Computer Science

CSC380: Principles of Data Science

Clustering

Xinchen Yu

Announcements

- Fill out SCS (https://scsonline.oia.arizona.edu/) – if 80% responses, will add 5 points to the homework with lowest grade.

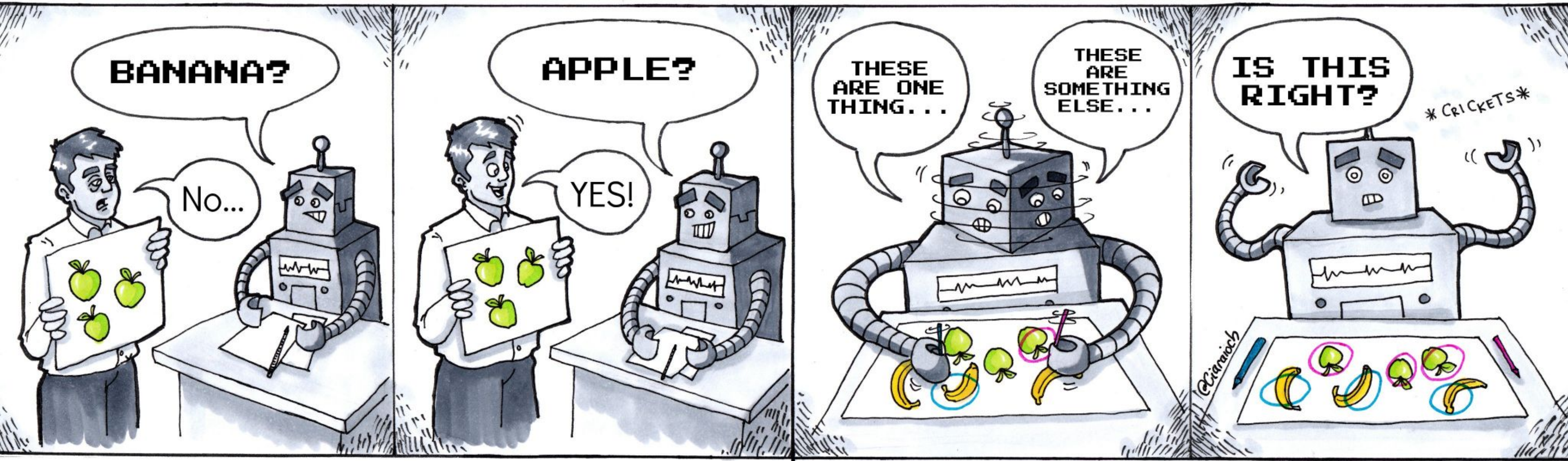
- No lecture next Tuesday, Apr 30
  - You can prepare final exam or work on practice problems in groups and I will do Q&A in person
  - Meinel Optical Sci, Rm 410 (same room)

Announcements

- Final exam
  - Time: Wednesday May 8, 3:30 - 5:30pm
  - Location: Meinel Optical Sci, Rm 410 (same room)
  - What you can bring:
    - one letter size cheat sheet, you can use double sides
    - calculator (not necessary)

Announcements

- ~20 questions and 50% questions will be before midterm.
- Practice questions has been out, keys will be out next week
- No coding questions
- How to prepare
  - **Slides**
  - Practice problems (helpful but do not only rely on it!)
  - HW questions before midterm



Supervised Learning

Unsupervised Learning

Task 1 : Group These Set of Document into 3 Groups based on meaning

Doc1 : Health , Medicine, Doctor

Doc 2 : Machine Learning, Computer

Doc 3 : Environment, Planet

Doc 4 : Pollution, Climate Crisis

Doc 5 : Covid, Health , Doctor

Task 1 : Group These Set of Document into 3 Groups based on meaning

Doc1 : Health , Medicine, Doctor

Doc 2 : Machine Learning, Computer

Doc 3 : Environment, Planet

Doc 4 : Pollution, Climate Crisis

Doc 5 : Covid, Health , Doctor

Task 1 : Group These Set of Document into 3 Groups based on meaning

Doc1 : Health , Medicine, Doctor

Doc 5 : Covid, Health , Doctor

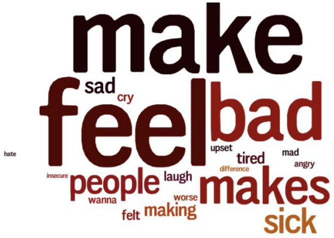
Doc 3 : Environment, Planet

Doc 4 : Pollution, Climate Crisis

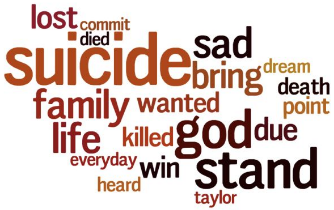
Doc 2 : Machine Learning, Computer

Task 2: Topic modeling

- Provides a summary of a corpus.
- Collected n tweets containing the keyword “bullying”, “bullied”, etc.
- Extracts k topics: each topic is a list of words with importance weights.
  - A set of words that co-occurs frequently throughout.



“feelings”

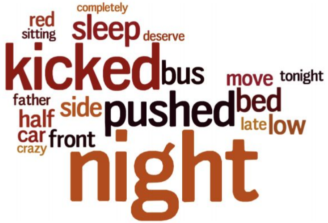


“suicide”

“family”

“school”

“verbal bullying”



“physical bullying”

Learning from bullying traces in social media.
Jun-Ming Xu, Kwang-Sung Jun, Xiaojin Zhu, and Amy Bellmore.
In the Conference of North American Chapter of the Association for Computational Linguistics: Human Language Technologies (**NAACL HLT**), 2012. [pdf]

What is unsupervised learning?

