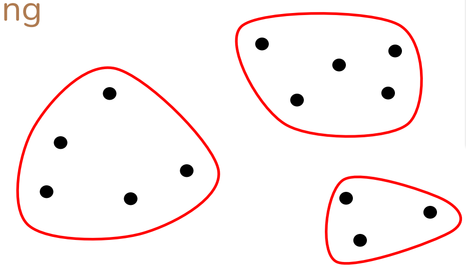
Clustering

Clustering

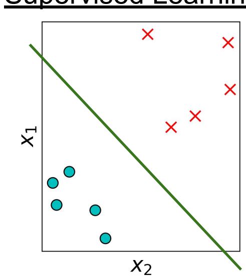


It can be useful to partition points into groups of similar points

Clustering is the process of finding groups of points, such that points in the same group are as similar to each other as possible and as dissimilar to points in other groups as possible

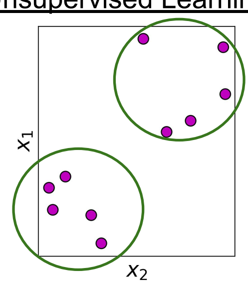
Supervised Learning vs. Unsupervised Learning

Supervised Learning



X is 10x2 array, y is 10x1 array

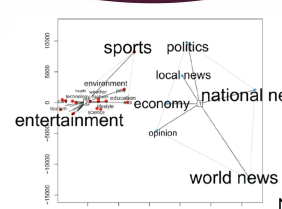
Unsupervised Learning



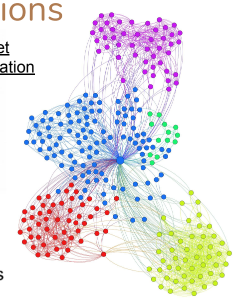
X is 10x2 array

Clustering Applications

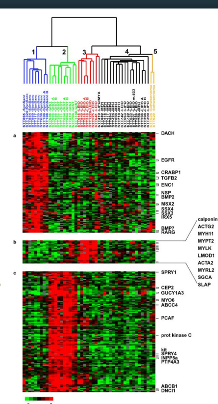
Market Segmentation



News Articles



Social Network Analysis



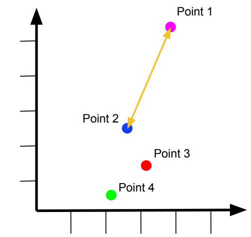
Gene Expression

Clustering Applications

- Feature quantization: group together many features into a few clusters
- Exploratory (data) science
- First pass before manually annotating data with labels

Distance Measure in 2D - Euclidean Distance

Point 1	3.8	5.4
Point 2	2.6	2.6
Point 3	3.1	1.5
Point 4	2.1	0.5



$$distance(\text{Point 1}, \text{Point 2}) = \sqrt[2]{(3.8 - 2.6)^2 + (5.4 - 2.6)^2}$$

$$distance(\text{Point } a, \text{Point } b) = \sqrt[2]{(a_1 - b_1)^2 + (a_2 - b_2)^2}$$

Distance Measure in Higher Dimensions

Point 1	3.8	5.4	4.7	5.0	...	4.2
Point 2	2.6	2.6	2.6	2.6	...	2.6
Point 3	3.1	1.5	2.2	1.9	...	2.7
Point 4	2.1	0.5	1.2	0.9	...	1.7

$$distance(\text{Point } a, \text{Point } b) = \sqrt[2]{\sum_{i=1}^{d} (a_i - b_i)^2}$$

