Clustering and k-Means

Jiahui Chen
Department of Mathematical Sciences
University of Arkansas

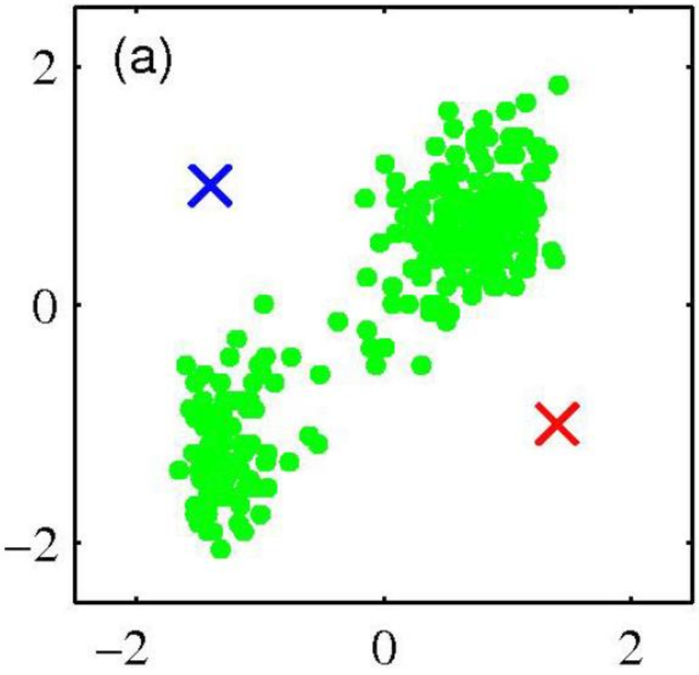
Introduction

- **Clustering**
  - Unsupervised learning (can be used for semi-supervised learning too)
  - Requires no labels
  - Detect patterns
    - Group emails
    - Websites
    - Regions of images, …
  - Useful when you do not know what you are looking for

