K-Means

Guowei Wei
Department of Mathematics
Michigan State University

References:
Duc D. Nguyen's lecture notes
David Sontag's note
Wikipedia

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

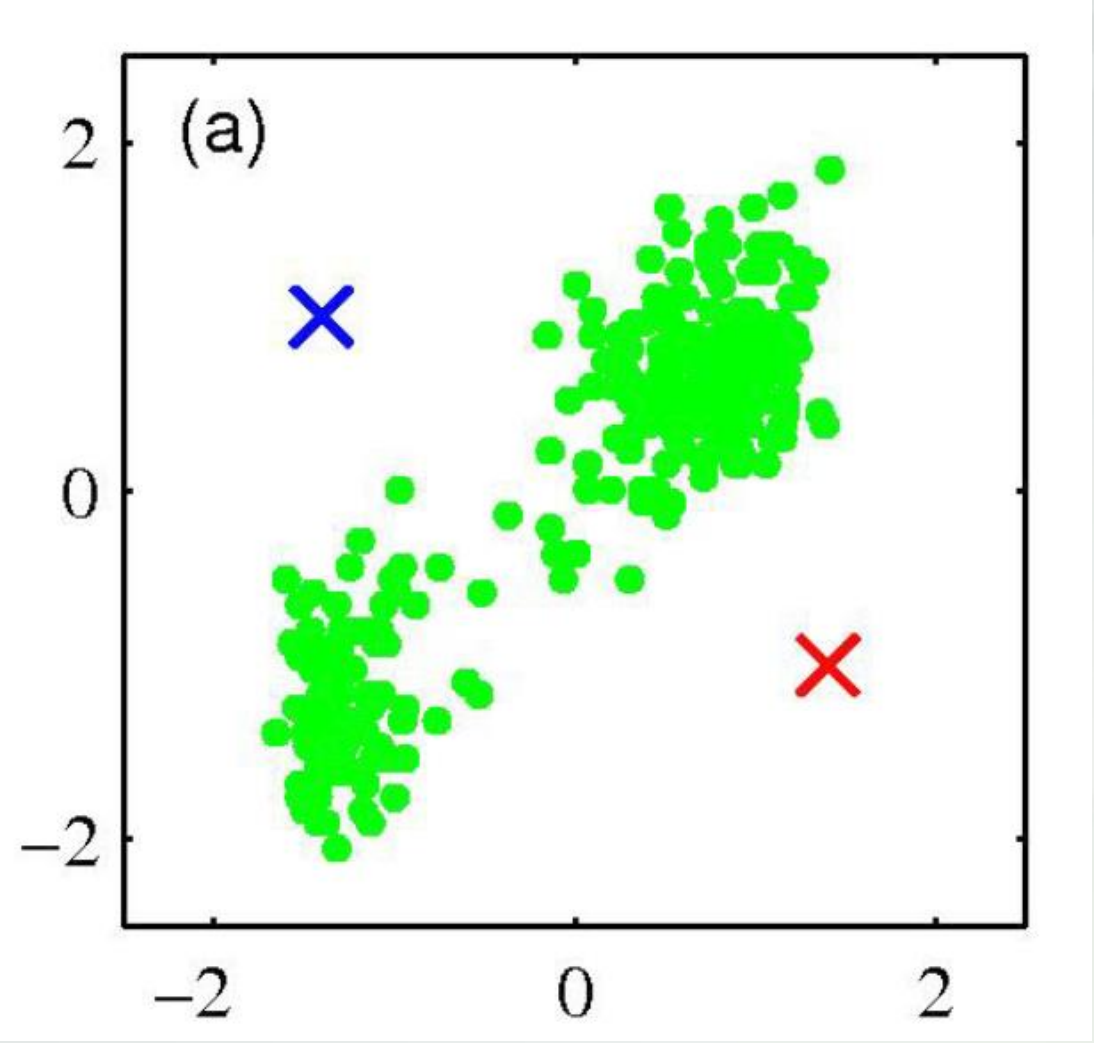
Examples

- Image segmentation



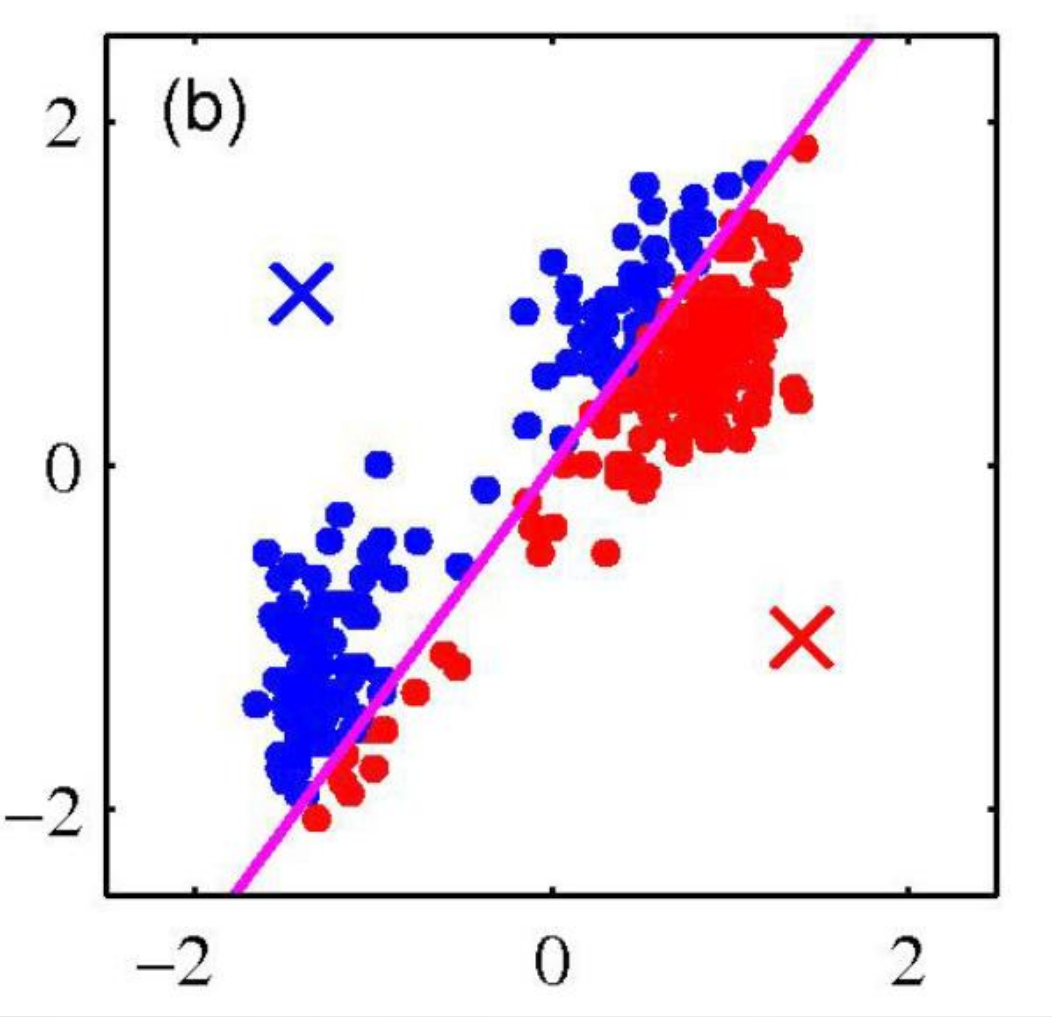

http://vladlen.info/publications/feature-space-optimization-for-semantic-video-segmentation/

