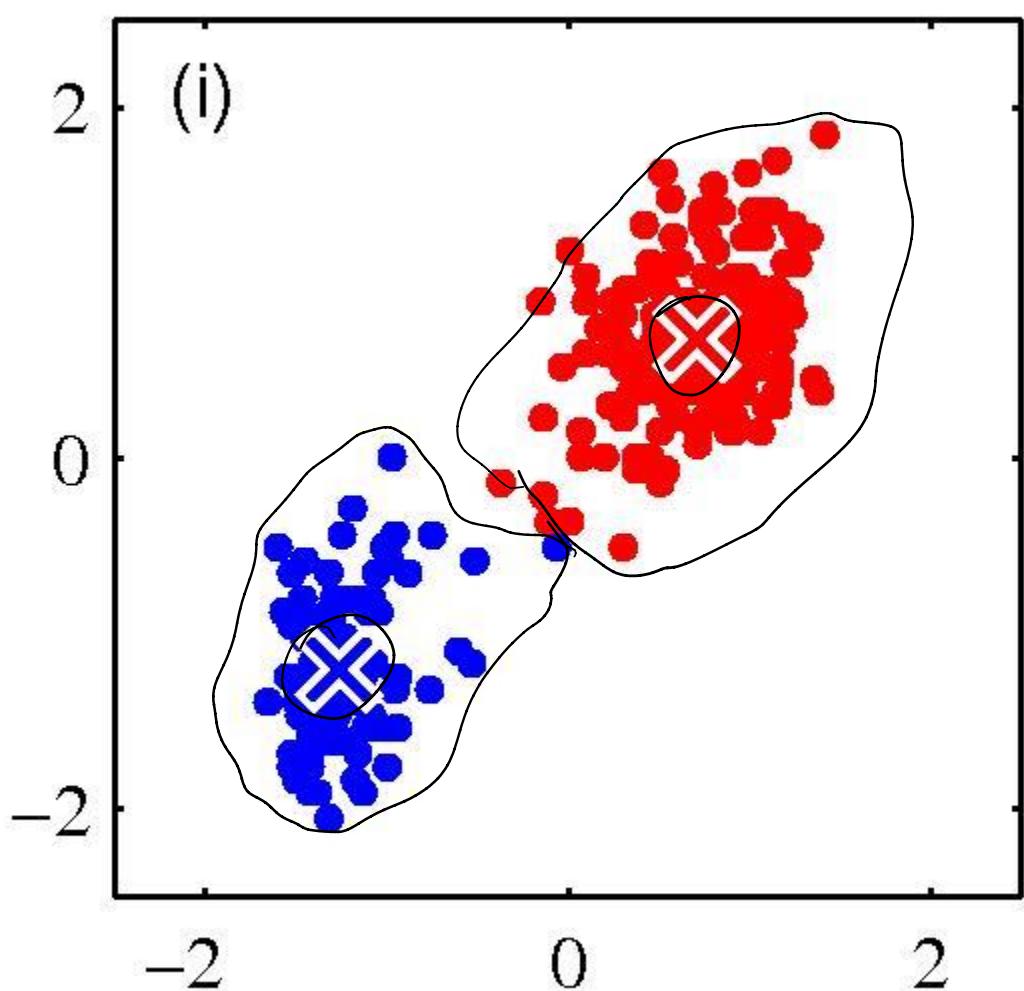
k -means clustering: Example



k -means clustering: Example



k -means clustering: Example



Stopping Criterion

$$\forall i = 1: K, \quad \|u_i^{t+1} - u_i^t\|_2 \leq \epsilon$$

$$c_i := n_j, j \in c_i$$

$$u_i = \frac{\sum_{j \in c_i} n_j}{|c_i|}$$



$$\arg \min_{x \in \mathbb{R}^n} \underbrace{\sum_{i=1}^K \|x_i - x\|_2^2}_{g(x)} = 0.$$

$$\begin{aligned} \nabla g(x) &= 2 \sum_{i=1}^K (x - x_i) \\ Kx &= \sum_{i=1}^K x_i \\ x &= \frac{1}{K} \sum_{i=1}^K x_i \end{aligned}$$