Examples

- Image segmentation

Algorithm



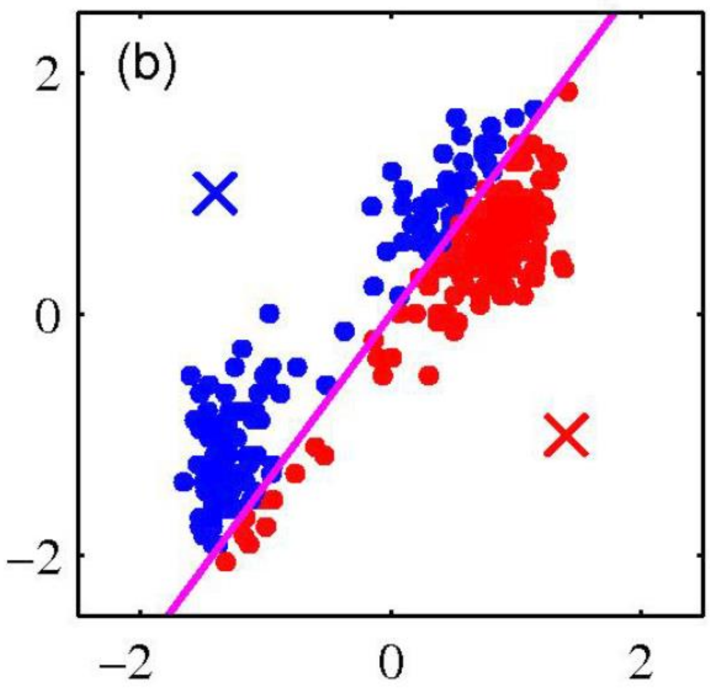
- Want to group in

$k = 2$ clusters

- Pick 2 random point as

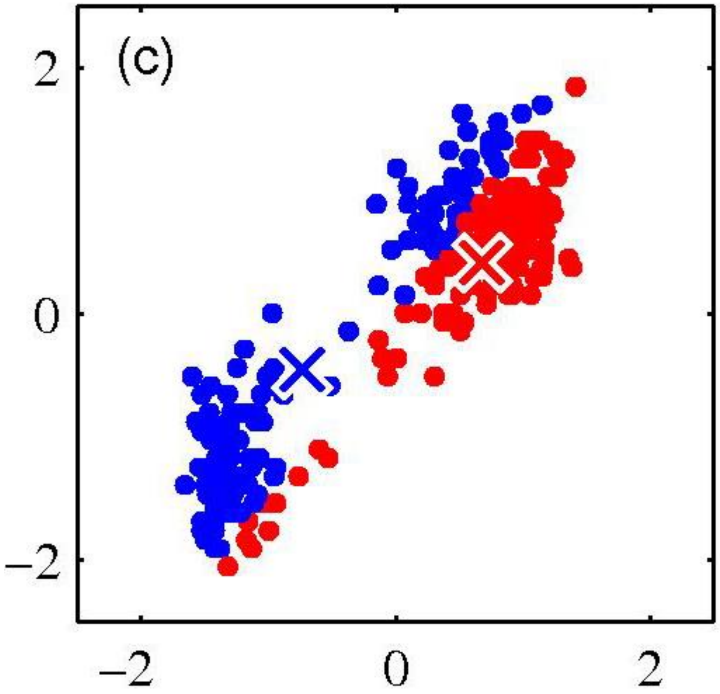
cluster centers

Algorithm



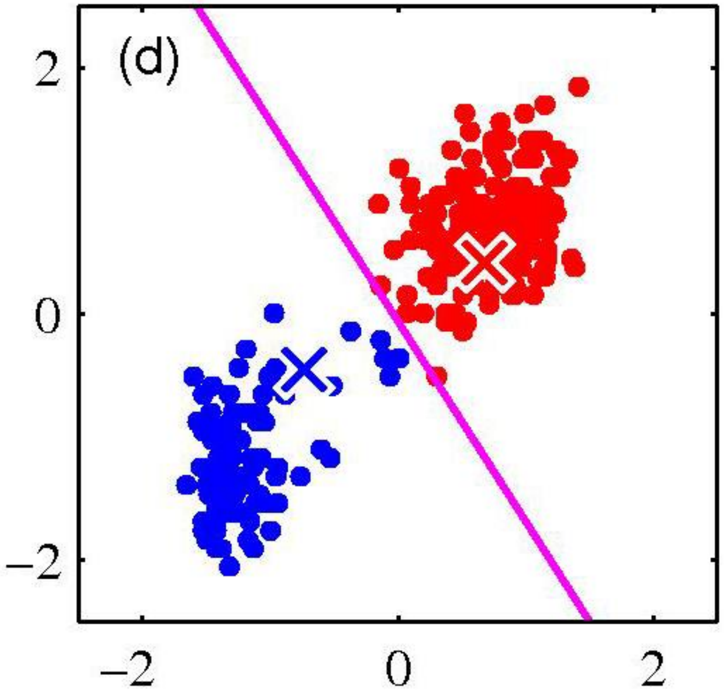
- Assign data points to closest cluster center

Algorithm



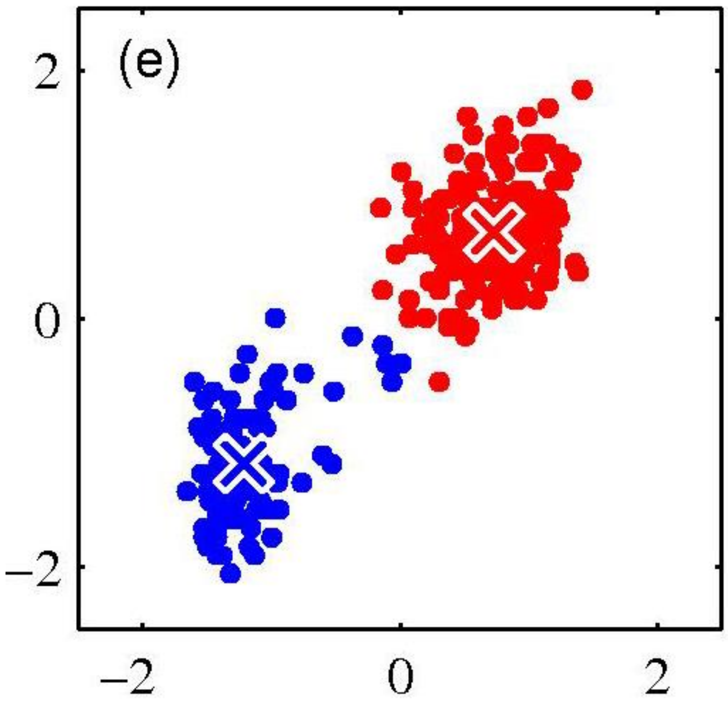
- Change cluster center to the average of the assigned data points

Algorithm



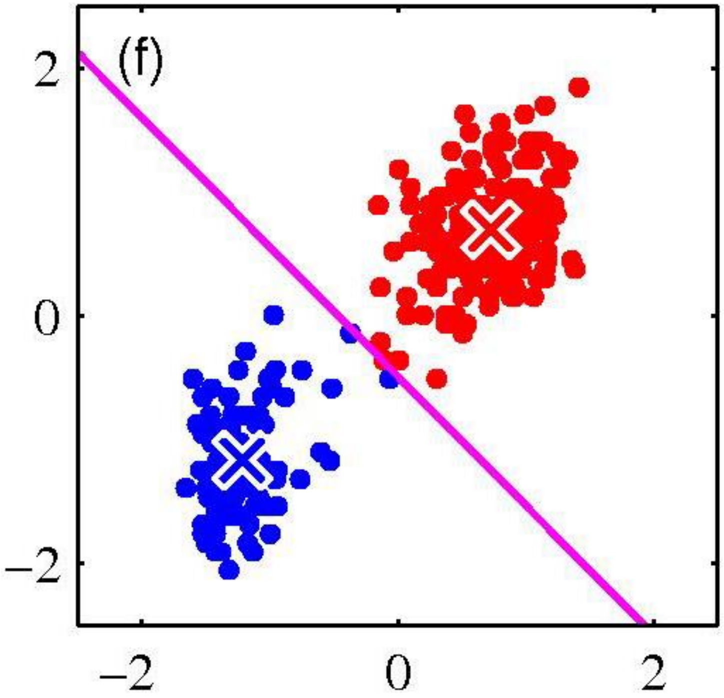
- Repeat until no change in the cluster center

