Computer Science

CSC380: Principles of Data Science

Clustering

Credit:
- Jason Pacheco,
- Kwang-Sung Jun,
- Chicheng Zhang
- Xinchen yu

1

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

3

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

Doc1 : Health , Medicine, Doctor

Doc 5 : Covid, Health , Doctor

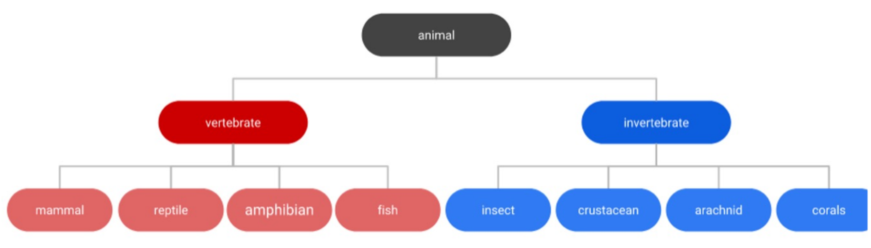
Doc 3 : Environment, Planet

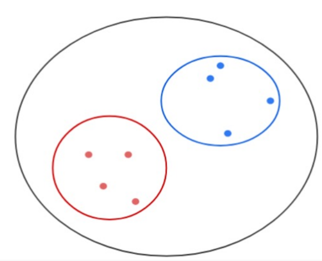
Doc 4 : Pollution, Climate Crisis

Doc 2 : Machine Learning, Computer

5

Hierarchical Clustering





7

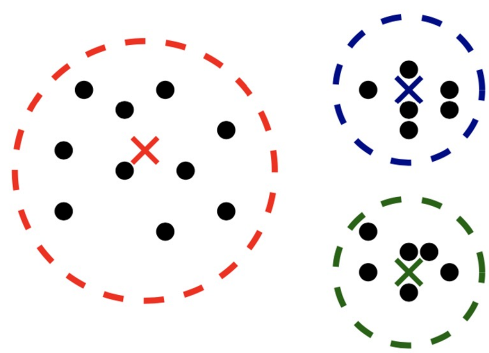
What is unsupervised learning?

- 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. (unsupervised)
  - **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
    - Soft clustering – a topic might be 30% in cluster 1, and 70% in cluster 2.

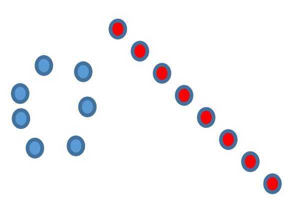
Clustering

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

  $S = \{x_1, \dots, x_n\}$
- Output
  - partition $\{G_i\}_{i=1}^{k}$ s.t. $S = \cup_i G_i$ (disjoint union).
  - often, we also obtain 'centroids'



Sometimes it is trickier to define centroid



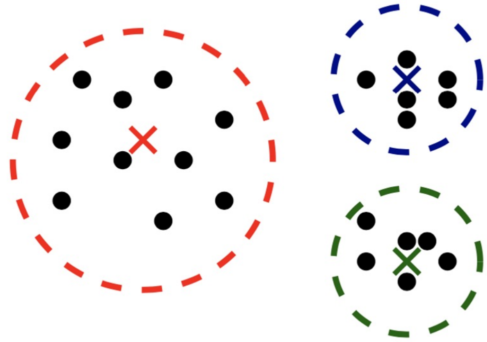
Sometimes addressed by
Spectral methods,
or dimensionality reductions

Warmup

10

- For a set of points $S = \{p_1 \ldots p_n\}$, find a point c minimizing

$$\sum_i dist^2\ (p_i, c) = \sum_i (p_i - c)^2$$



Solution $c = \frac{1}{n}\left(\sum x(p_i), \sum y(p_i)\right)$

Center of mass, centroid

Other common disance functions:

1) minimize radius of enclosing ball. (that is, minimize distance from center of cluster to furthest point
2) Minimize distances $\sum distance(p_i, c)$

Total quality of clustering: Max, sum or sum of squares of distances of all clusters