Algorithm



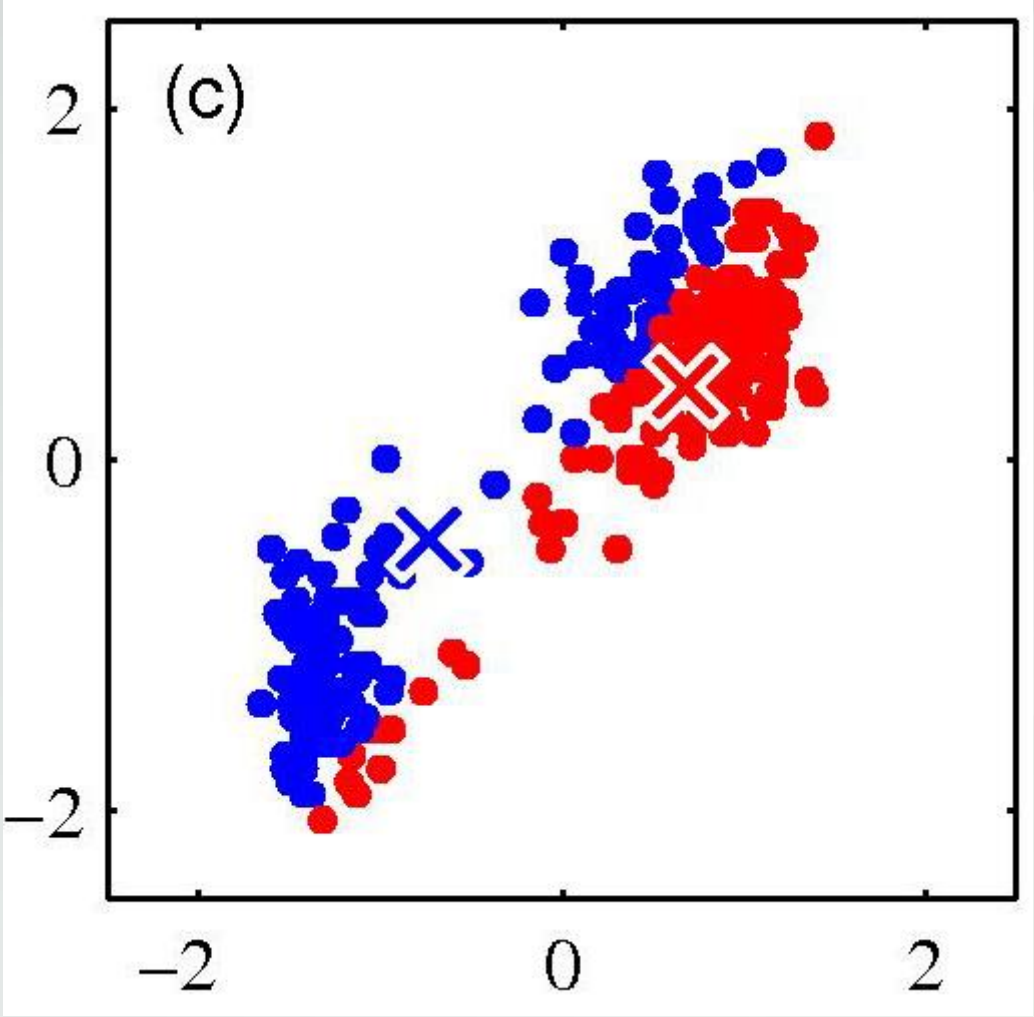
- Want to group in $k = 2$ clusters
- Pick 2 random point as cluster centers

Algorithm



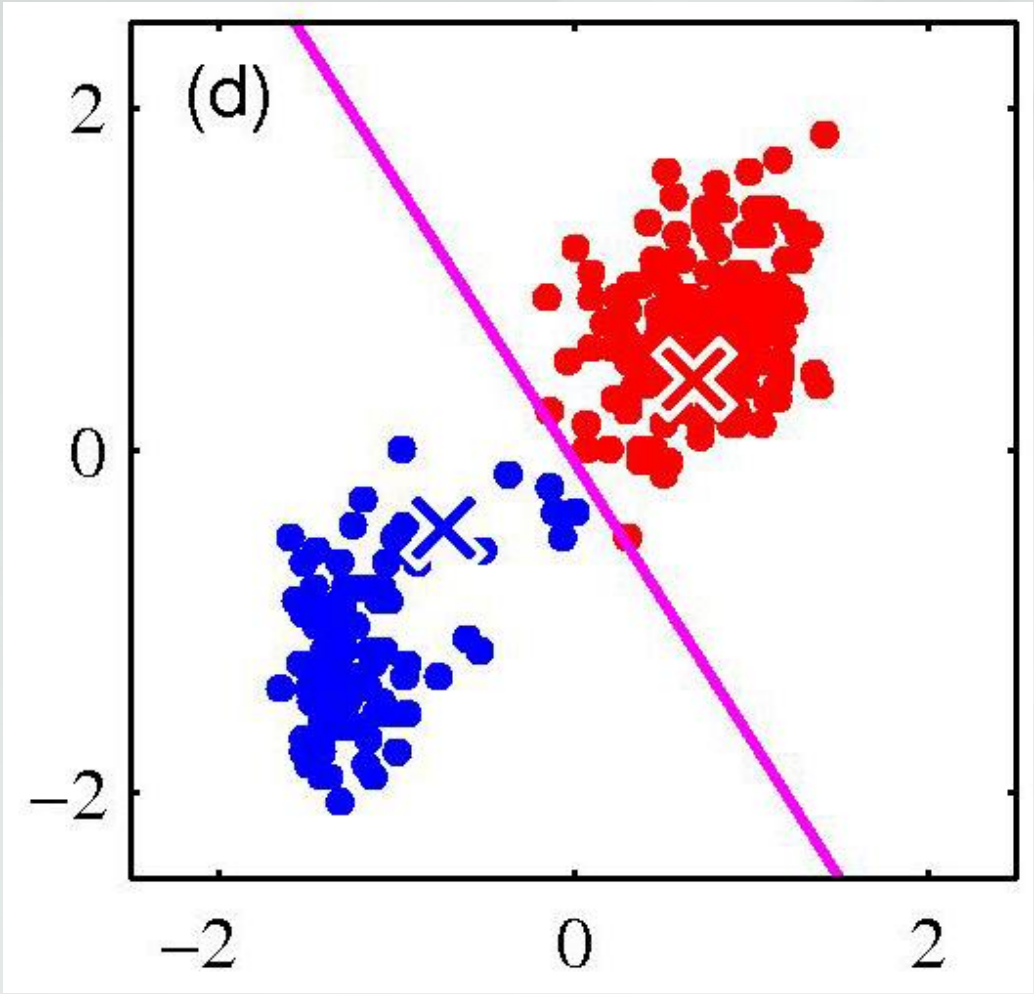
- Assign data points to closest cluster center