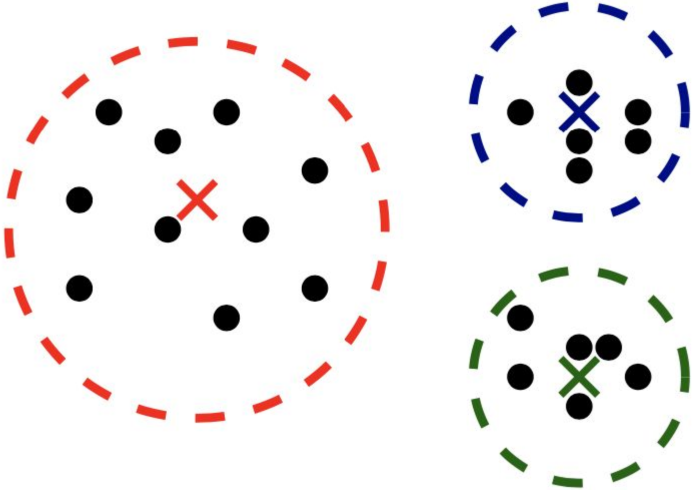
- Learning with unlabeled data
- What can we expect to learn?
  - **Clustering**: obtain partition of the data that are well-separated.
    - a preliminary classification without predefined class labels.
  - **Components**: extract common components
    - e.g., topic modeling given a set of articles: each article talks about a few topics => extract the topics that appear frequently.
- How can we use?
  - As a summary of the data
    - **Exploratory data analysis**: what are the **patterns** even without labels?
  - As a 'preprocessing techniques'
    - e.g., extract useful **features** using soft clustering assignments

Clustering

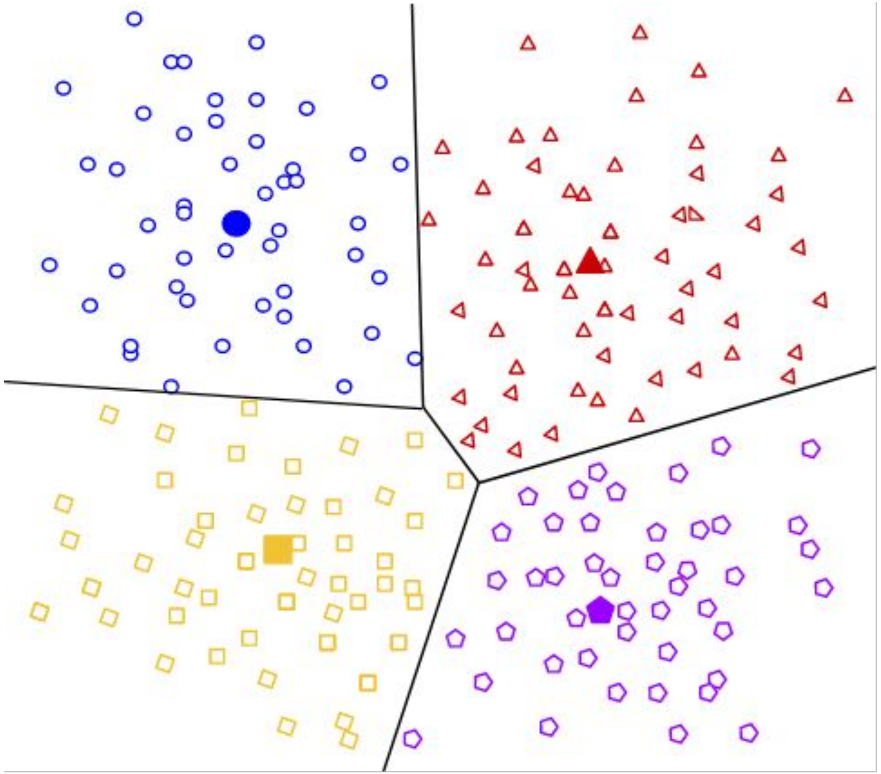
- Input: k: the number of clusters (hyperparameter)

$$S = \{x_1, \ldots, x_n\}$$

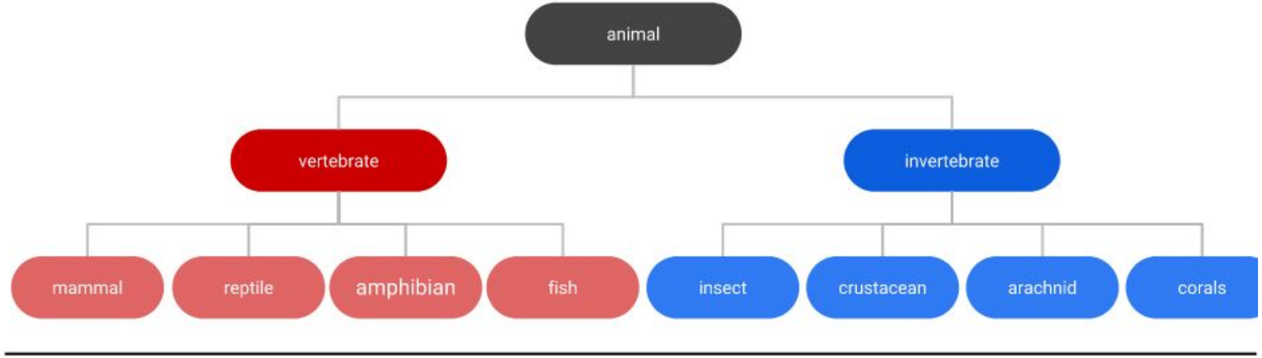
- Output
  - partition $\{G_i\}_{i=1}^{k}$ s.t. $S = \cup_i G_i$ (disjoint union).
  - often, we also obtain 'centroids'