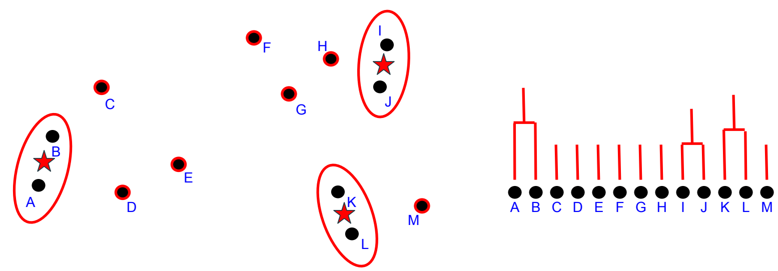
Clustering Algorithms

- Hierarchical (agglomerative) clustering
- *k*-means
- Gaussian mixture models

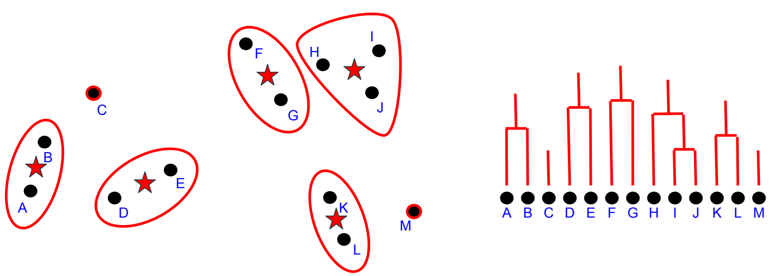
Hierarchical Clustering Algorithm

- Assign each point to its own cluster
- Repeat until the desired number of clusters is reached:
  - ➢ Merge together the two closest clusters

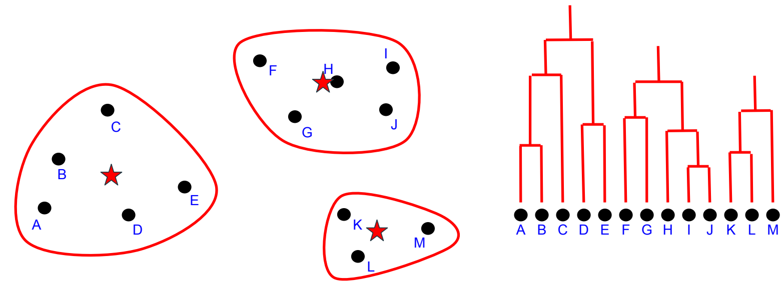
Hierarchical Clustering Example



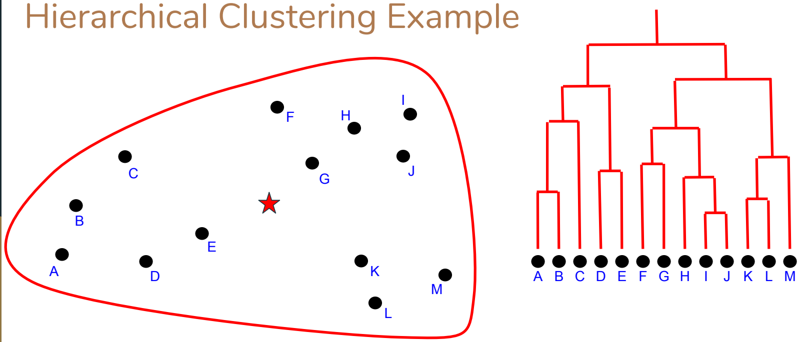
Hierarchical Clustering Example



Hierarchical Clustering Example



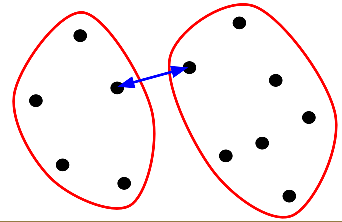
Hierarchical Clustering Example



Distance Between Clusters

- Single linkage

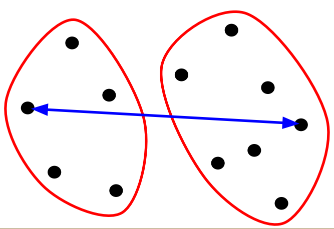
  The distance between two clusters is the distance between the closest pair of points (one from each cluster) in the clusters