Algorithm



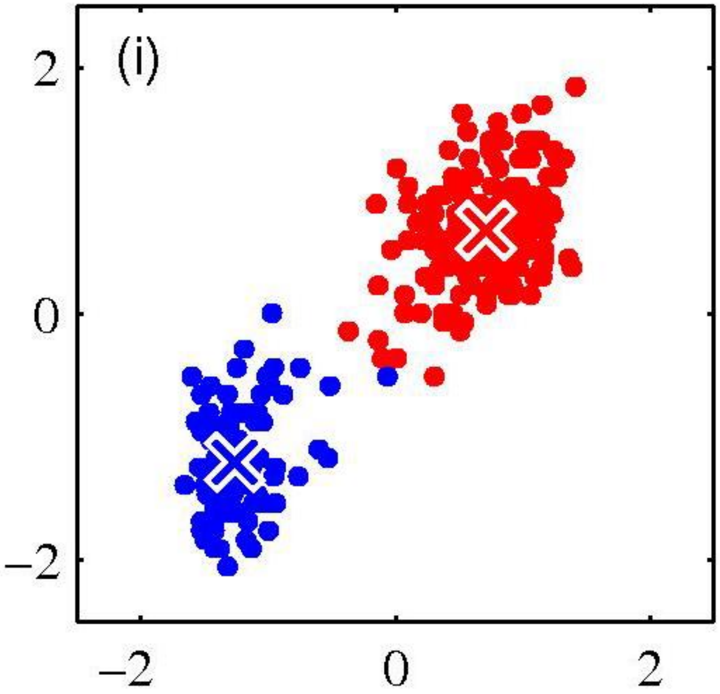
- Repeat until no change in the cluster center

Algorithm



- Repeat until no change in the cluster center
- Show a dividing boundary

Algorithm



- Repeat until no change in the cluster center

Algorithm

- Summary:
  1. Pick k random points as cluster centers
  2. Repeat:
     a) Assign data points to closest cluster center
     b) Change the cluster center to the average of its data points
  3. Until no change in the cluster centers
- Can use distance metrics as discussed in *k*-NN lecture: Euclidean, Manhattan, Minkowski, etc.

K-Means Property

- Always converge in a finite number of iterations

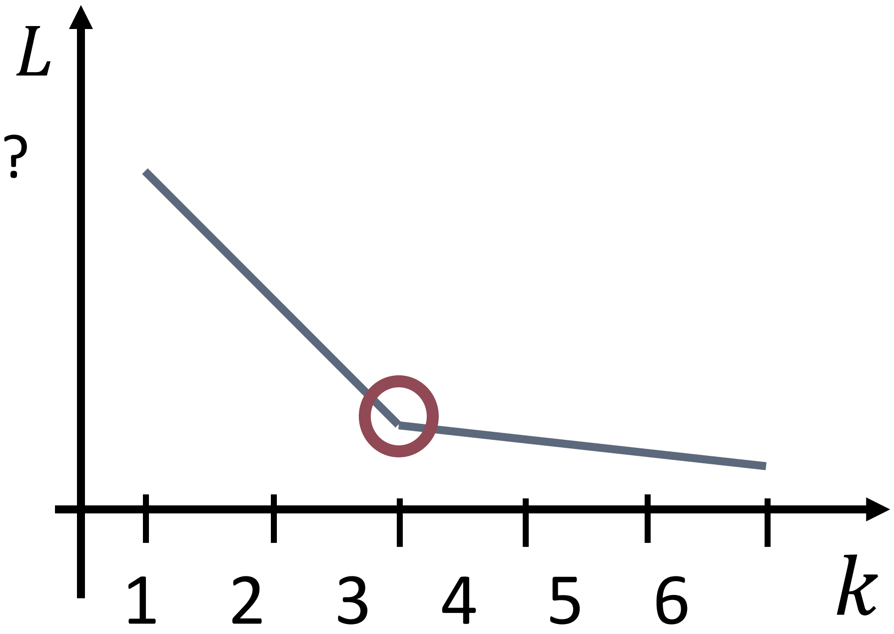
Given a finite set of data points $\boldsymbol{x} = \{x_1, x_2, \dots, x_n\}$ in R^d, the k-means cluster aim to find a partition $\boldsymbol{S} = \{S_1, S_2, \dots, S_k\}$ $(k < n)$. The mean square error (MSE) is minimized

$$\operatorname*{Arg\,min}_{\boldsymbol{S}} \sum_{j=1}^{k} \sum_{\boldsymbol{x} \in S_j} \left\| \boldsymbol{x} - \mu_j \right\|^2$$

if $$\mu_j = \frac{1}{\|S_j\|} \sum_{\boldsymbol{x} \in S_j} \boldsymbol{x}$$

How to Choose K

- Should not do it automatically
- Can we do cross-validation?
- Visualization
- Based on additional information of the data
- Plot the cost functions and use the elbow observation



Silhouette Score

This metric measures how similar an object is to its own cluster compared to other clusters

$S(i) = 1$: The sample is far away from the neighboring clusters

$S(i) = 0$: The sample is on or very close to the decision boundary between two neighboring clusters

$S(i) < 0$: The sample might have been assigned to the wrong cluster

The Sihouette Score $S(i)$ for sample i:

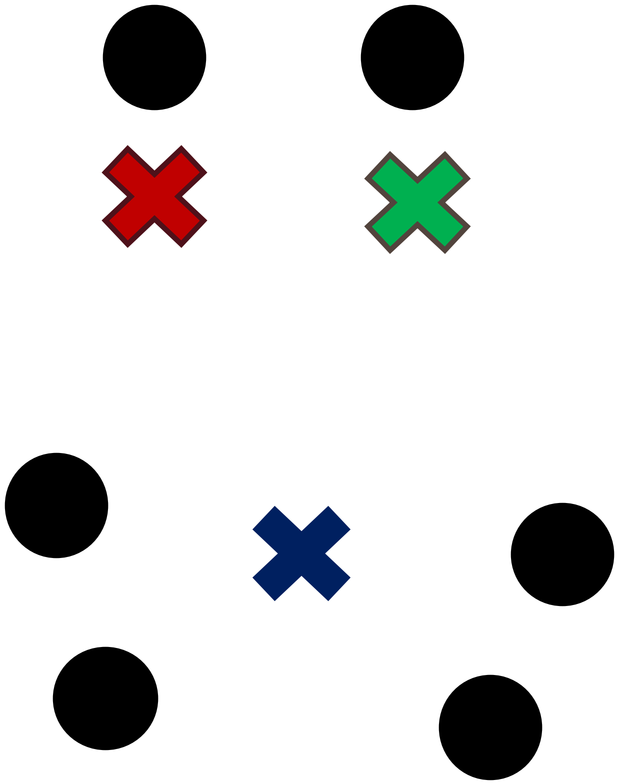
$$S(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

$a(i)$: calculate the average distance from all other points in the same cluster

$b(i)$: calculate the average distance from all points in the nearest cluster that i is not a part of.