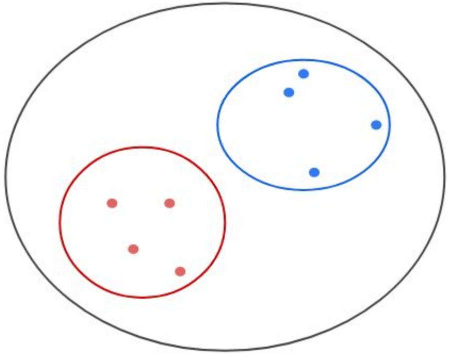


Centroid-based Clustering



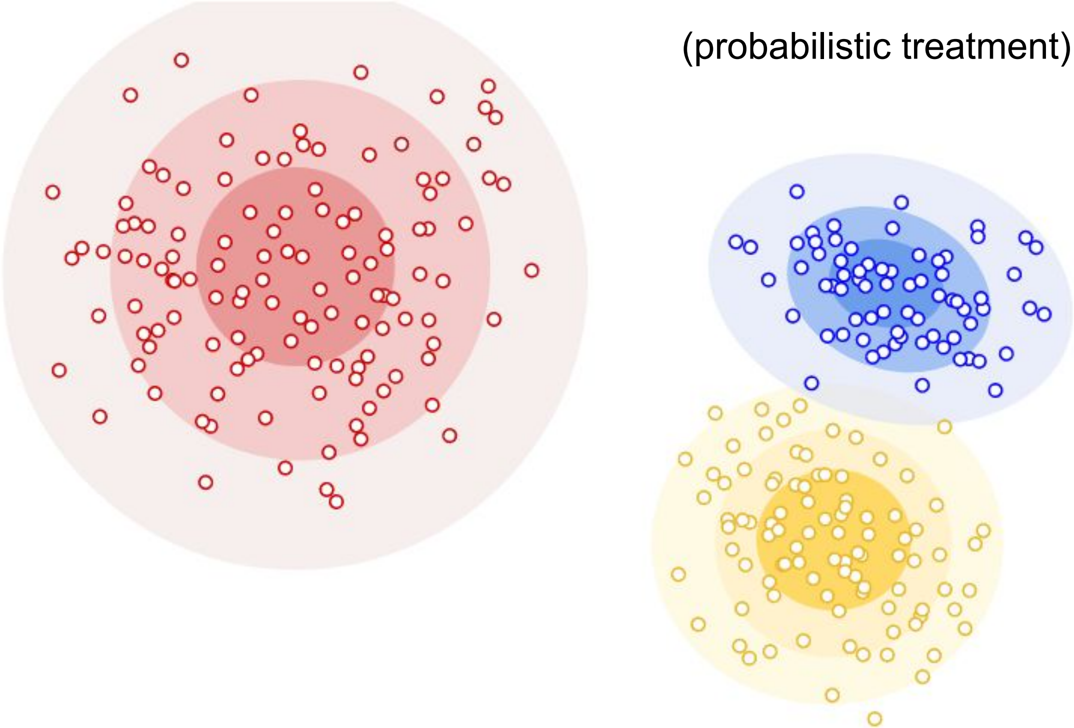
Hierarchical Clustering





Distribution-based Clustering

(probabilistic treatment)

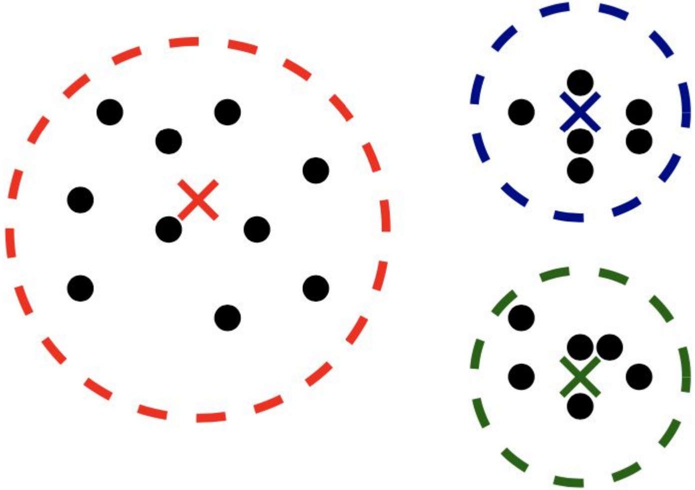


Clustering

- Input: k: the number of clusters (hyperparameter)

$$S = \{x_1, \ldots, x_n\}$$

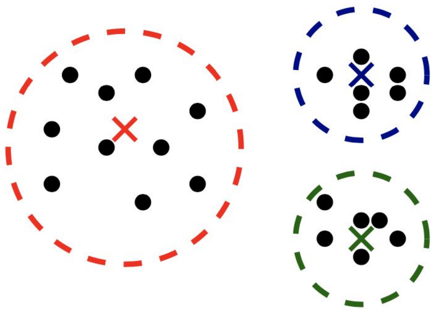
- Output
  - partition $\{G_i\}_{i=1}^{k}$ s.t. $S = \cup_i G_i$ (disjoint union).
  - often, we also obtain 'centroids'



- Q: if we are given the groups, what would be a reasonable definition of centroids?
  - The **point** that has the minimum average **distance** to the datapoints?
  - The **datapoint** that has the minimum average **distance** to the datapoints?
  - The **point** that has the minimum average **squared distance** to the datapoints?