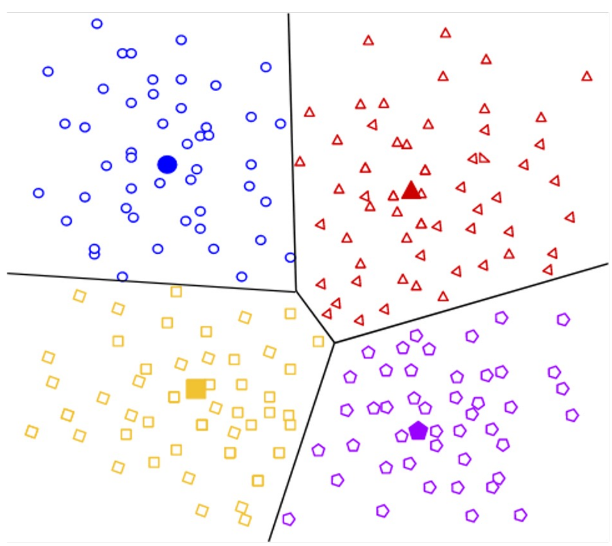
10

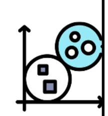
Centroid-based Clustering

If the locations of the centauroids is fixed, Then clusters are created in an obvious way: Each data point is assigned to the nearest centroid.

So the cluster of the red triangles are all data points whose distance to red triangle <distance to other centroids.

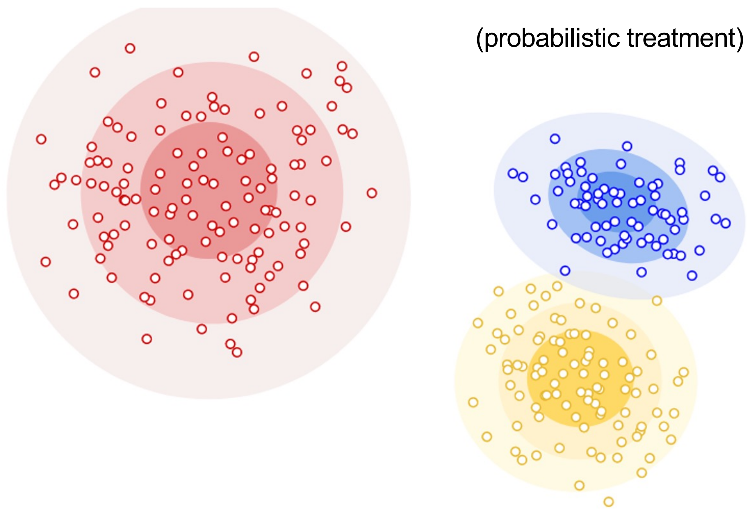
Question: How to pick centroid's location?





11

Distribution-based Clustering



12

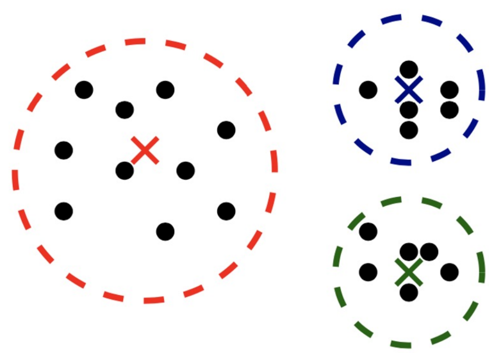
Clustering

13

- 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'



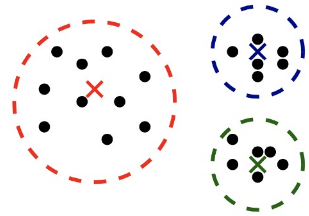
- Q: if we are given the groups, what would be a reasonable definition of centroids?
  - The **point** that has the minimum average **distance** to the datapoints?
  - The **datapoint** that has the minimum average **distance** to the datapoints?
  - The **point** that has the minimum average **squared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

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

13

k-means Clustering

14



Lloyd's algorithm: solve it approximately (heuristic)

Observation: The chicken-and-egg problem.

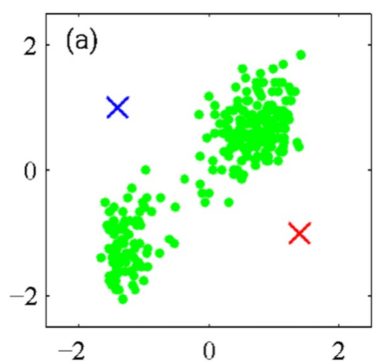
- If you knew the cluster assignments… just find the centroids as the average
- If you knew the centroids… make cluster assignments by the closest centroid.