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

$$n = \underset{n}{\operatorname{argmin}} \sum_{i=1}^3 \|n - n_i\|^2$$

$$= \underset{n}{\operatorname{argmin}} \|n - n_1\|^2 + \|n - n_2\|^2 + \|n - n_3\|^2$$

$$= \underset{n}{\operatorname{argmin}} \left[3\|n\|^2 - 2(\langle n, n_1 \rangle + \langle n, n_2 \rangle + \langle n, n_3 \rangle) \right]$$

$$= \underset{n}{\operatorname{argmin}} \left[3\|n\|^2 - 2 \langle n, n_1 + n_2 + n_3 \rangle \right]$$

$$6n - 2(n_1 + n_2 + n_3)$$

$$\Rightarrow n = \frac{n_1 + n_2 + n_3}{3}$$

$C: \{n_1, n_2, \dots, n_l\}$

$$M = \frac{n_1 + n_2 + \dots + n_l}{l} = \frac{\sum_{i=1}^l n_i}{l}$$

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

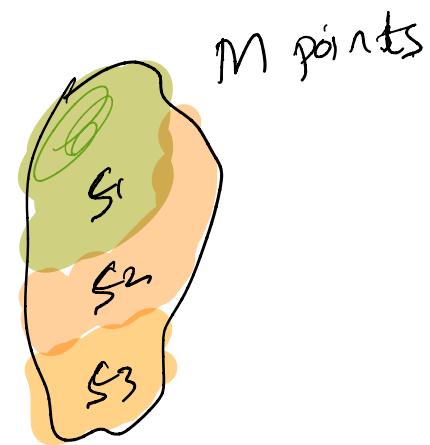


- Minimize the distance of each input point to the mean of the cluster/partition that contains it

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

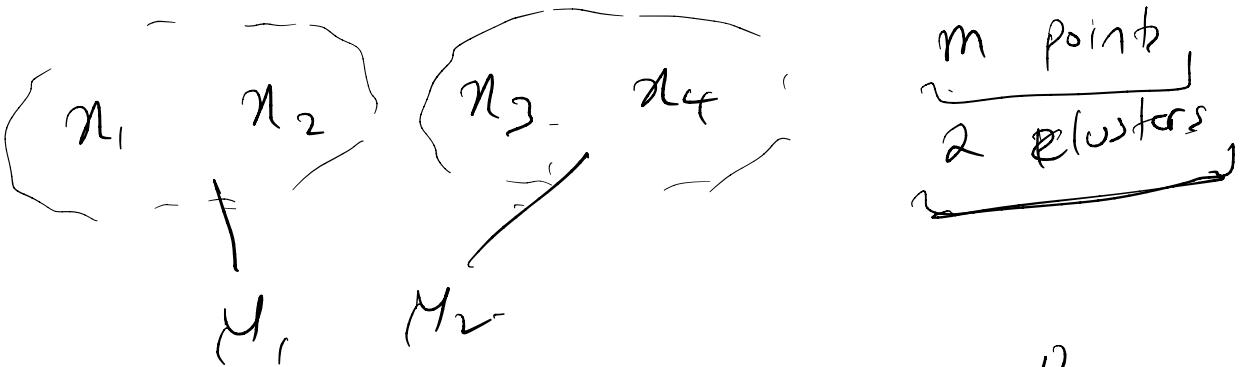
where

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



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

$$S_i \cap S_j = \emptyset$$



$$\|x_1 - M_1\|_2^2 + \|x_2 - M_1\|_2^2 + \|x_3 - M_2\|_2^2 + \|x_4 - M_2\|_2^2$$

$$M_1 = \frac{x_1 + x_2}{2}$$

$$M_2 = \frac{x_3 + x_4}{2}$$

Clustering

$$= 2^m - 2$$

$$= (m!) + \binom{m}{2} + \dots + \binom{m}{k}$$

k -means Clustering as Optimization



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

$$\min_{\substack{S_1, \dots, S_k \\ |S_i \cap S_j| = \emptyset \\ \cup S_i = [M]}} \sum_{i=1}^k \sum_{j \in S_i} \|x^{(j)} - \mu_i\|^2$$