=> Turns out, the last one corresponds to the average point!

k-means Clustering



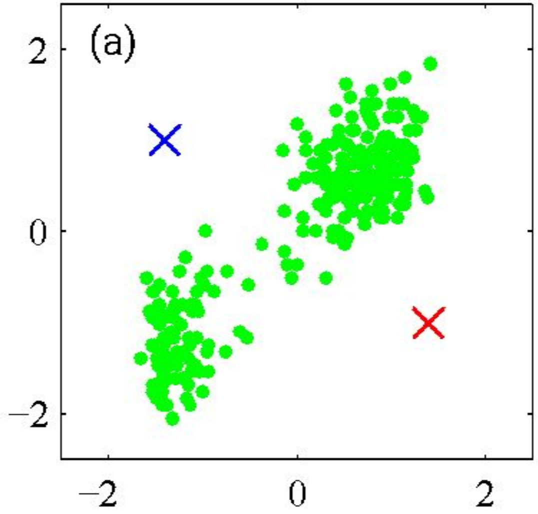
Lloyd's algorithm: solve it approximately (heuristic)

Observation: The chicken-and-egg problem.

- If you knew the cluster assignments… just find the centroids as the average
- If you knew the centroids… make cluster assignments by the closest centroid.

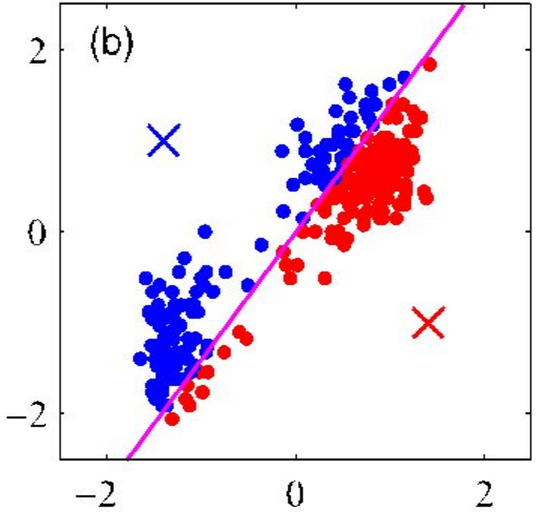
Why not: start from some centroids and then alternate between the two?

Initialization

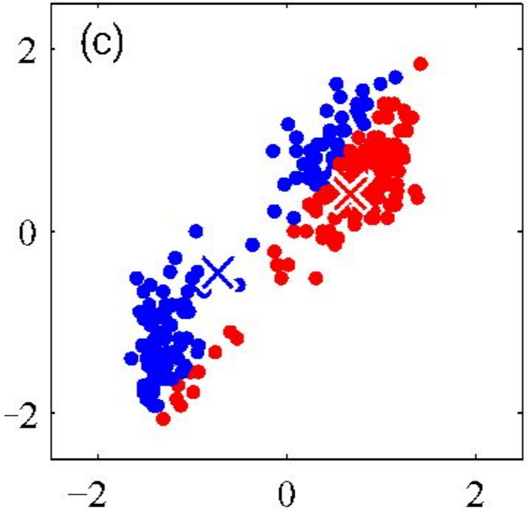


Arbitrary/random initialization of c_1 and c_2

Iteration 1

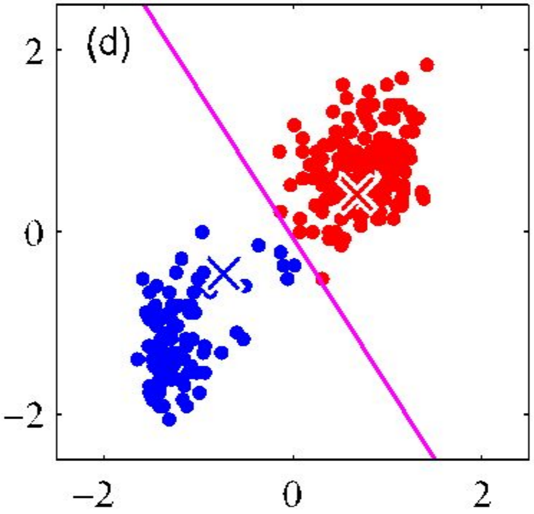


(A) update the cluster assignments.

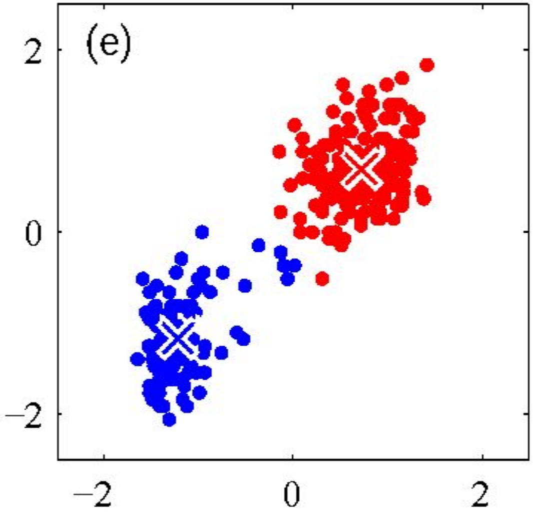


(B) Update the centroids $\{c_j\}$

Iteration 2

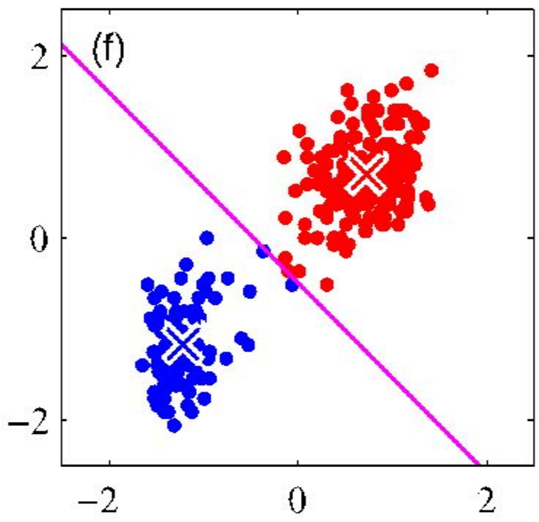


(A) update the cluster assignments.