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

14

Initialization

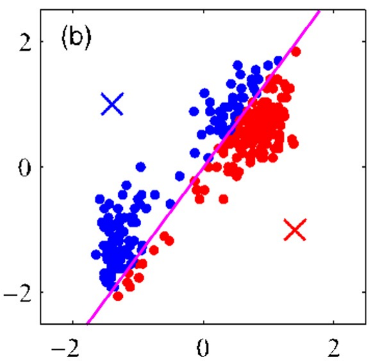
15



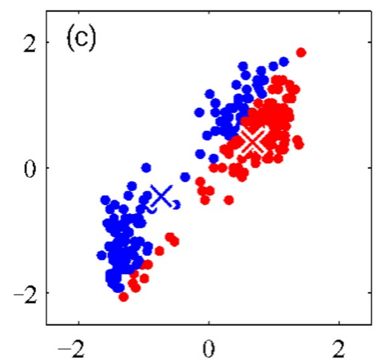
Arbitrary/random initialization of c_1 and c_2

15

Iteration 1



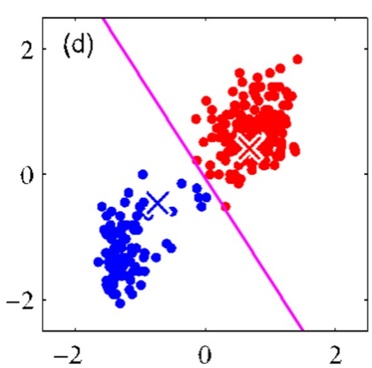
(A) update the cluster assignments.



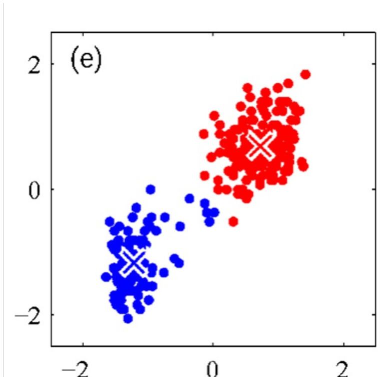
(B) Update the centroids $\{c_j\}$

16

Iteration 2



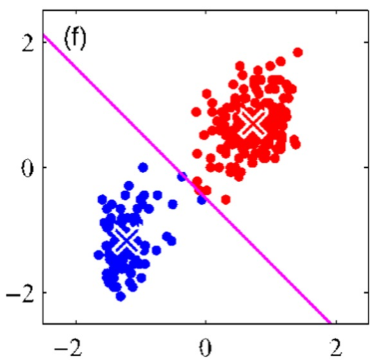
(A) update the cluster assignments.



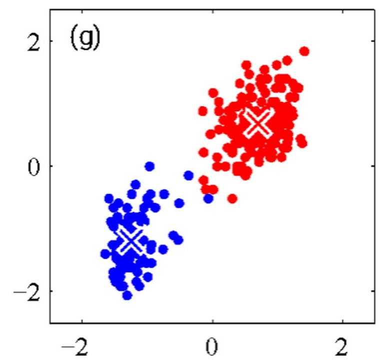
(B) Update the centroids $\{c_j\}$

17

Iteration 3



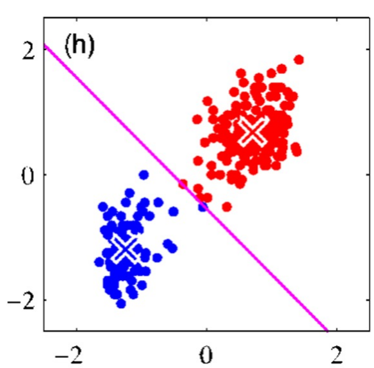
(A) update the cluster assignments.