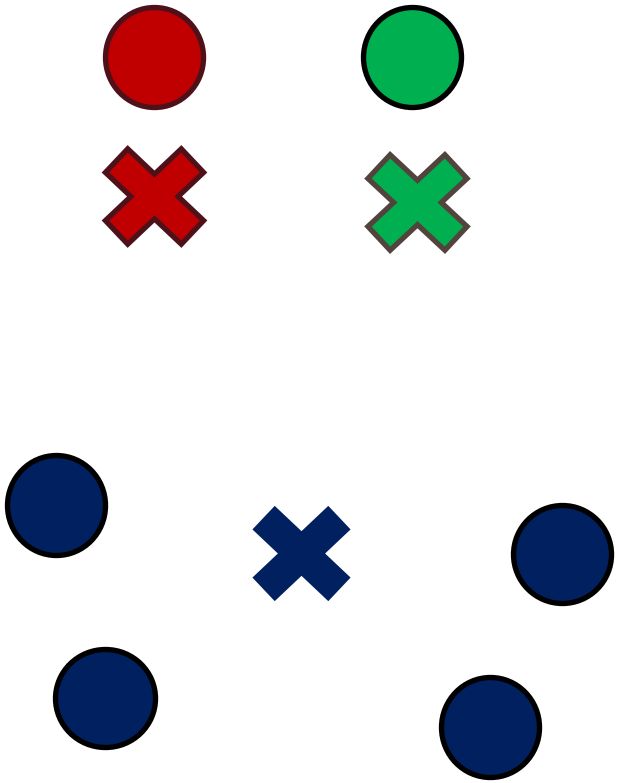
K-Means Property

- Usefulness of k-means depend on what you pick



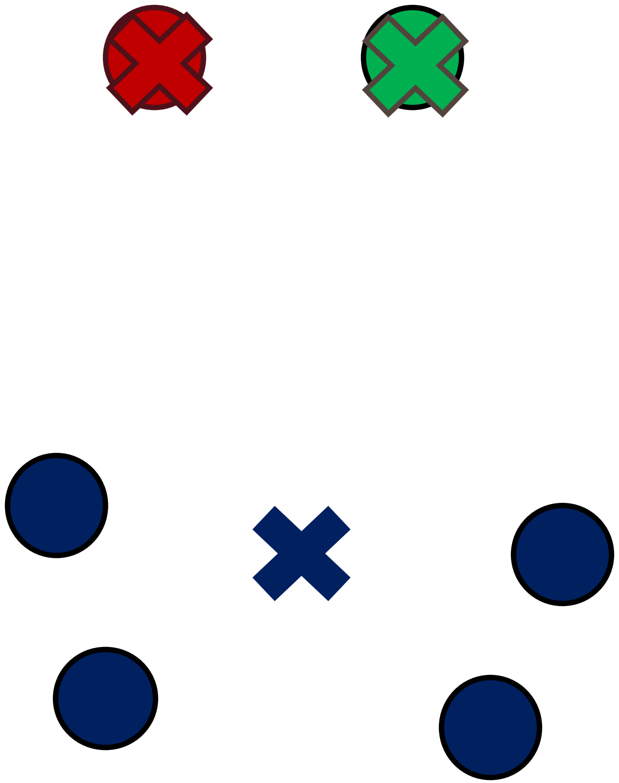
K-Means Property

- Usefulness of k-means depend on what you pick



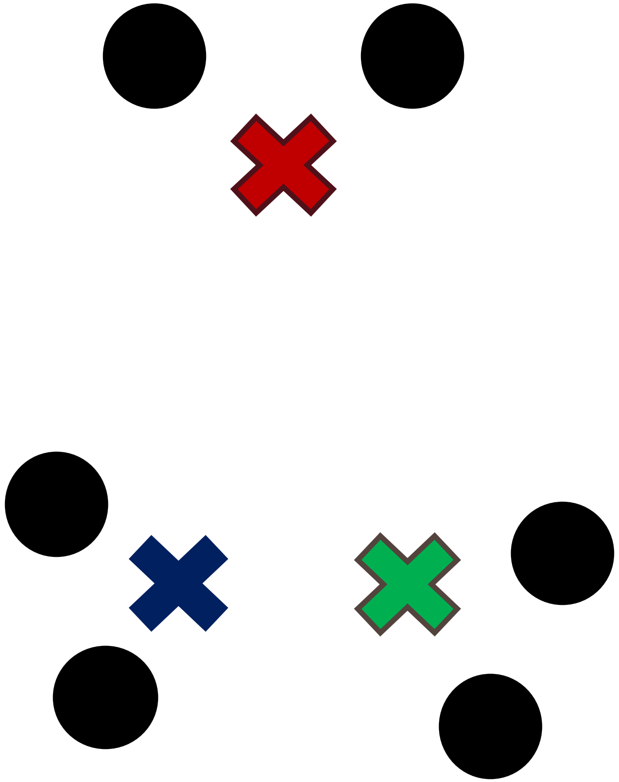
K-Means Property

- Usefulness of k-means depend on what you pick



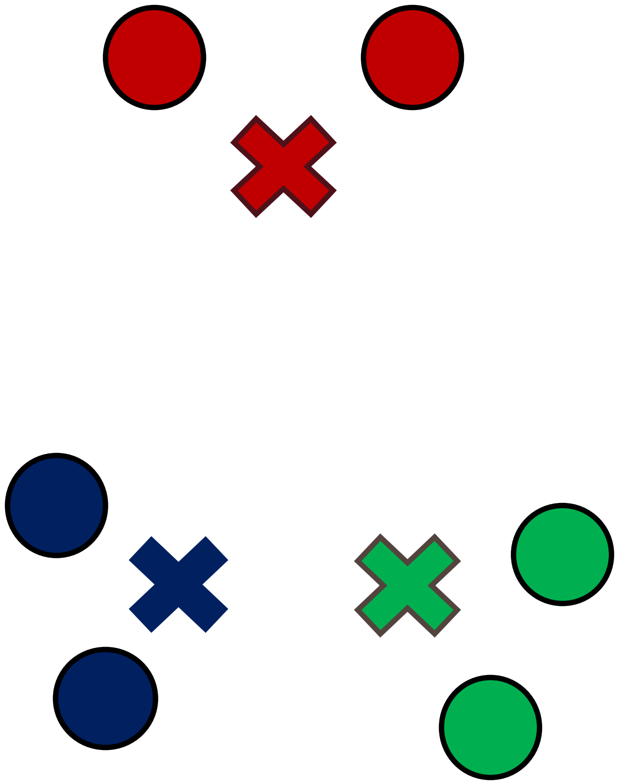
K-Means Property