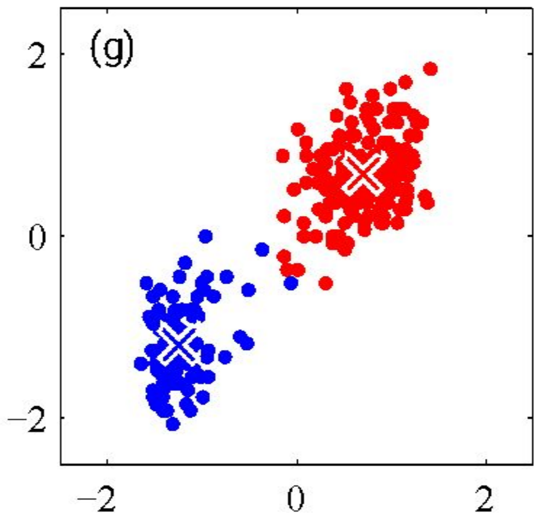


(B) Update the centroids $\{c_j\}$

Iteration 3

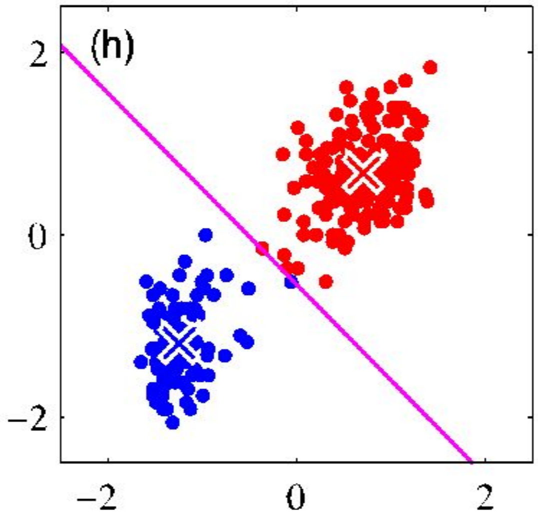


(A) update the cluster assignments.

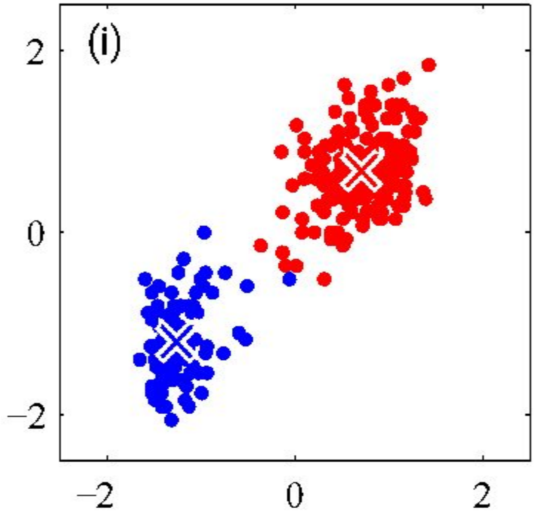


(B) Update the centroids $\{c_j\}$

Iteration 4

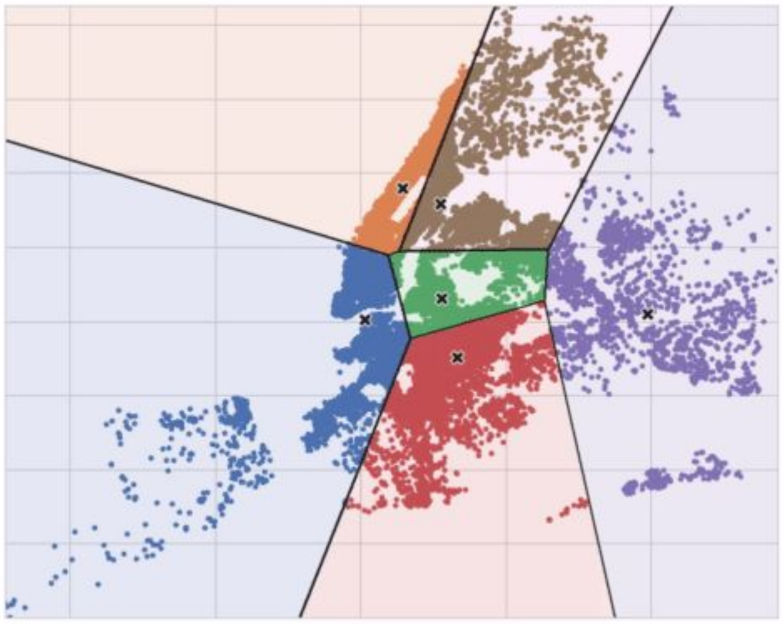


(A) update the cluster assignments.