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

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

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

$$f(S_1, \dots, S_k, M_1, \dots, M_k) = \sum_{i=1}^k \sum_{j \in S_i} \|x^{(j)} - M_i\|^2$$

k -Means as Optimization

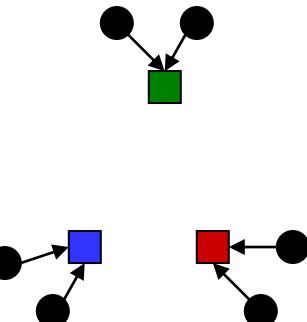
- Consider the k -means objective function

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

Data points
 \downarrow
 $\phi(x, S, \mu)$
 \uparrow
 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



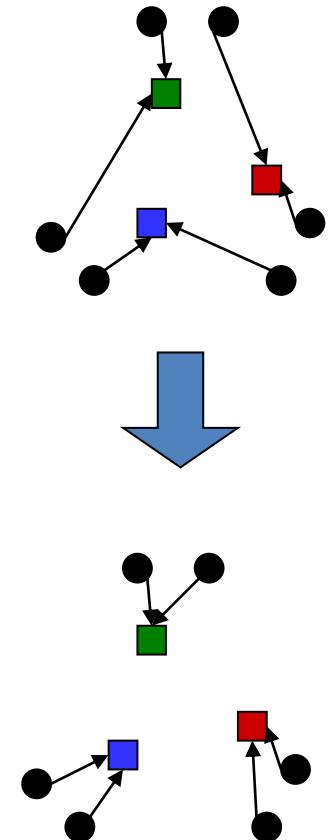
Phase I: Update Assignments

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

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

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

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



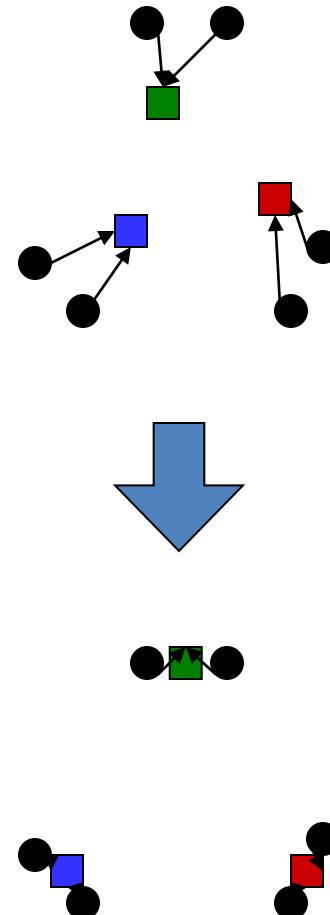
Phase II: Update Means

- Move each mean to the average of its assigned points

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

- Also can only decrease total distance...
 - Why?

$$\begin{aligned} \underline{\mu}_i &= \min_{\mu} \sum_{j \in S_i} \|\mu - x^{(j)}\|^2 \\ &= \sum_{j \in S_i} \frac{n}{|S_i|} \end{aligned}$$