Distance Between Clusters

- Complete linkage

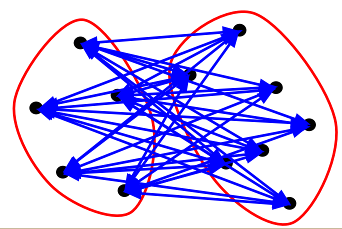
  The distance between two clusters is the distance between the farthest pair of points (one from each cluster) in the clusters



Distance Between Clusters

- Average linkage

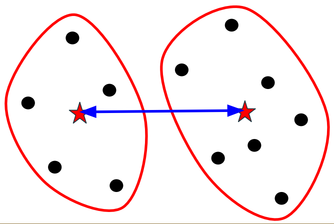
  The distance between two clusters is the average distance between all pairs of points (one from each cluster) in the clusters



Distance Between Clusters

- Centroid linkage

  The distance between two clusters is the distance between the centroids of each cluster



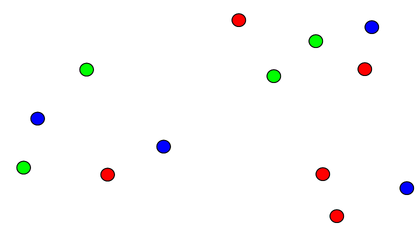
Clustering Algorithms

- Hierarchical (agglomerative) clustering
- *k*-means
- Gaussian mixture models

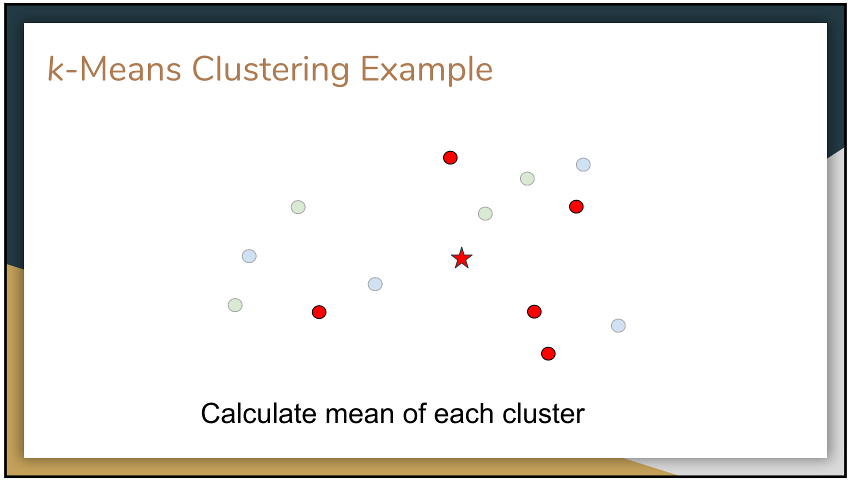
k-Means Clustering Algorithm

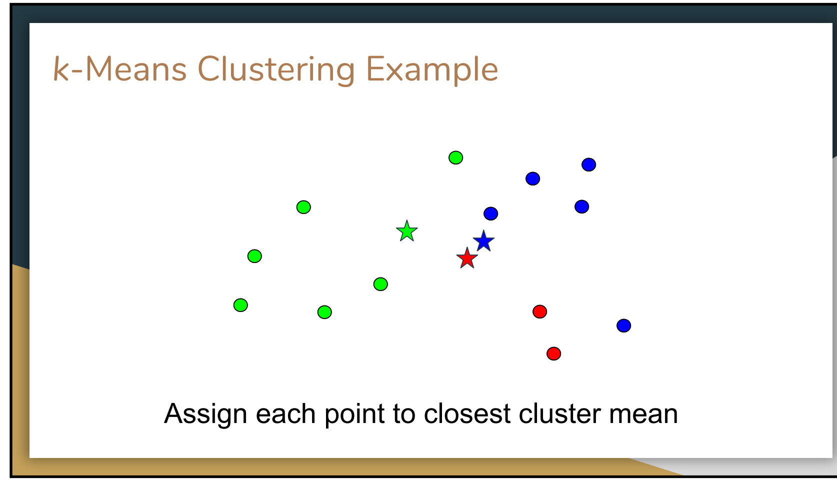
- Randomly assign each point to one of *k* clusters
- Repeat until convergence:
  - ➢ Calculate *mean* of each of the *k* clusters
  - ➢ Assign each point to the cluster with the closest *mean*