(B) Update the centroids $\{c_j\}$

Iterating until Convergence



Animation from [Kaggle]

k-means clustering

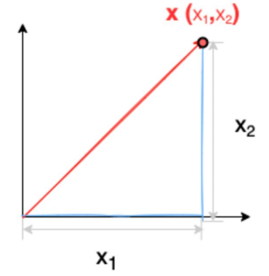
Input: k: num. of clusters, $S = \{x_1, \dots, x_n\}$

[Initialize] Pick $c_1, \dots, c_k$ as randomly selected points from S (see next slides for alternatives)

For t=1,2,…,max_iter

- ***[Assignments]*** $\forall x \in S, \quad a_t(x) = \arg\min_{j \in [k]} \left\| x - c_j \right\|_2^2$
- If $\mathrm{t} \neq 1 \quad \text{AND} \quad a_t(x) = a_{t-1}(x), \forall x \in S$
  - break

$$||x||_2 = \sqrt{x_1^2 + x_2^2}$$



k-means clustering

Input: k: num. of clusters, $S = \{x_1, \dots, x_n\}$

[Initialize] Pick $c_1, \dots, c_k$ as randomly selected points from S (see next slides for alternatives)

For t=1,2,…,max_iter

- ***[Assignments]*** $\forall x \in S, \quad a_t(x) = \arg\min_{j \in [k]} \|x - c_j\|_2^2$
- If t $\neq 1$ $\;\text{AND}\;$ $a_t(x) = a_{t-1}(x), \forall x \in S$
  - break
- ***[Centroids]*** $\forall j \in [k], \quad c_j \leftarrow \text{average}(\{x \in S : a_t(x) = j\})$

Output: $c_1, \dots, c_k$ and $\{a_t(x_i)\}_{i \in [n]}$

But,

It may converge to a local rather than global minimum.

number of clusters

number of cases

centroid for cluster *j*

case *i*

$$\text{objective function} \leftarrow J = \sum_{j=1}^{k} \sum_{i=1}^{n} \underbrace{\left\| x_i^{(j)} - c_j \right\|^2}_{\text{Distance function}}$$

Local optima

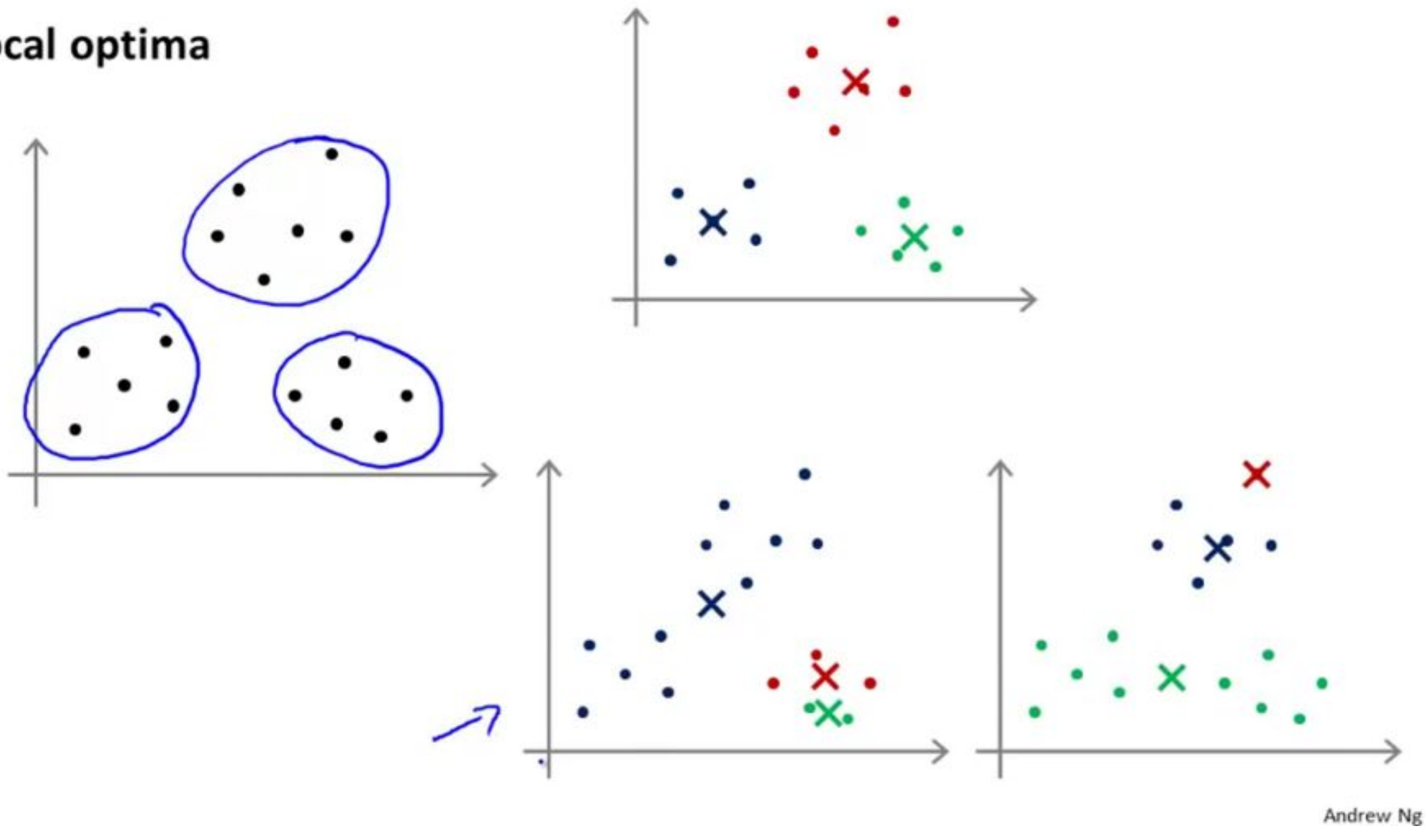


Image from Andrew NG Coursera Machine Learning Course

Issue 1: Unreliable solution

- You usually get suboptimal solutions
- You usually get different solutions every time you run.