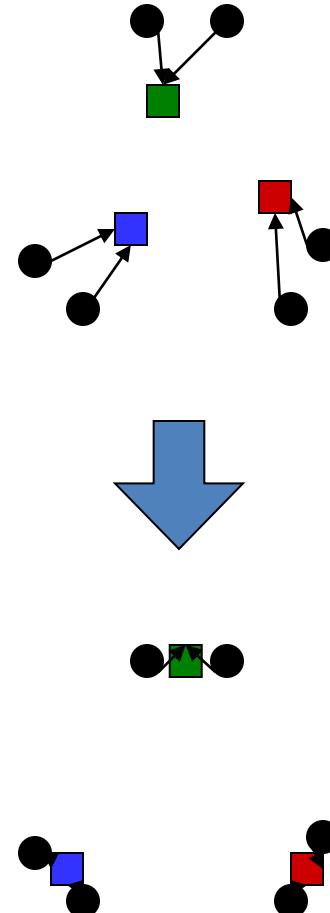


Phase II: Update Means

- 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



Alternating Minimization

K-means

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

Initialize x^0, y^0

$$x = s_1, \dots, s_n$$

$$y = M_1, \dots, M_n$$

Iterate: $t = 1 : T$

$$x^{t+1} = \min_x f(x, y^t) \quad \leftarrow \text{assignment}$$

$$y^{t+1} = \min_y f(x^{t+1}, y) \quad \leftarrow \text{means}$$

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

Theorem
 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

$$f(s_1, \dots, s_k, m_1, \dots, m_k) = \sum_{i=1}^k \sum_{j \in S_i} \|x^{(j)} - m_i\|^2$$

$$\min_{m_1, \dots, m_k} \sum_{i=1}^k \sum_{j \in S_i} \|x^{(j)} - m_i\|^2$$

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

↑
Computing / Centroids

$$\min_{s_1, \dots, s_k} \sum_{i=1}^k \sum_{j \in S_i} \|x^{(j)} - m_i\|^2 \quad \leftarrow \text{Assignment}$$

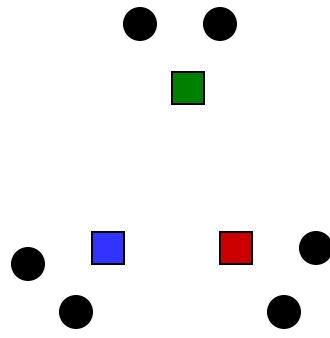
$s_i = \{ j : \|x_j - m_i\| \leq \|x_j - m_l\| \text{ for all other centroids } l \}$

Initialization

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

Initialization

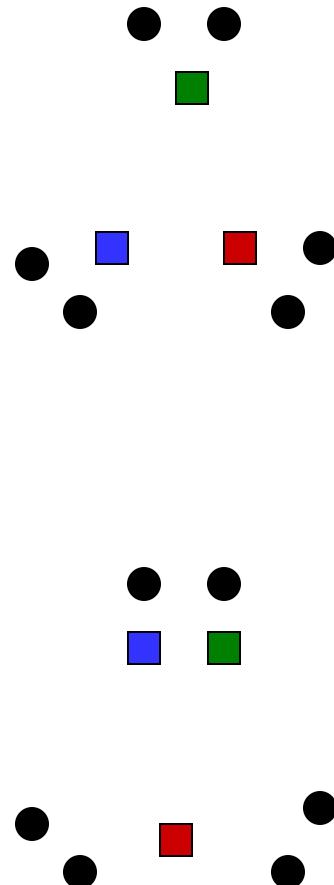
- K-means is sensitive to initialization
 - It does matter what you pick!
 - What can go wrong?



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)



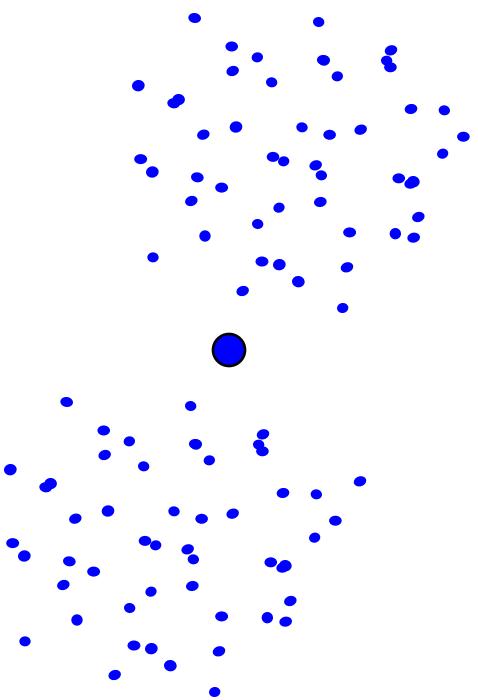
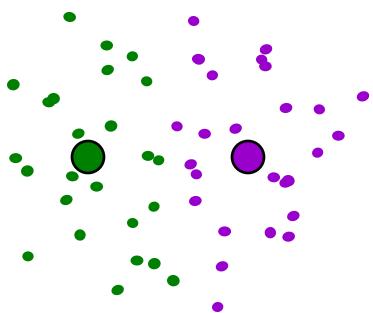
- Not clear how to figure out the "best" k in advance
- Want to choose k to pick out the interesting clusters, but not to overfit the data points
 - Large k doesn't necessarily pick out interesting clusters
 - Small k can result in large clusters than can be broken down further