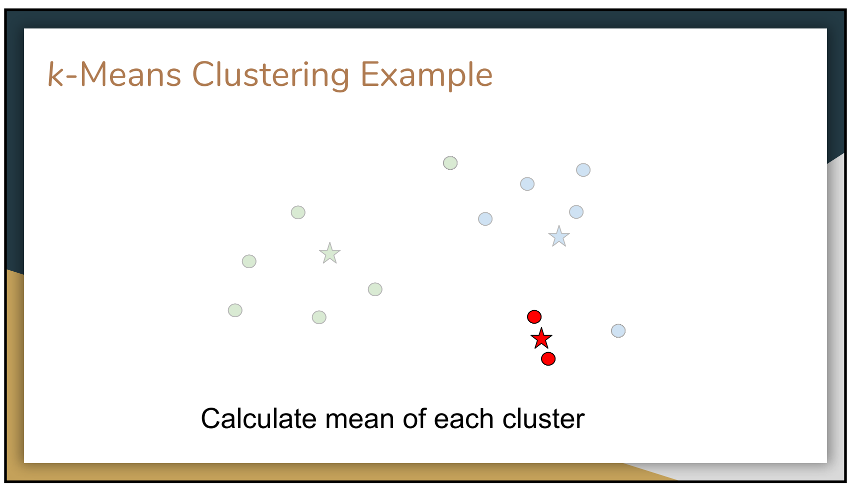
k-Means Clustering Example

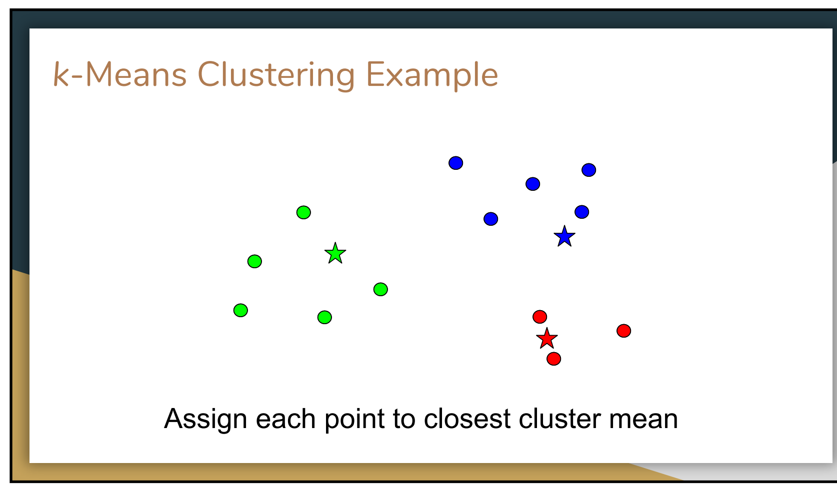


Randomly assign each point to one of *k* clusters

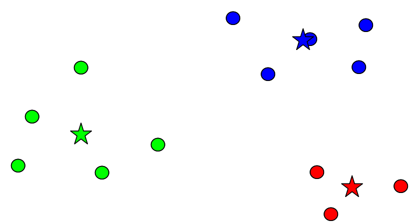

k-Means Clustering Example
Calculate mean of each cluster


k-Means Clustering Example
Assign each point to closest cluster mean


k-Means Clustering Example
Calculate mean of each cluster


k-Means Clustering Example
Assign each point to closest cluster mean

k-Means Clustering Example



Convergence

Clustering Problem

- For a given number of clusters, k, we measure a clustering's quality as the sum of the distances between each point and the mean of the point's cluster

$$\sum_{i=1}^{k} \sum_{\mathbf{x} \in i^{\text{th}}\text{cluster}} (\mathbf{x} - \mu_i)^2$$