Algorithm



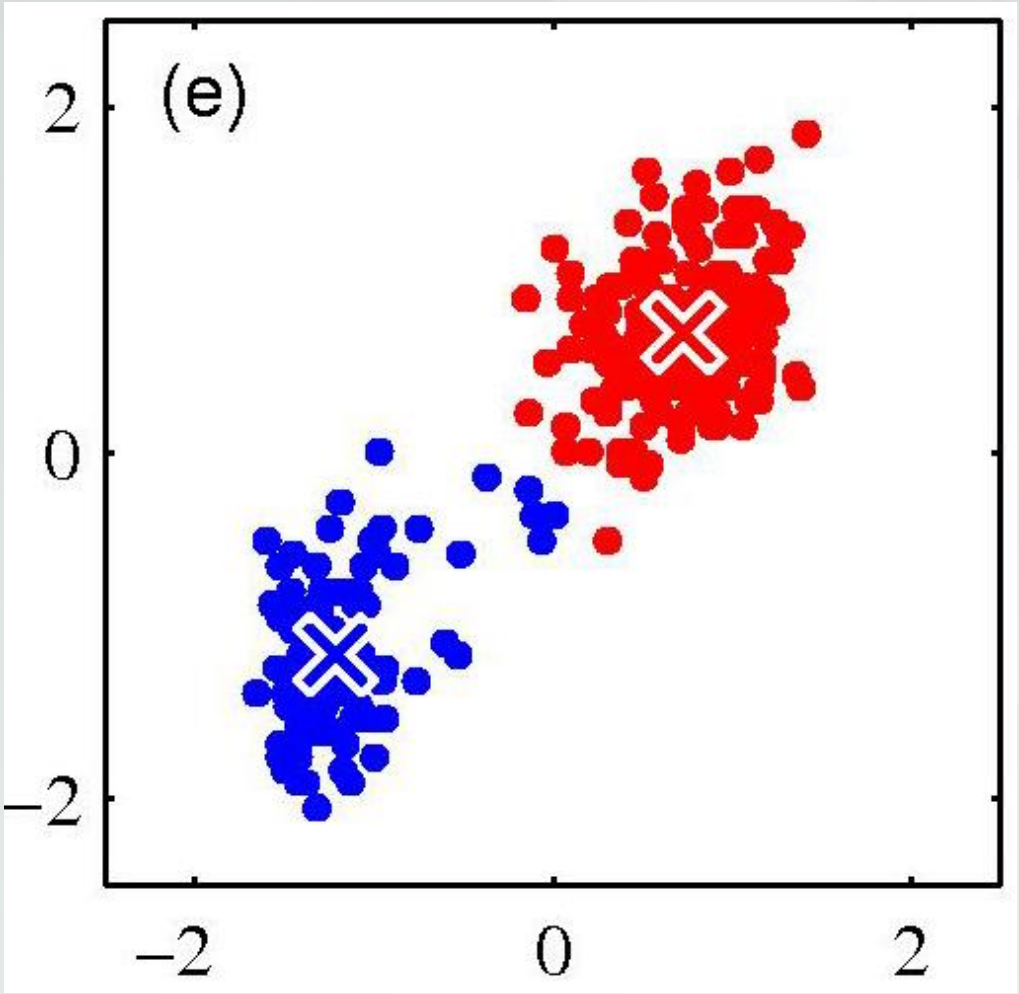
- Change cluster center to the average of the assigned data points

Algorithm



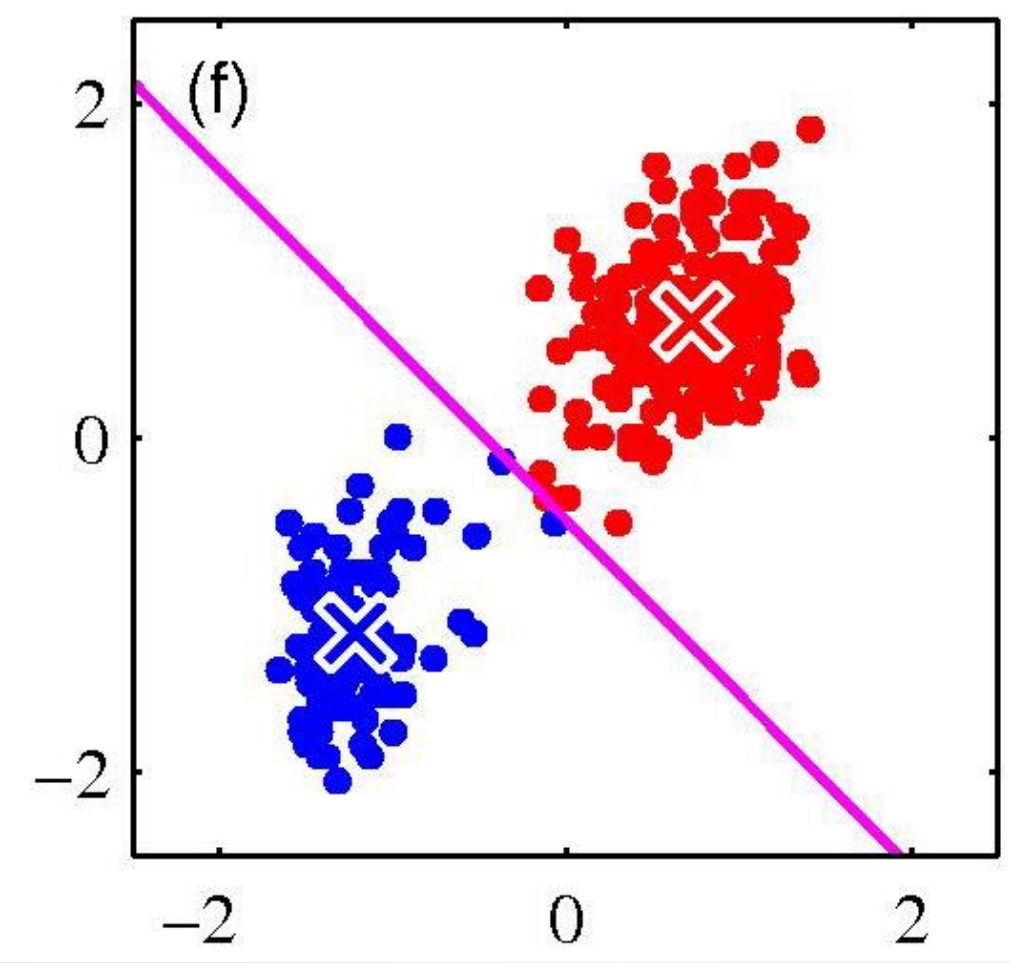
- Repeat until no change in the cluster center