- Usefulness of k-means depend on what you pick



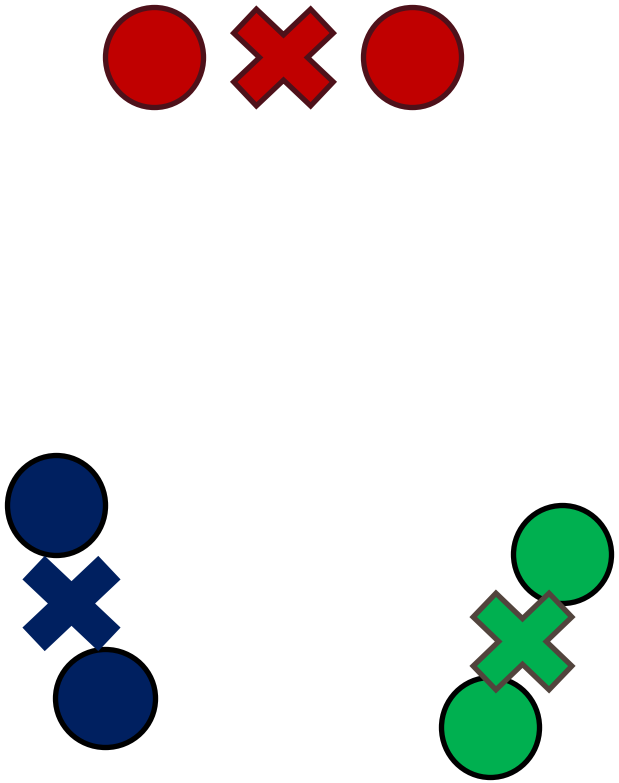
K-Means Property

- Usefulness of k-means depend on what you pick



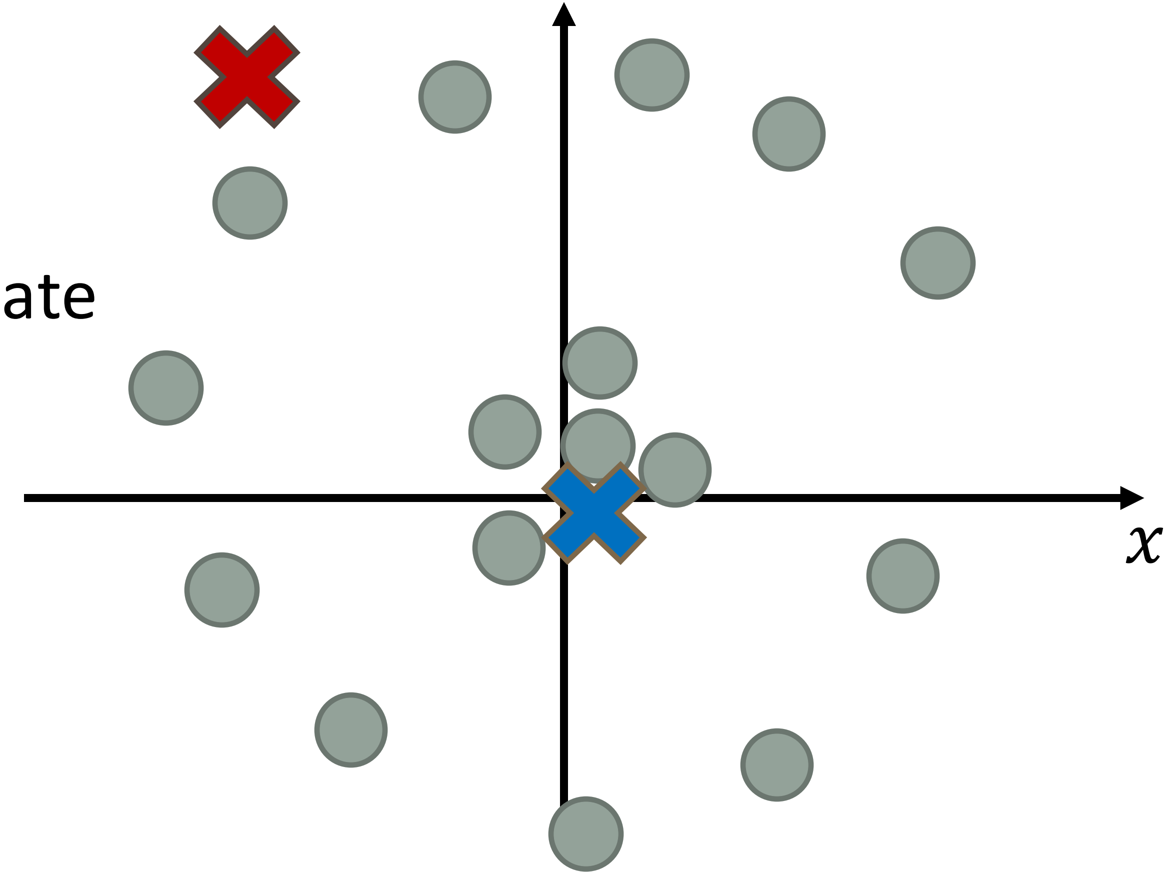
K-Means Property

- Usefulness of k-means depend on what you pick



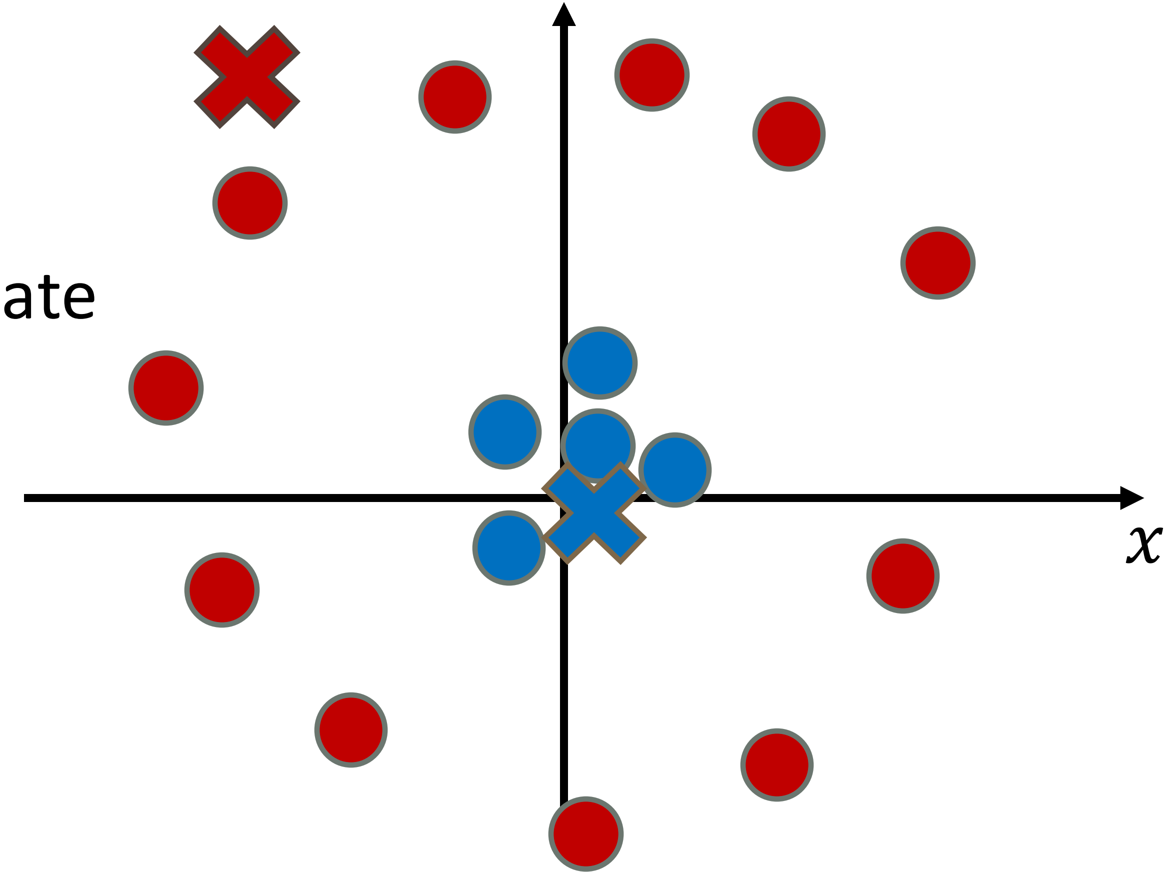
K-Means Property

Changing coordinate might be helpful



K-Means Property