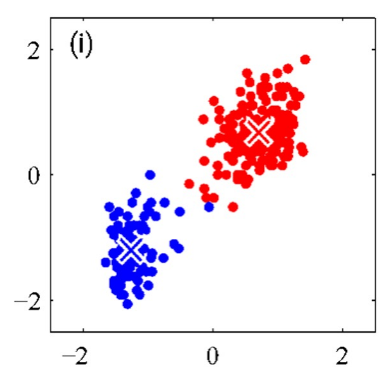
(B) Update the centroids $\{c_j\}$

18

Iteration 4



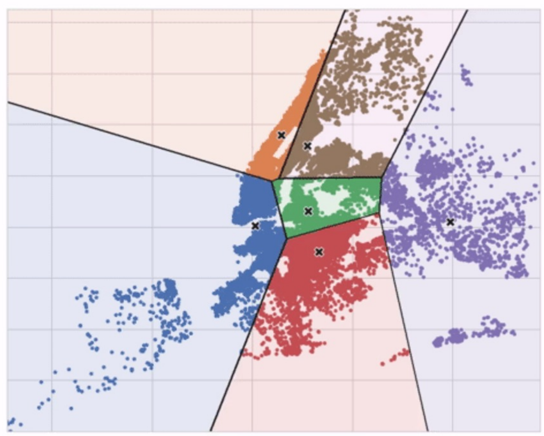
(A) update the cluster assignments.



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

19

Iterating until Convergence



Animation from Kaggle

20

k-means clustering

21

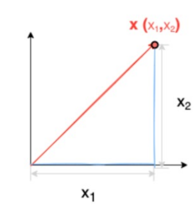
Input: k: num. of clusters, $S = \{x_1, \dots, x_n\}$

[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

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



21

k-means clustering

22

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

Calculate for each data point the nearest cluster head

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

Cluster j are all data points with $a_t(x) = j$

- ***[Centroids]*** $\forall j \in [k], \quad c_j \leftarrow \text{average}(\{x \in S: a_t(x) = j\})$

Could be replaced by
center of smallest disk containing
points points in ths cluster

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

22

But,

It may converge to a local rather than global minimum.

number of clusters

number of cases

centroid for cluster *j*

case *i*

objective function $\leftarrow$

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

23

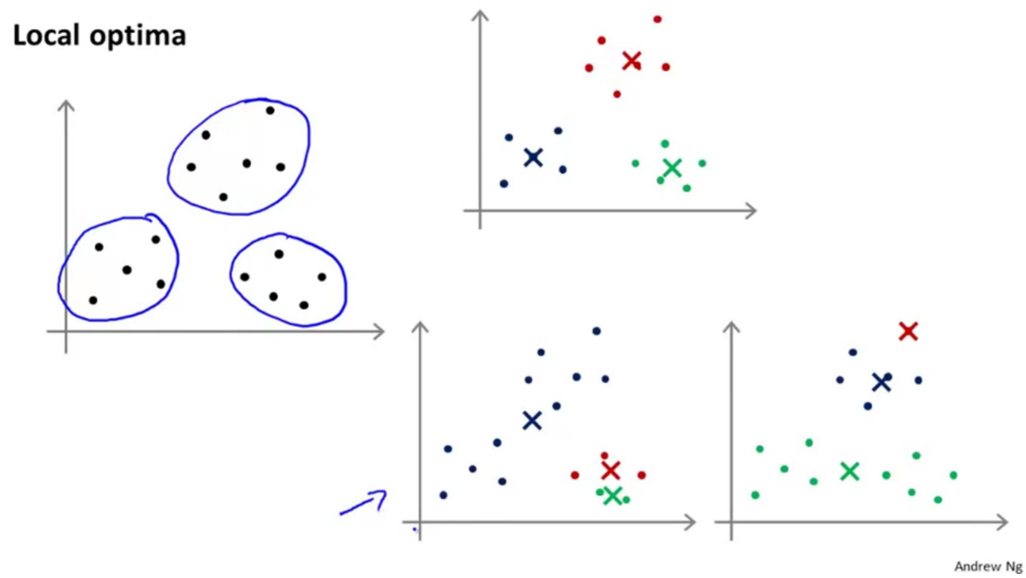


Image from Andrew NG Coursera Machine Learning Course

24

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 $c_1, \ldots, c_k$

25

Alternative initialization

26

- ***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.

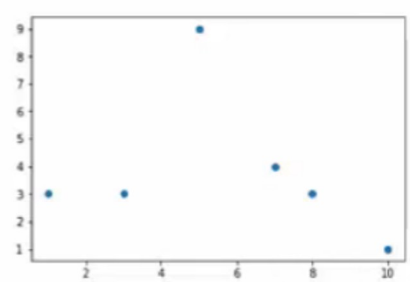
More likely to choose x_i that is farthest from already-chosen centroids.