Algorithm



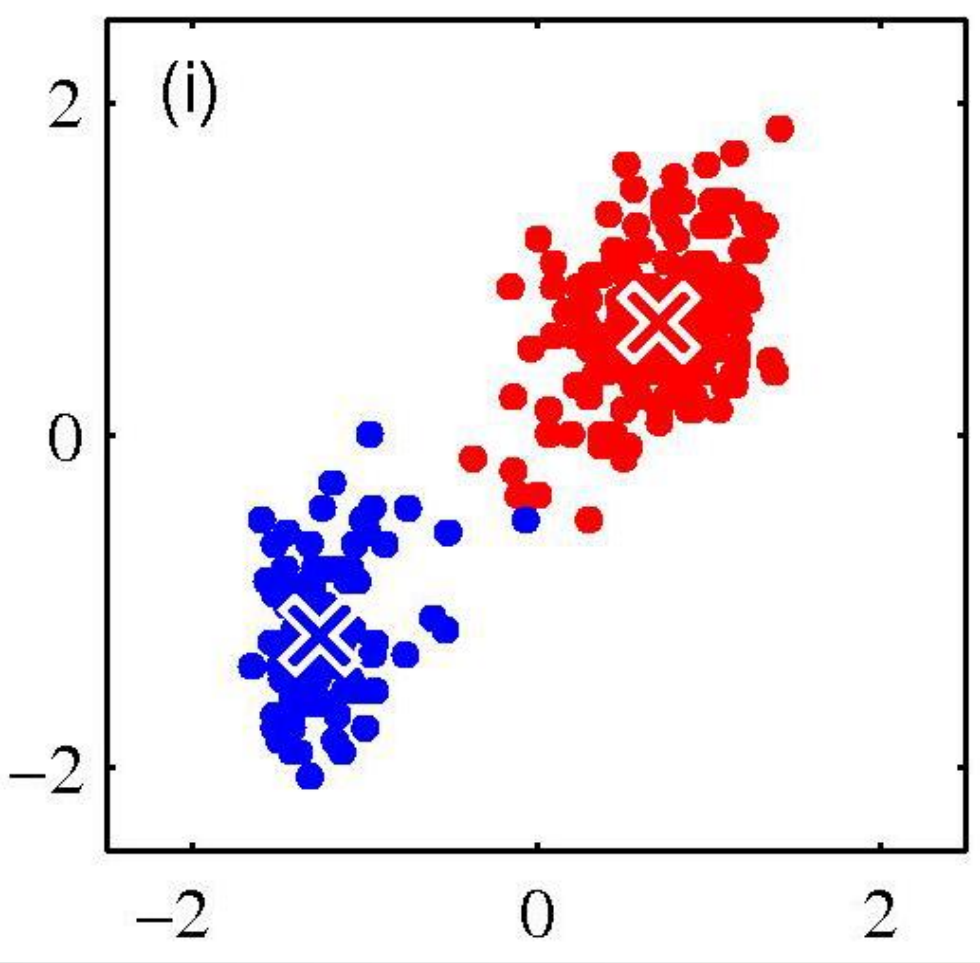
- Repeat until no change in the cluster center

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

$$\operatorname{Arg}\min_{\boldsymbol{S}} \sum_{j=1}^{k} \sum_{\boldsymbol{x} \in S_j} \|\boldsymbol{x} - \mu_j\|^2$$

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

K-means Property

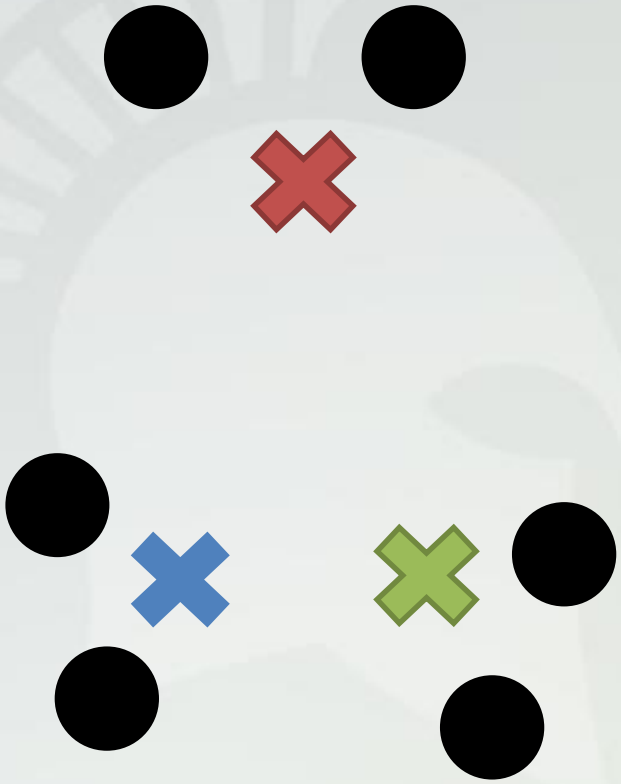
- Usefulness of k-means depend on what you pick