- **Standard practice**: Run it 50 times and take the one that achieves the smallest objective function
  - Recall: $\min_{c_1,\ldots,c_k} \sum_{i=1}^{n} \min_{j\in[k]} \left\|x_i - c_j\right\|_2^2$

    Each run of algorithm outputs $c_1, \ldots, c_k$. Compute this to evaluate the quality!

- And/or, change the initialization (next slide)
  - Idea: ensure that we pick a widespread $c_1, \ldots, c_k$

Alternative initialization

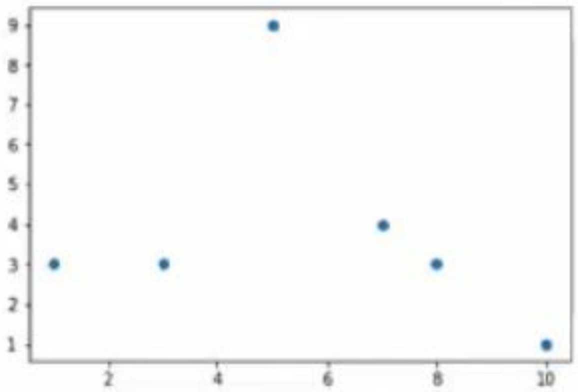
- ***k*-means++**
  - Pick $c_1 \in \{x_1, \dots, x_n\}$ uniformly at random
  - For $j = 2, \dots, k$
    - Define a distribution $\forall i \in [n]$, $\mathbb{P}(c_j = x_i) \propto \min_{j'=1,\dots,j-1} \|x_i - c_{j'}\|_2^2$
    - Draw c_j from the distribution above.

More likely to choose x_i that is farthest from already-chosen centroids.

=> has a mathematical guarantee that it will be better than an arbitrary starting point!

Suppose we have the small dataset [(7,4),(8,3),(5,9),(3,3),(1,3),(10,1)] to which we wish to assign 3 clusters.
We begin by randomly selecting (7,4) to be a cluster center.

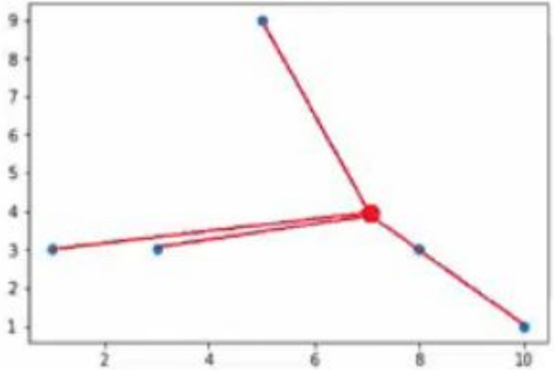
x	$\min(d(x, z_i)^2)$
(7,4)	
(8,3)	
(5,9)	
(3,3)	
(1,3)	
(10,1)	



From Sara Jensen's Youtube Channel

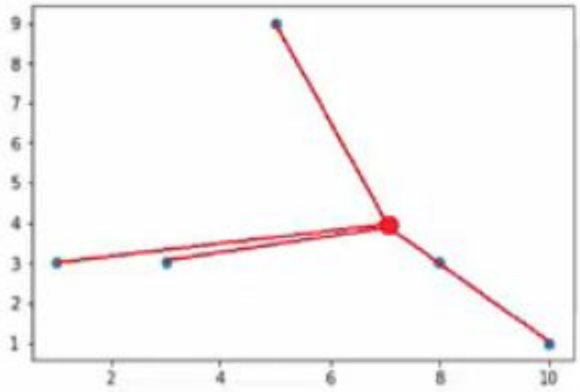
Suppose we have the small dataset [(7,4),(8,3),(5,9),(3,3),(1,3),(10,1)] to which we wish to assign 3 clusters.
We begin by randomly selecting (7,4) to be a cluster center.

x	$\min(d(x, z_i)^2)$
(7,4)	-
(8,3)	2
(5,9)	29
(3,3)	17
(1,3)	37
(10,1)	18



Suppose we have the small dataset
[(7,4),(8,3),(5,9),(3,3),(1,3),(10,1)] to which we wish to assign 3 clusters.
We begin by randomly selecting (7,4) to be a cluster center.

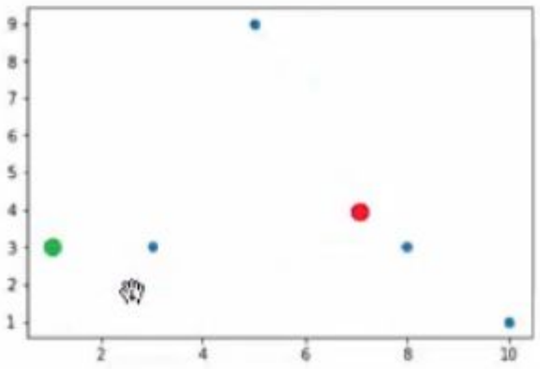
x	prob
(7,4)	-
(8,3)	2/103
(5,9)	29/103
(3,3)	17/103
(1,3)	37/103
(10,1)	18/103



Suppose we have the small dataset [(7,4),(8,3),(5,9),(3,3),(1,3),(10,1)] to which we wish to assign 3 clusters.

We add (1,3) to the list of cluster centers.