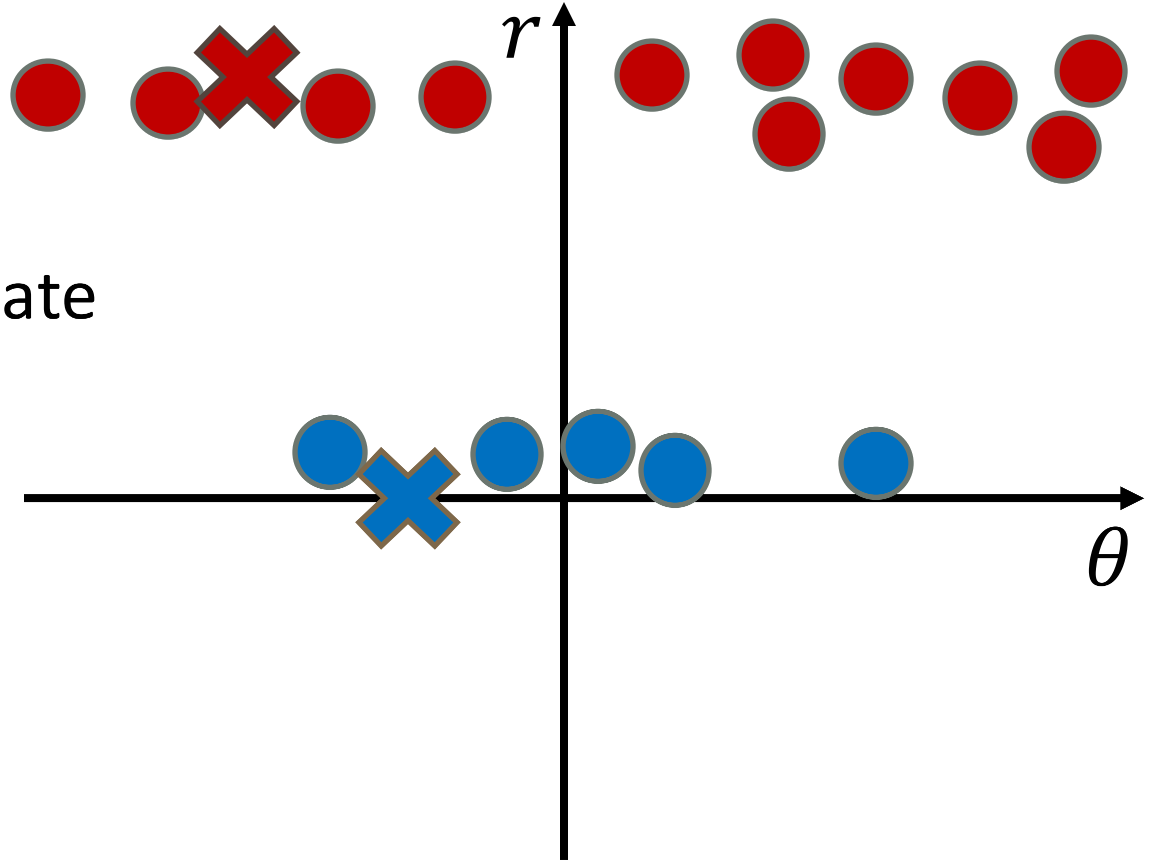
Changing coordinate might be helpful



K-Means Property

Changing coordinate might be helpful

- $x = r\cos\theta$
- $y = r\sin\theta$

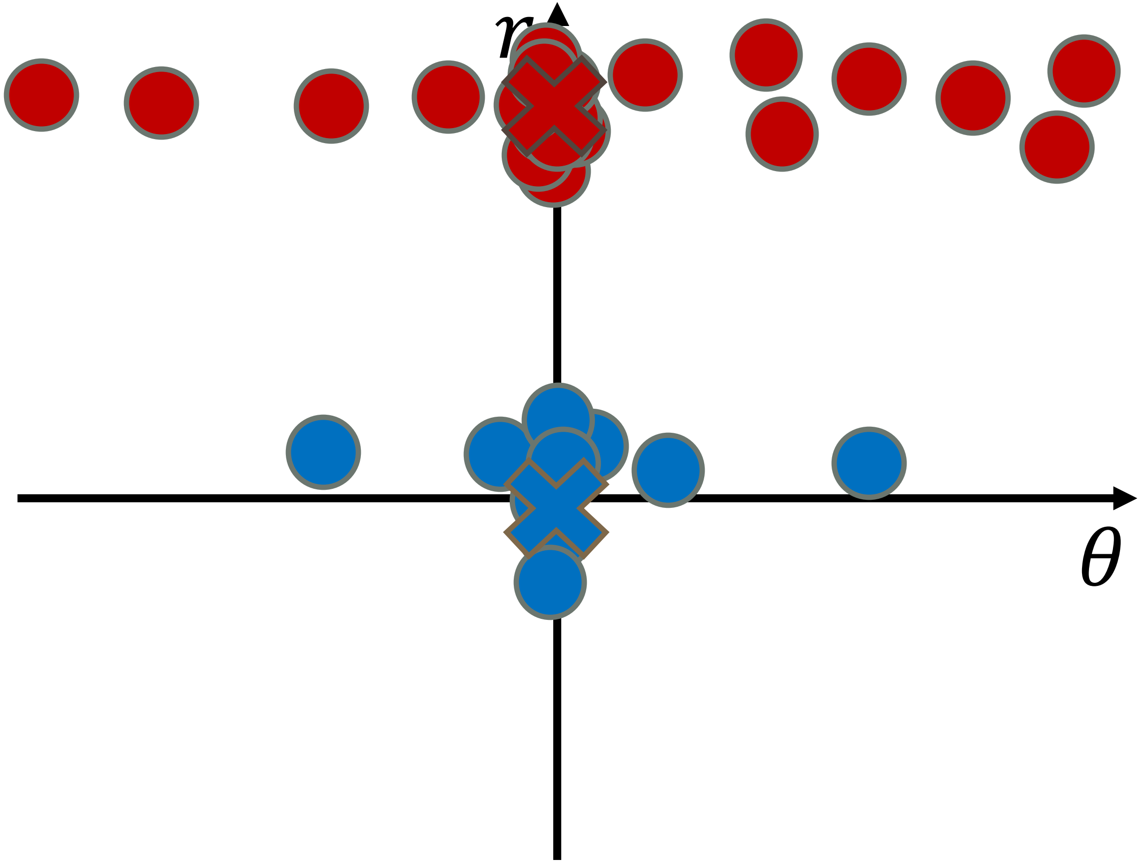


Dimensionality Reduction

PCA

tSNE

UMAP

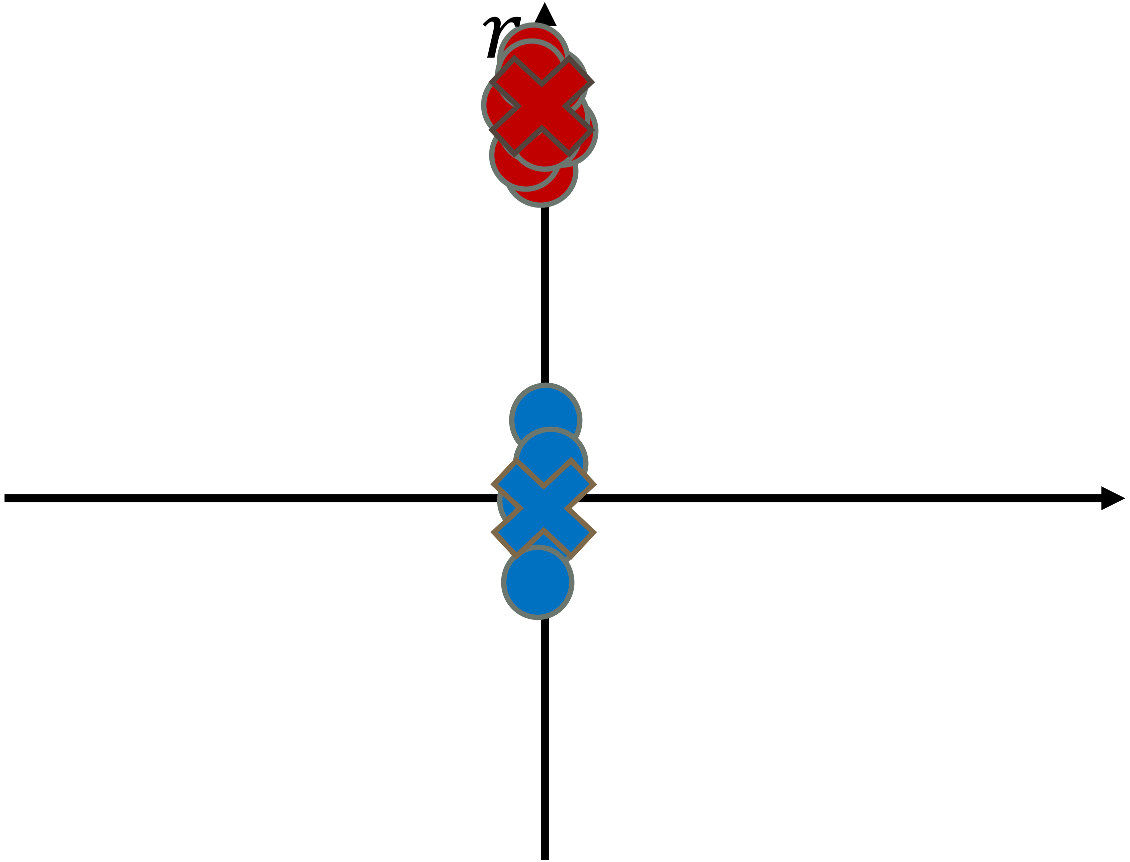


Dimensionality Reduction

PCA

tSNE

UMAP



Discussions

- Various metrics can be applied and lead to various variations, such as Minkowski weighted k-means, etc.
- Dimensionality reduction can be critical for the successful application of K-means, especially in datasets with a large number of features, which can lead to the "curse of dimensionality".
- The choice of initial cluster centroids can significantly affect the final clustering outcome, as K-means can converge to local minima.