K-means Property

- Usefulness of k-means depend on what you pick

K-means Property

- Usefulness of k-means depend on what you pick

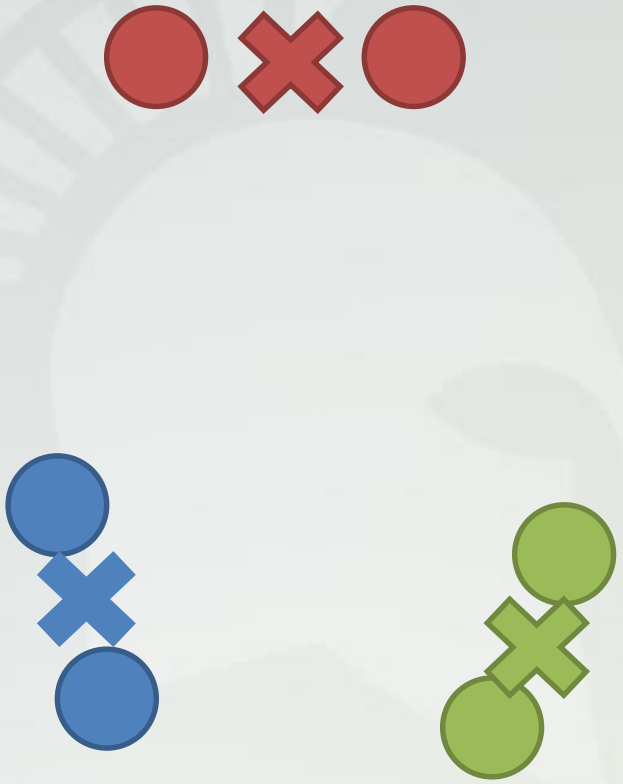


K-means Property

- Usefulness of k-means depend on what you pick

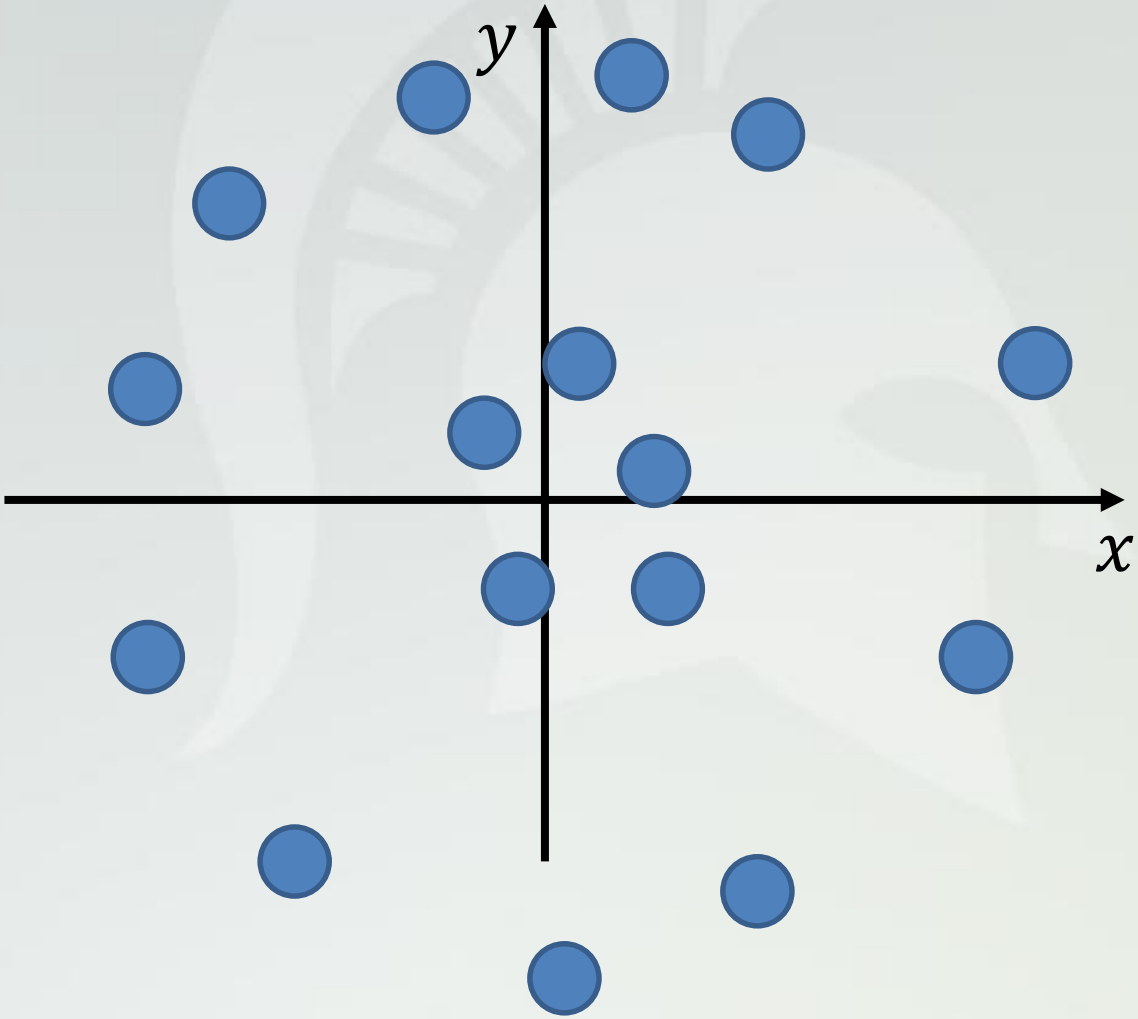
K-means Property

- Usefulness of k-means depend on what you pick



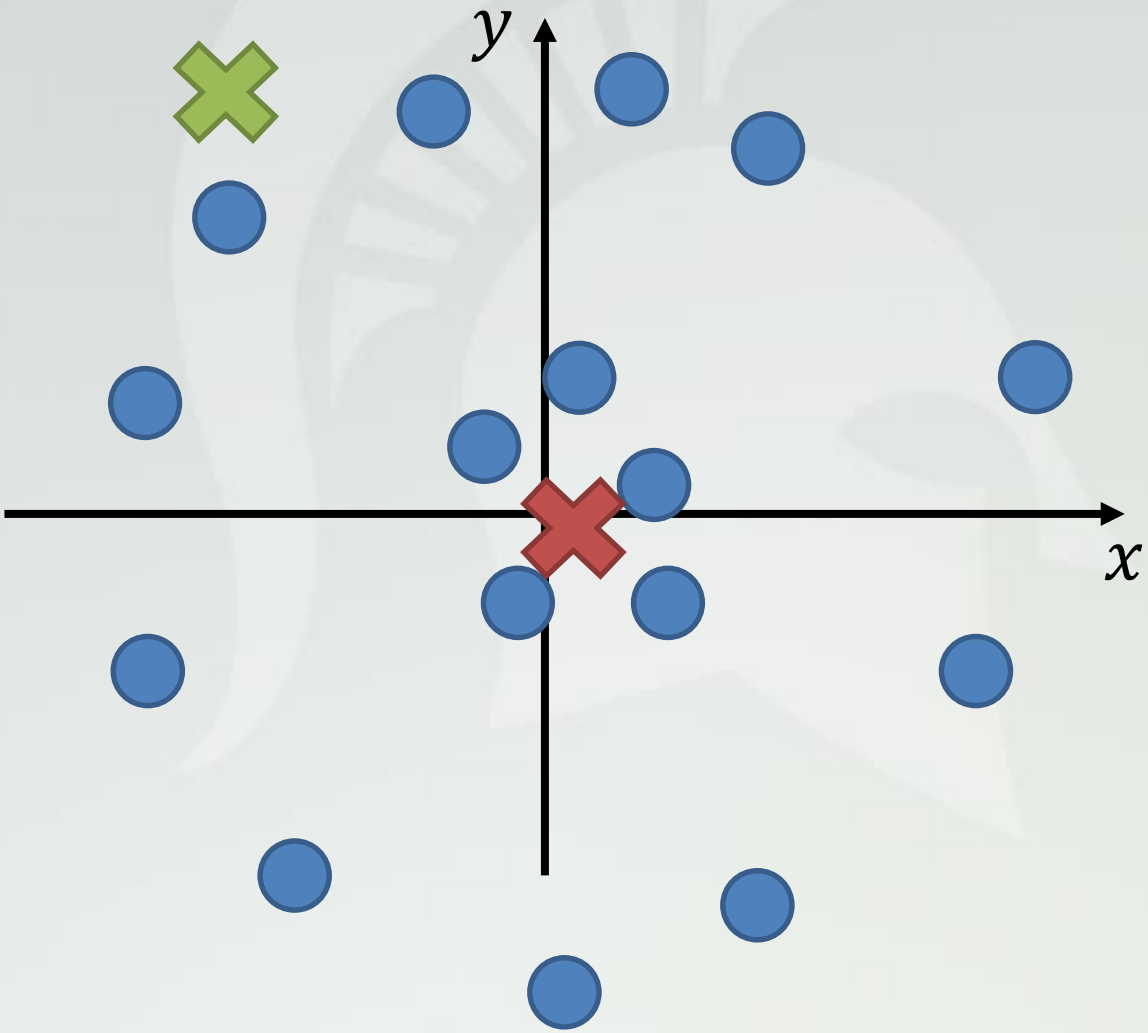