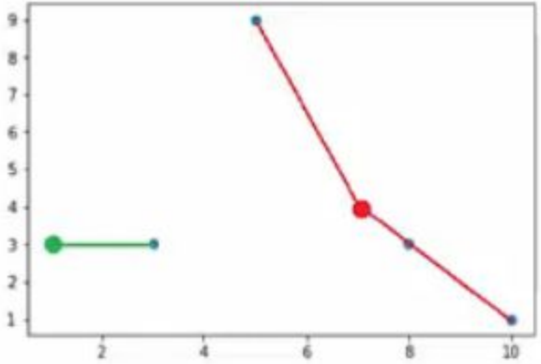
x	$\min(d(x,z_i)^2)$
(7,4)	-
(8,3)	
(5,9)	
(3,3)	
(1,3)	-
(10,1)	



Suppose we have the small dataset [(7,4),(8,3),(5,9),(3,3),(1,3),(10,1)] to which we wish to assign 3 clusters.

We add (1,3) to the list of cluster centers.

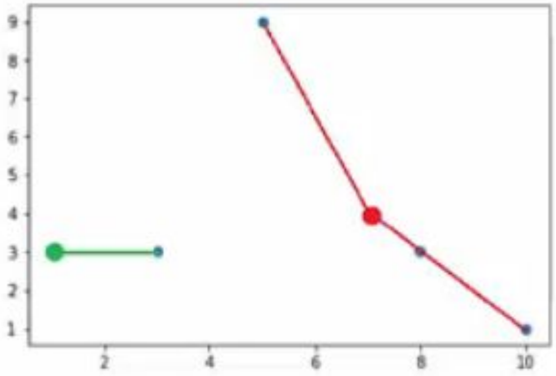
x	$\min(d(x, z_i)^2)$
(7,4)	-
(8,3)	2
(5,9)	29
(3,3)	4
(1,3)	-
(10,1)	18



Suppose we have the small dataset [(7,4),(8,3),(5,9),(3,3),(1,3),(10,1)] to which we wish to assign 3 clusters.

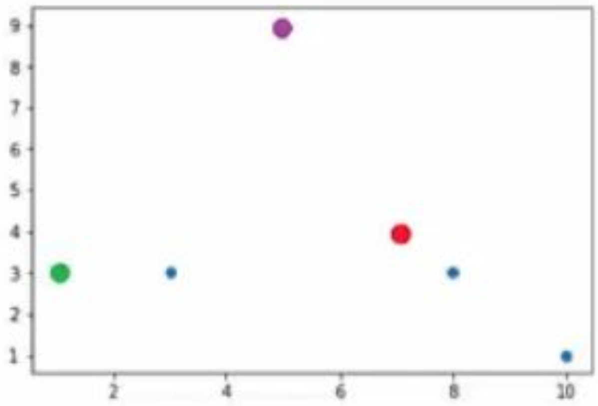
We add (1,3) to the list of cluster centers.

x	prob
(7,4)	-
(8,3)	2/53
(5,9)	29/53
(3,3)	4/53
(1,3)	-
(10,1)	18/53



Suppose we have the small dataset
[(7,4),(8,3),(5,9),(3,3),(1,3),(10,1)] to which we wish to assign 3 clusters.
We add (5,9) to the list of cluster centers.

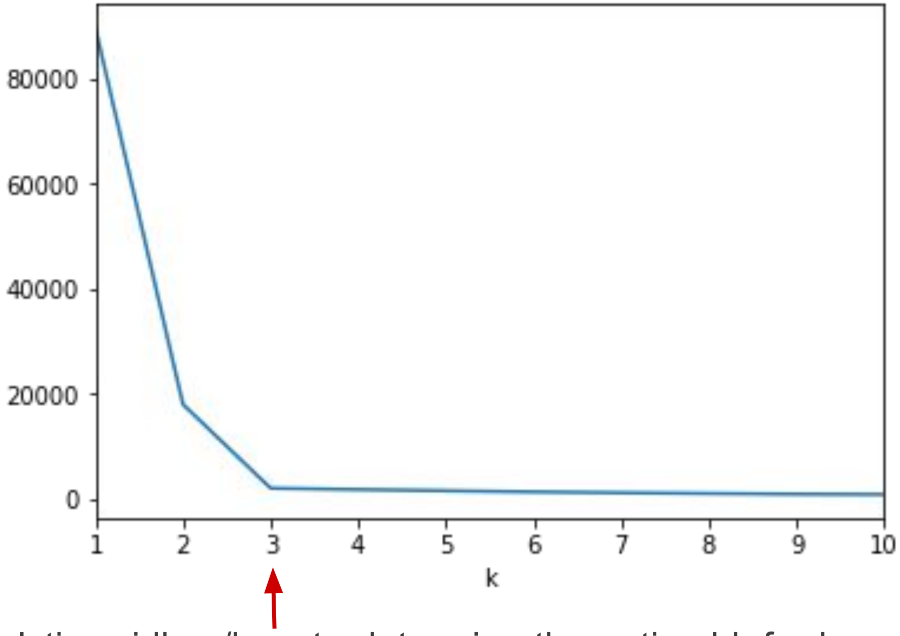
x	prob
(7,4)	-
(8,3)	
(5,9)	-
(3,3)	
(1,3)	-
(10,1)	



Issue 2: Choose k

- No principled way.
- Elbow method: calculate Within-Cluster-Sum of Squared Errors (WSS) and choose k where WSS starts to diminish.

Objective function



https://medium.com/analytics-vidhya/how-to-determine-the-optimal-k-for-k-means-708505d204eb