Recent work

Local Optima \Rightarrow Bad Initialization.



- Assignment
- centroid



k -Means Summary



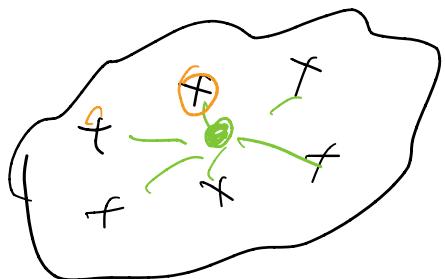
- Guaranteed to converge
 - But not to a global optimum
- Choice of k and initialization can greatly affect the outcome
- Runtime: $\underbrace{O(kMn)}$ per iteration (Bound)
- Popular because it is fast, though there are other clustering methods that may be more suitable depending on your data

$$\mathcal{E} \rightarrow O\left(\frac{kMn}{\epsilon}\right)$$

K-medoids 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.

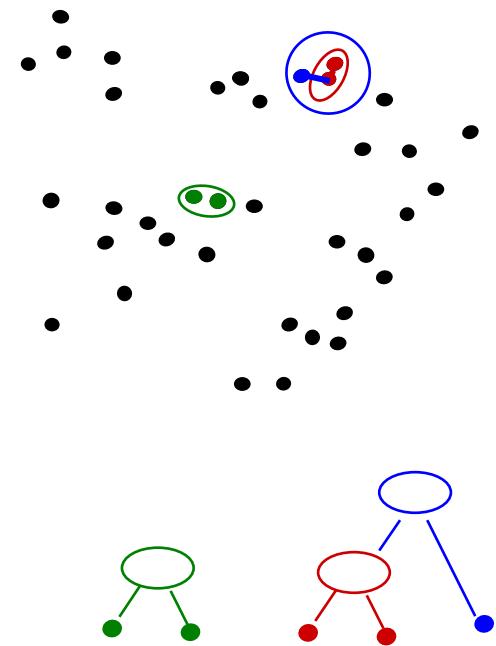
- ❑ Extensions of k-means: Other distance measures, e.g. the Bregman divergence