K-means Property

- Changing coordinate might be helpful



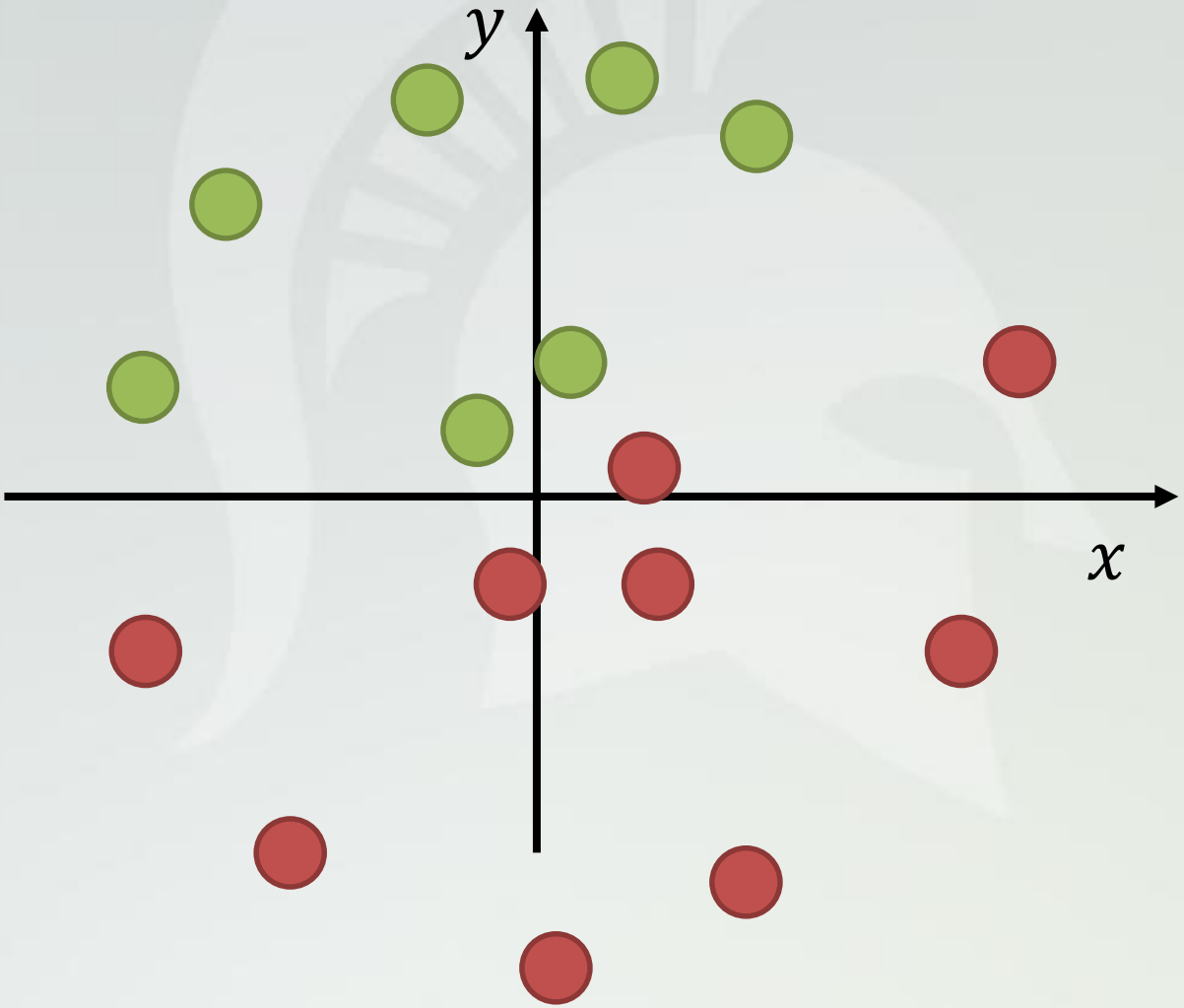
K-means Property

- Changing coordinate might be helpful



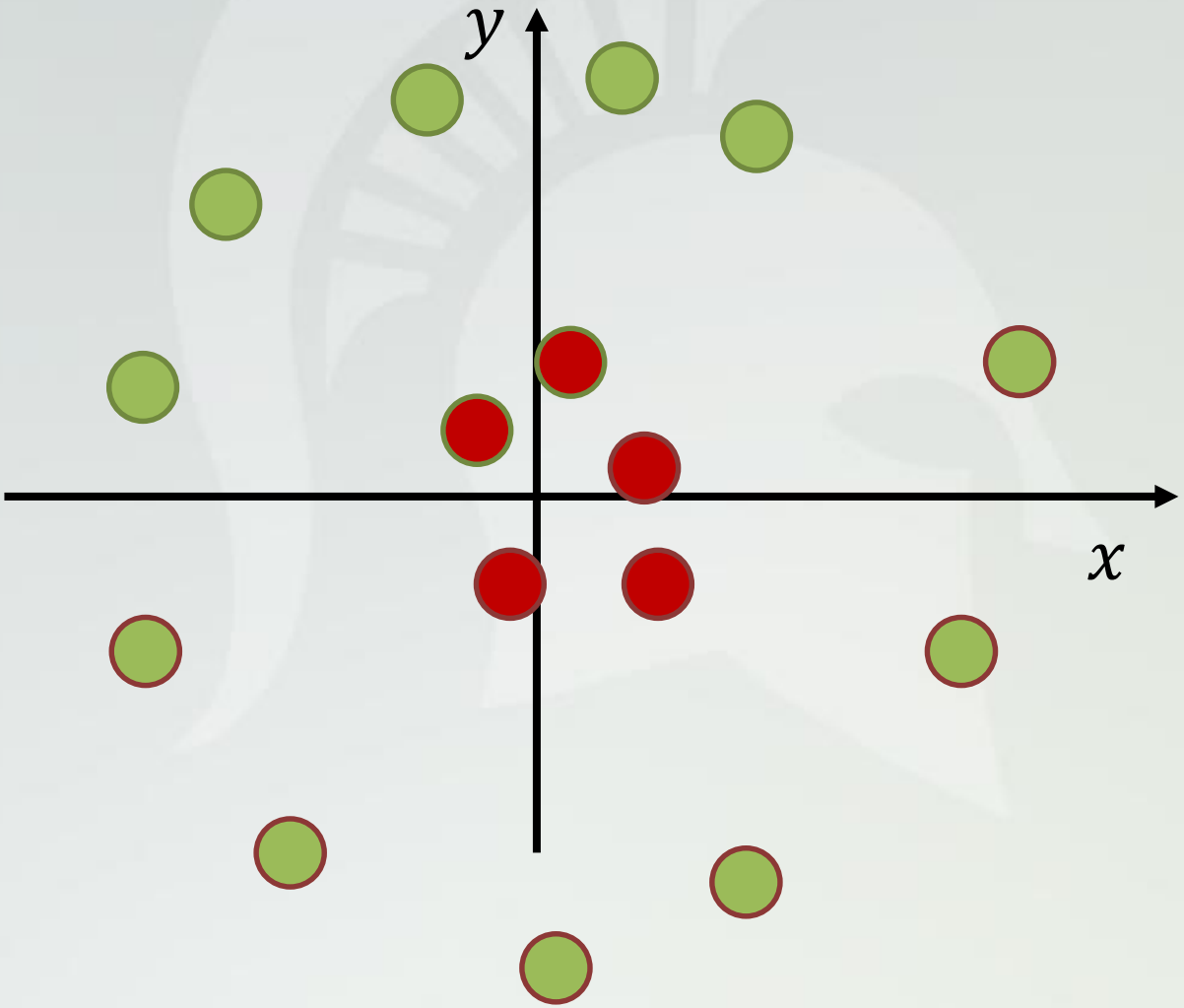
K-means Property

- Changing coordinate might be helpful



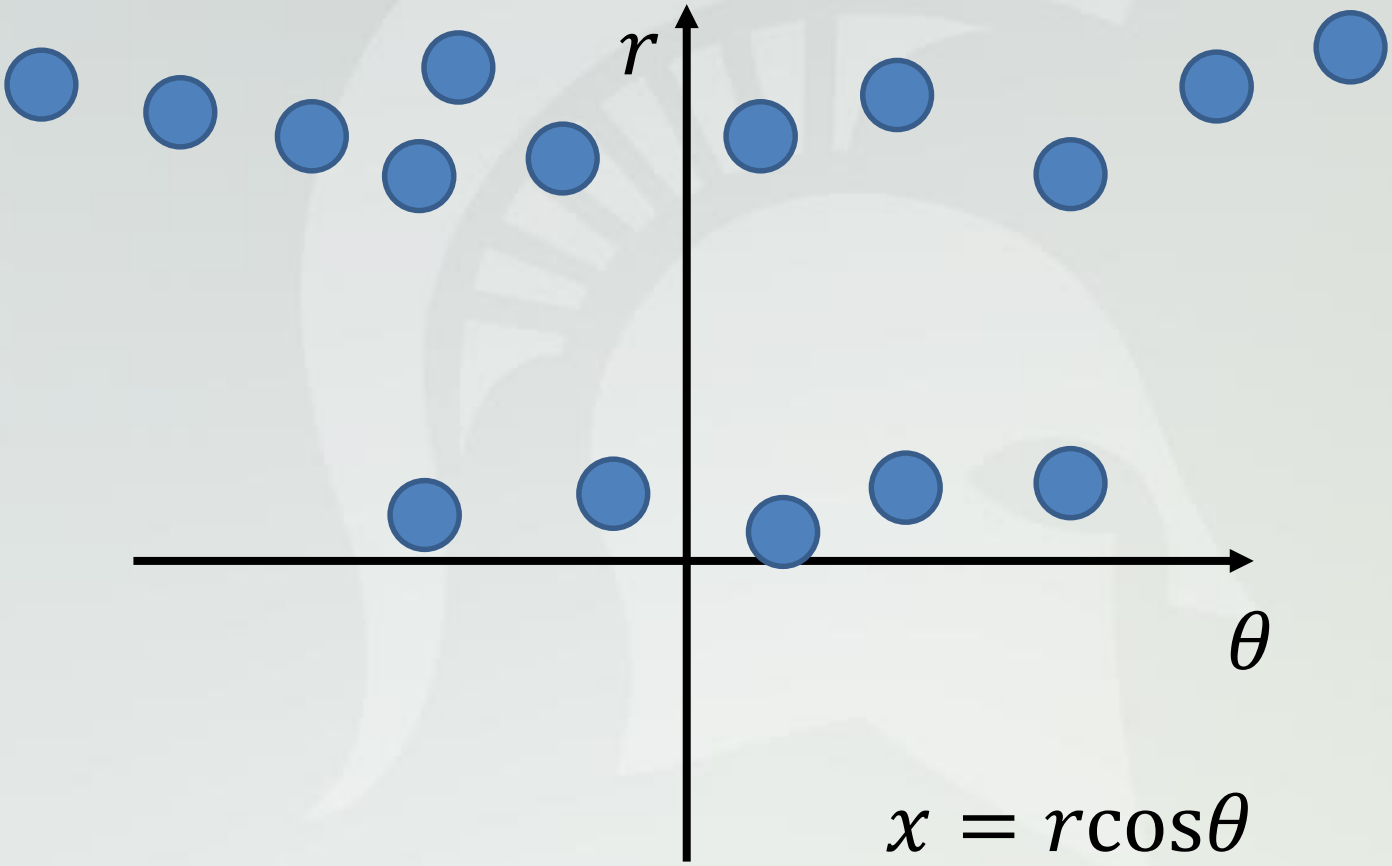
K-means Property

- Changing coordinate might be helpful



K-means Property