- *Clustering Problem:* Partition n data points into k clusters such that the total distance from each point to its cluster mean is minimized
- Clustering is an NP-complete problem

k-Means Heuristic

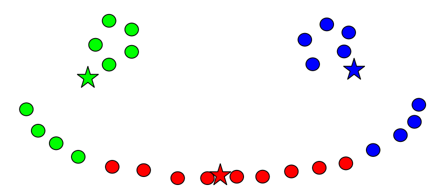
- Find a set of k means $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_k$ such that:

$$\underset{\mu_1, \mu_2, \ldots, \mu_k}{\operatorname{argmin}} \sum_{i=1}^{k} \sum_{\mathbf{x} \in i^{\text{th}}\text{cluster}} (\mathbf{x} - \mu_i)^2$$

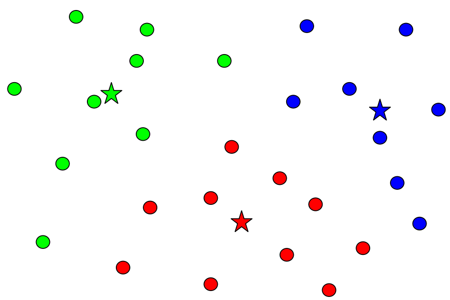
- *k*-means (Lloyd's) algorithm is one way to minimize this objective function
- Walks "downhill" of this function with each iteration
- Objective function is not convex: has local minima
- Algorithm finds local minimum depending on starting point

Thus, repeat algorithm with different random starting points!

Does *k*-Means Always Work?



Does *k*-Means Always Work?



Clustering Algorithms

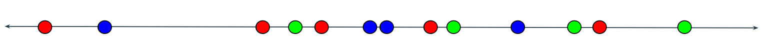
- Hierarchical (agglomerative) clustering
- *k*-means
- Gaussian mixture models

Model-Based Clustering

- Randomly assign each point to one of *k* clusters
- Repeat until convergence:
  - Calculate *model* of each of the *k* clusters
  - Assign each point to the cluster with the closest *model*