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

$$O(|S|^2)$$

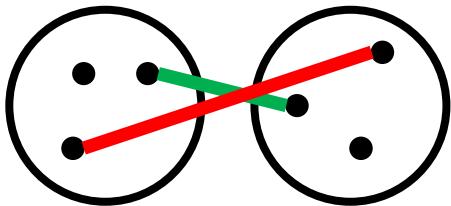
Hierarchical Clustering

- 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**

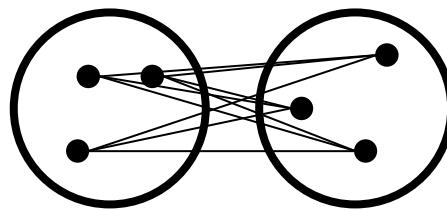


Agglomerative Clustering

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



Closest / farthest pair



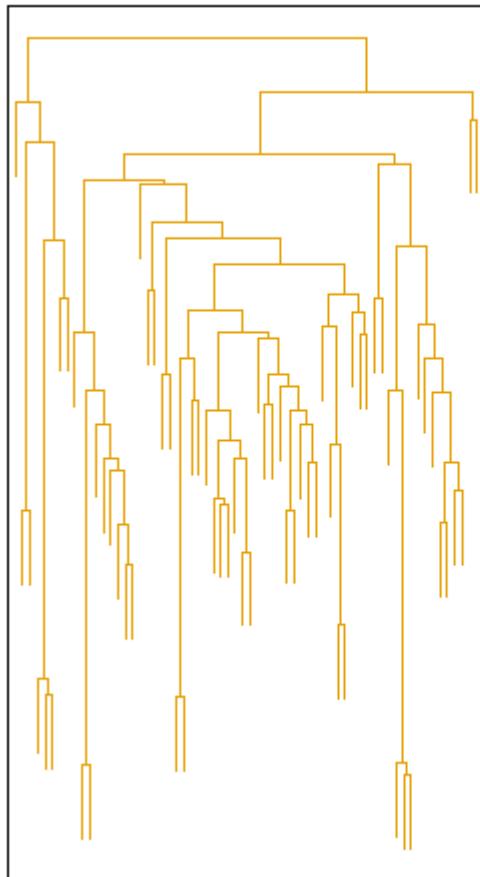
Average of all pairs