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

26

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)	

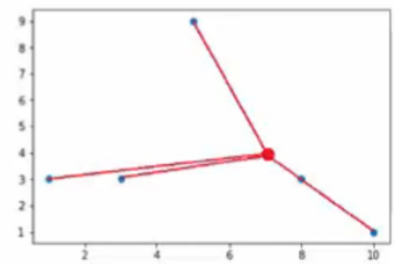


From Sara Jensen's Youtube Channel

27

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.

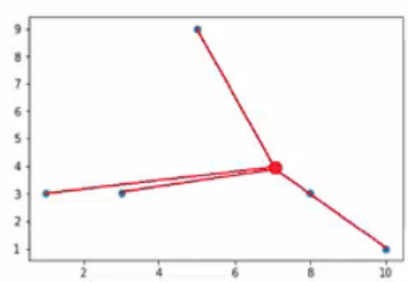
x	$\min(d(x, z_i)^2)$
(7,4)	-
(8,3)	2
(5,9)	29
(3,3)	17
(1,3)	37
(10,1)	18



28

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.

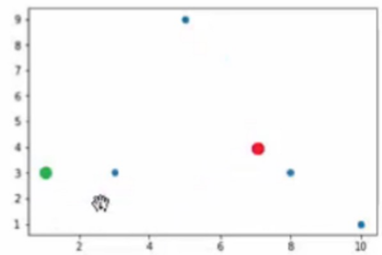
x	prob
(7,4)	-
(8,3)	$2/103$
(5,9)	$29/103$
(3,3)	$17/103$
(1,3)	$37/103$
(10,1)	$18/103$



29

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.

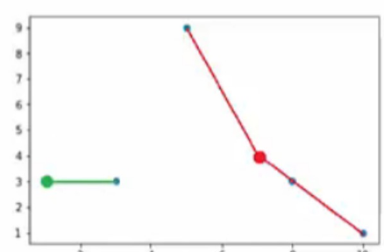
x	$\min(d(x, z_i)^2)$
(7,4)	-
(8,3)	
(5,9)	
(3,3)	
(1,3)	-
(10,1)	



30

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.

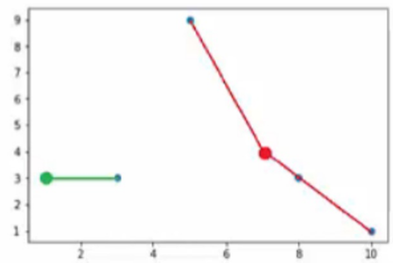
x	$\min(d(x, z_i)^2)$
(7,4)	-
(8,3)	2
(5,9)	29
(3,3)	4
(1,3)	-
(10,1)	18



31

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.

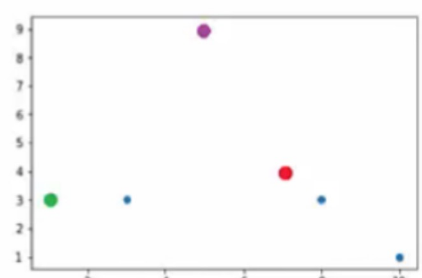
x	prob
(7,4)	-
(8,3)	2/53
(5,9)	29/53
(3,3)	4/53
(1,3)	-
(10,1)	18/53



32

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)	

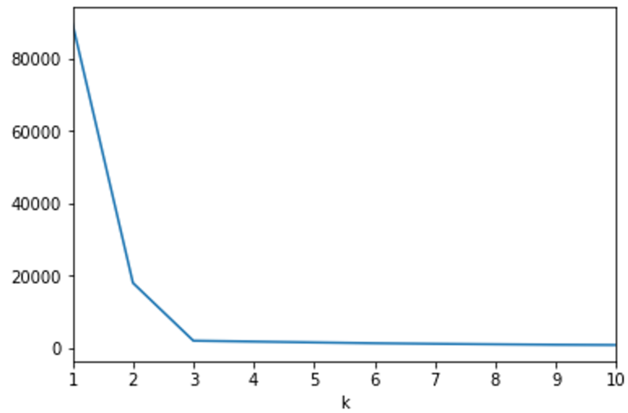


33

Issue 2: Choose k

- No principled way.
- Elbow method: see where you get saturation.

Objective function



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