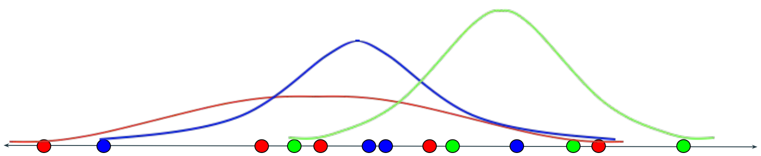
Example Clustering in 1-Dimension

Randomly assign each point to one of *k* clusters



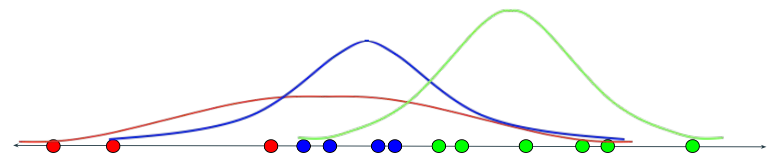
Example Clustering in 1-Dimension

Calculate model for each of the *k* clusters



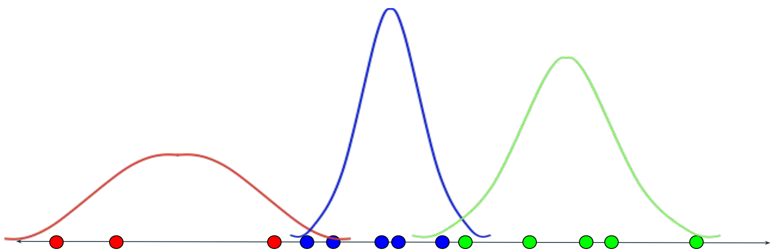
Example Clustering in 1-Dimension

Assign each point to the most likely model



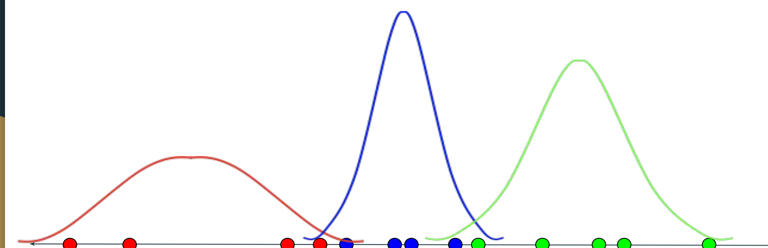
Example Clustering in 1-Dimension

Calculate model for each of the *k* clusters



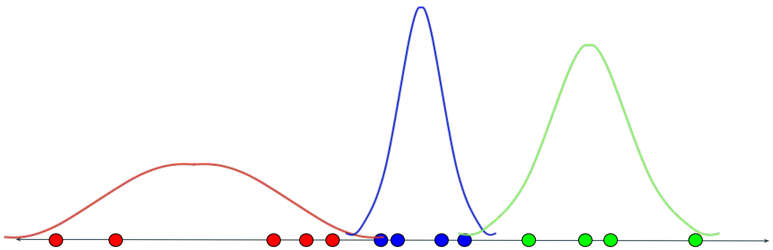
Example Clustering in 1-Dimension

Assign each point to the most likely model

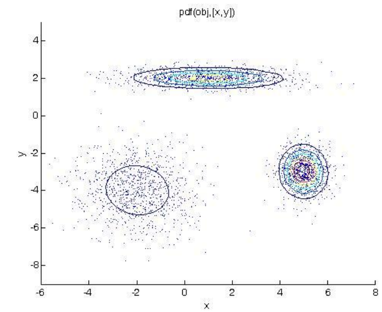


Example Clustering in 1-Dimension

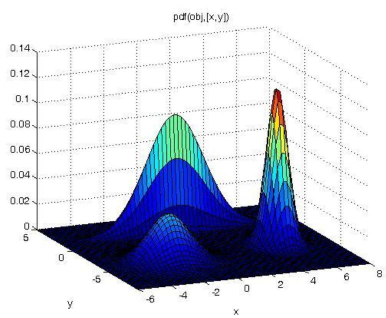
Calculate model for each of the *k* clusters



Example Clustering in 2-Dimensions



Example Clustering in 2-Dimensions

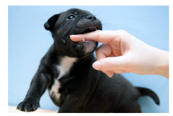


Clustering Data Examples vs. Features

	Feature 1	Feature 2	Feature 3	Feature 4	Feature 5		Feature *d*
Example 1	0.6	4.4	1.3	1.0	3.1	...	2.9
Example 2	1.5	2.6	5.2	0.8	2.7	...	1.6
Example 3	0.7	3.7	2.4	1.9	1.5	...	4.0
Example 4	0.3	3.0	0.2	1.3	4.9	...	0.9
…	…	…	…	…	…	…	…
Example *n*	0.5	3.4	1.8	0.5	2.2	...	3.1

Assessing Clustering

Challenge: clustering is subjective



Assessing Clustering

- Evaluate against ground truth labels
  - ➢ Trouble is, we normally don't have ground truth labels. If we did, we could have used *supervised* classification.

$$\mathrm{FMI} = \frac{\mathrm{TP}}{\sqrt{(\mathrm{TP} + \mathrm{FP})(\mathrm{TP} + \mathrm{FN})}}$$

- If we are clustering features, has it helped our classification task?
- High intra-class similarity, low inter-class similarity
- Human evaluation