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

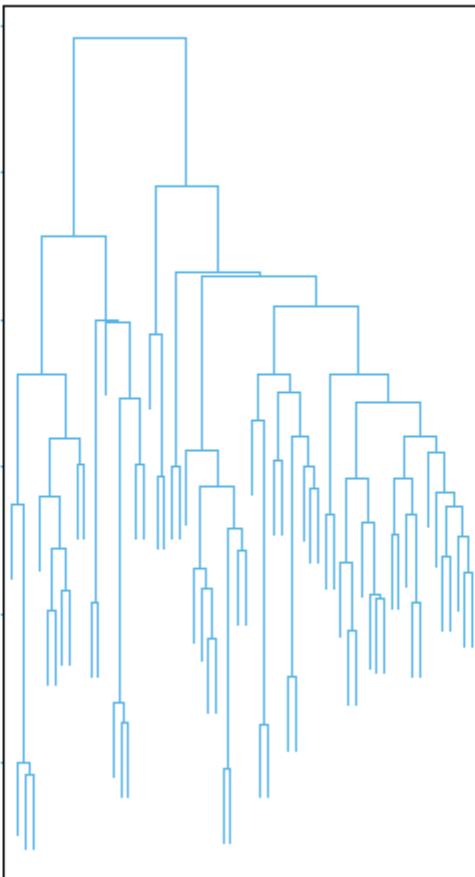
Clustering Behavior



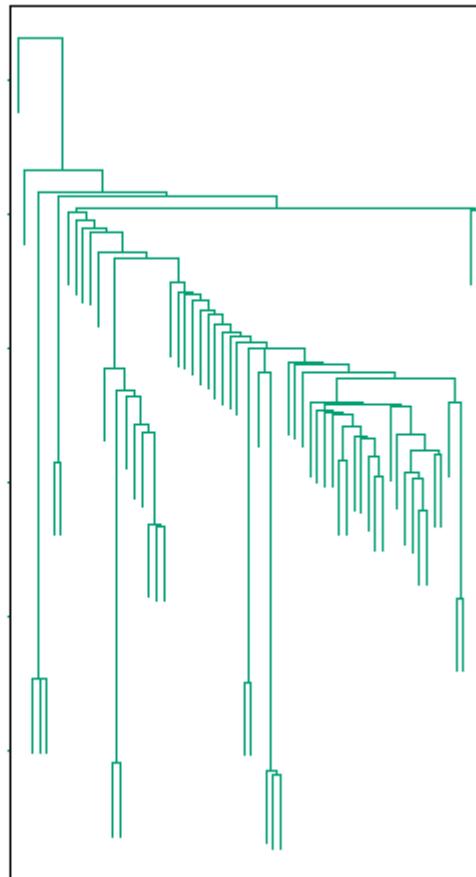
Average



Farthest



Nearest

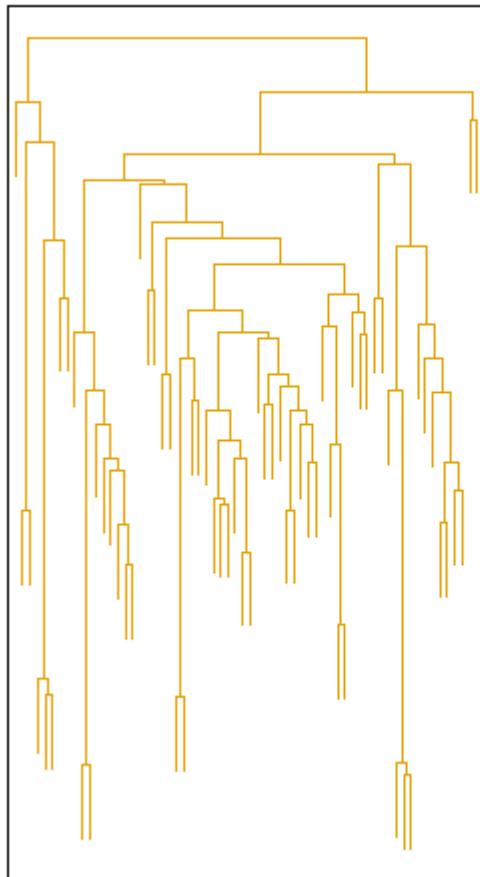


Mouse tumor data from [Hastie]

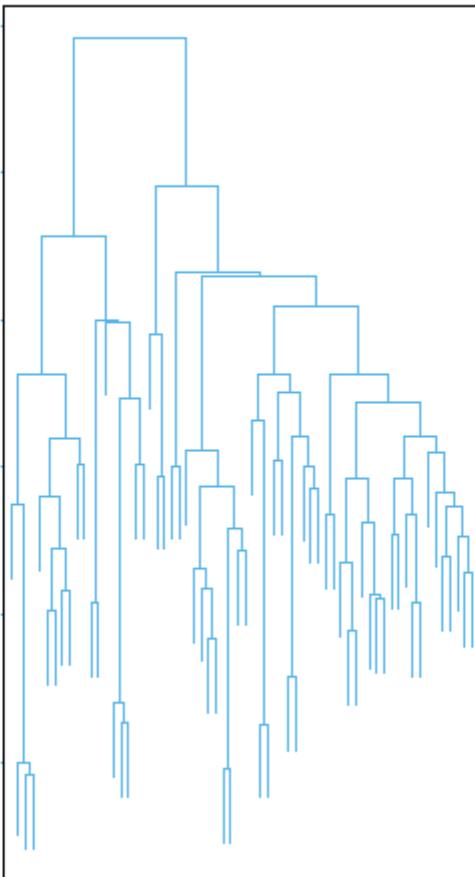
Clustering Behavior



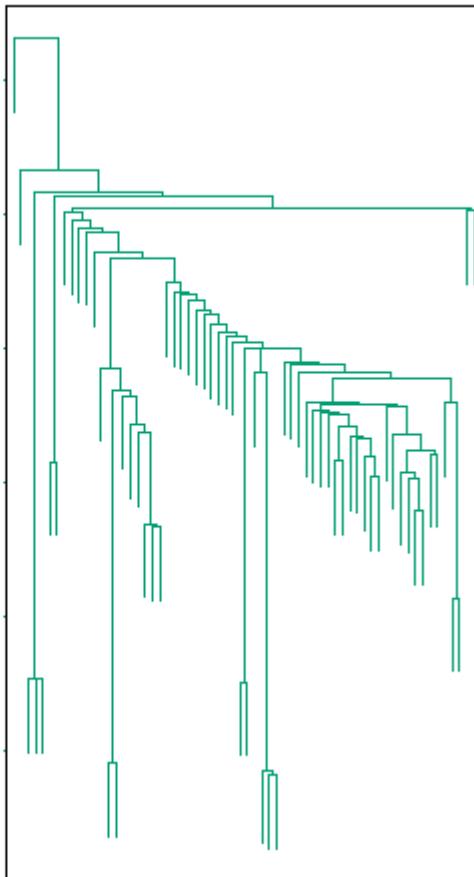
Average



Farthest



Nearest



Mouse tumor data from [Hastie]