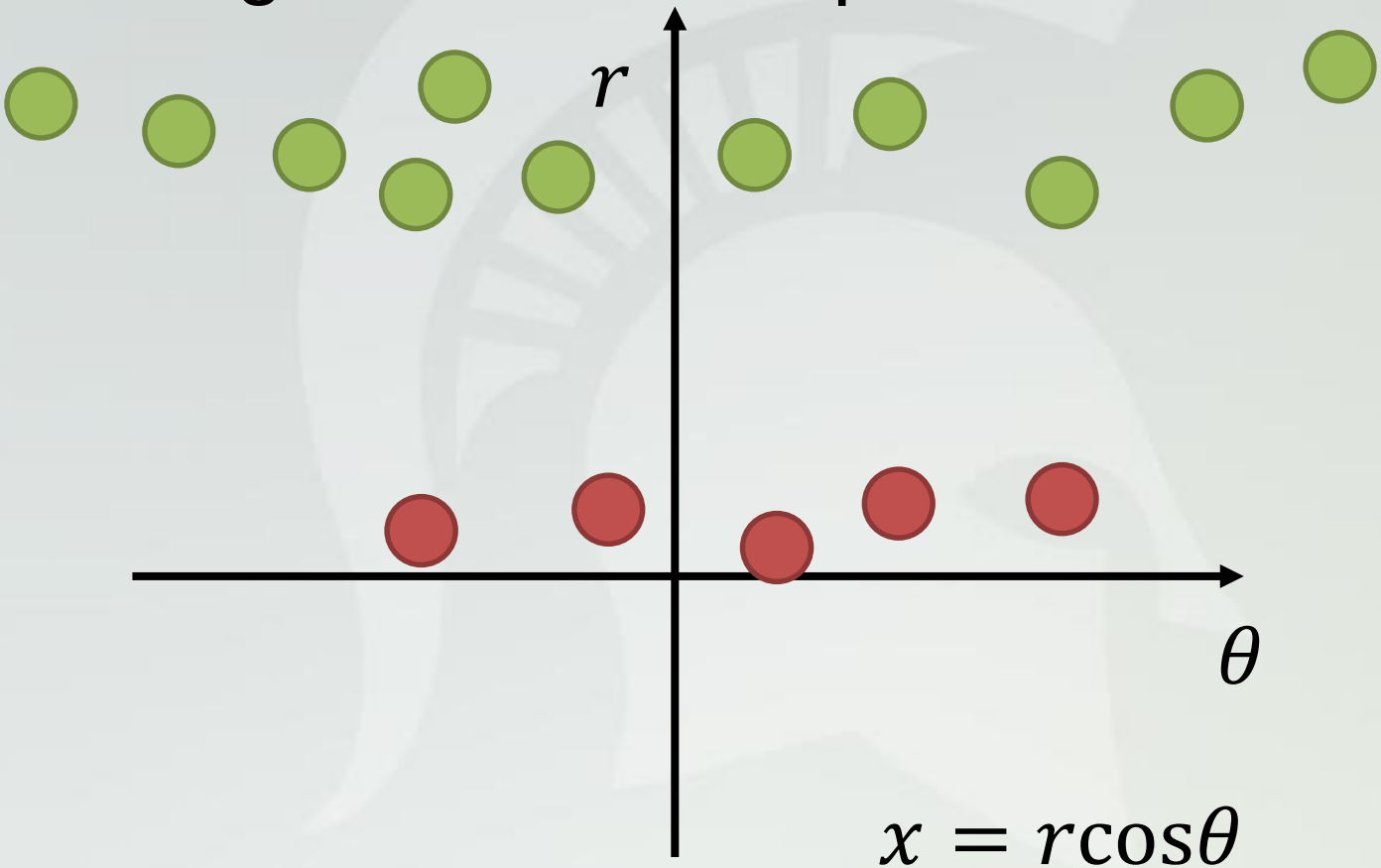
- Change Cartesian to polar coordinate



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

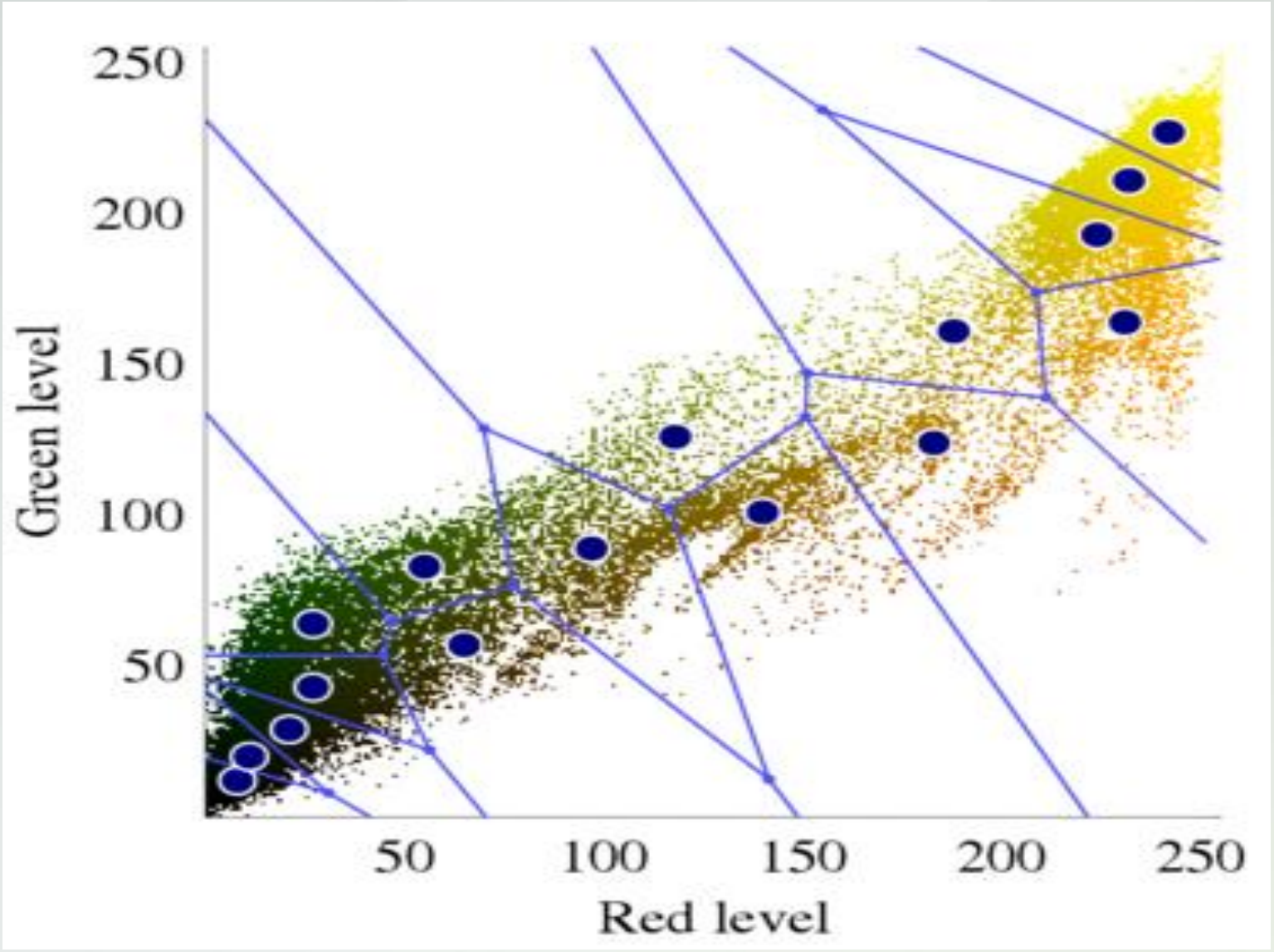
K-means Property

- Change Cartesian to polar coordinate



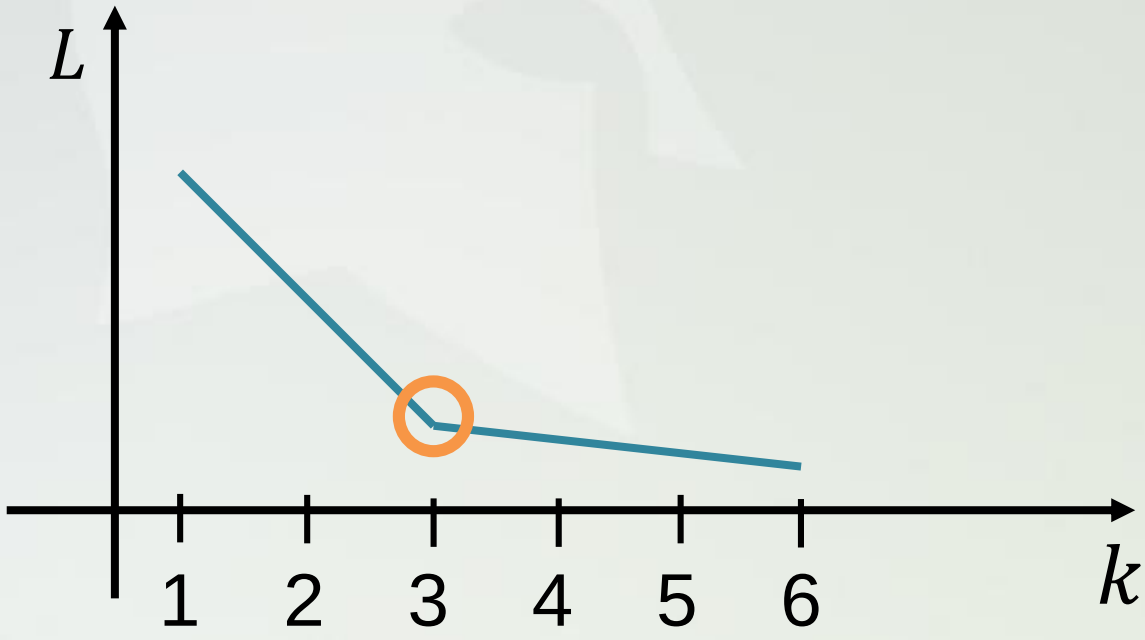
$$x = r\cos\theta$$
$$y = r\sin\theta$$

Vector quantization of colors present in the image above into Voronoi cells using *k*-means



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



Discussions

- Various metrics discussed in the K-NN can be applied and lead to various variations, such as Minkowski weighted k-means, etc.
- Complexity: NP hard in general and in R^d as $O(n^{dk+1})$
- k-means can be regarded as special case of
  1) Gaussian mixture model
  2) Principal component analysis
  3) ……