eMasters in Communication Systems

Prof. Aditya Jagannatham

Elective Module:

Advanced ML Techniques

Chapter 7

K clusters.

K Means Clustering

Clustering

- **Unsupervised learning**
- Requires data, but NO Labels.

Large amounts of Data

Linear Regression
Naive Bayes
SVM

Supervised Learning

Clustering

- **Unsupervised learning**
- Requires data, but **NO labels**

Clustering

- ***Detect patterns*** e.g. in
- **Group emails** or search results
- Customer **shopping patterns**
- **Regions** of images

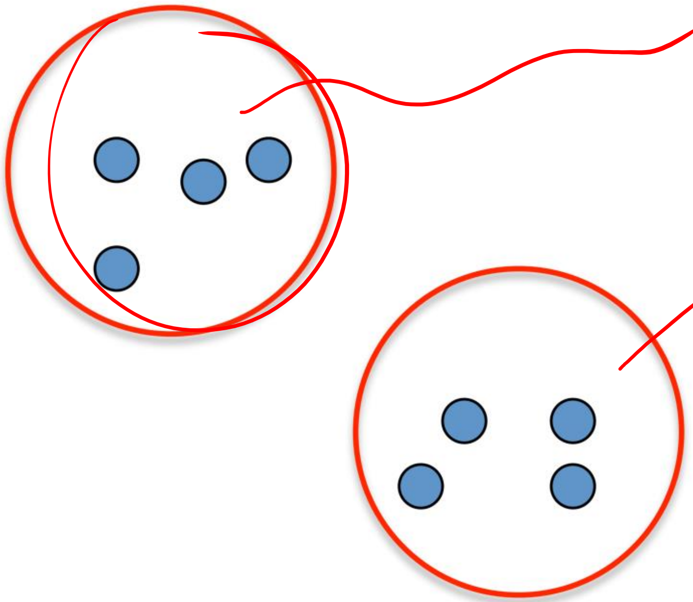
image processing

Clustering

- Basic idea: Group together similar instances.
- Example: 2D ***point patterns***

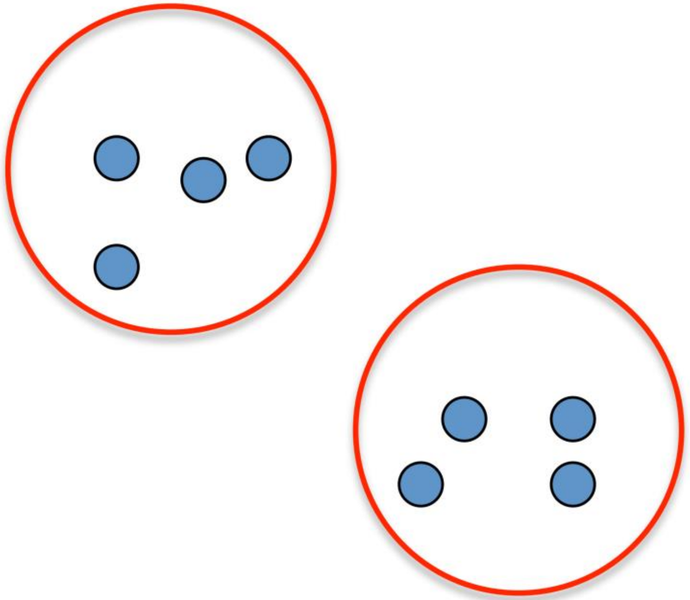
cluster 1

cluster 2



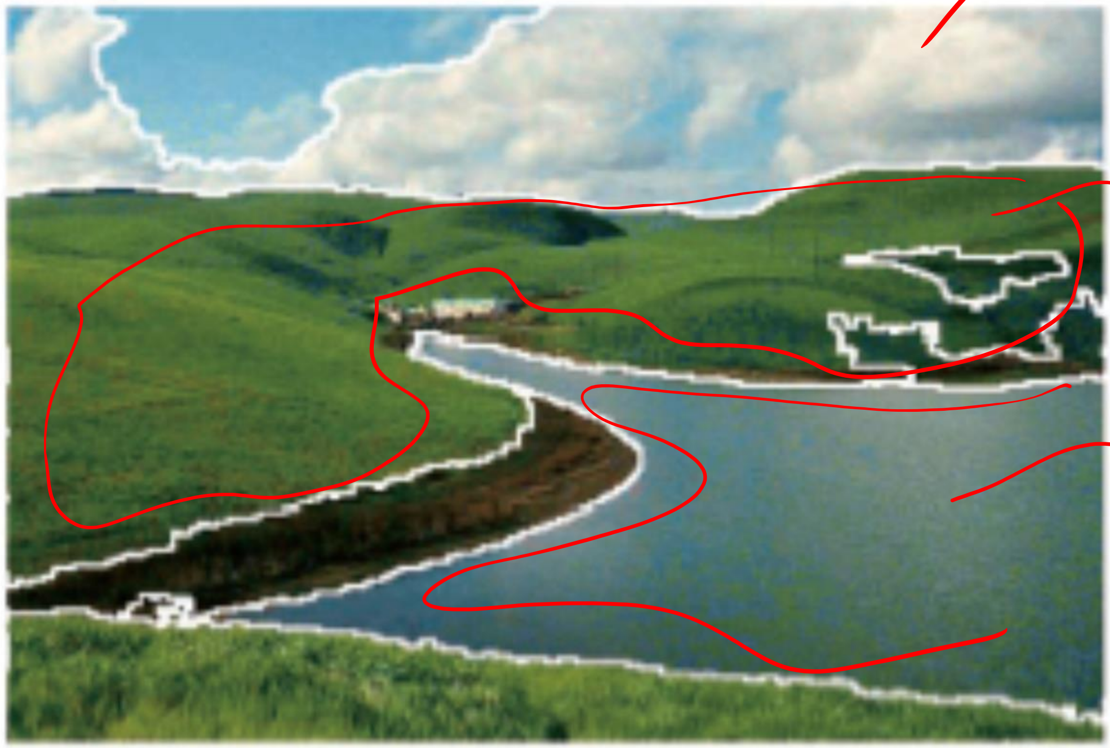
Clustering

- Basic idea: Group together ***similar instances***
- Example: 2D ***point patterns***



Clustering

- **Image segmentation**
  - Goal: Partition an image into perceptually similar regions



grassland

Waterbody

Clustering

- **Image segmentation**
  - Goal: Partition an image into **perceptually similar** regions

K-Means Algorithm

- K −means is an **iterative algorithm**
- Consider the ***dataset*** of n −dimensional vectors

M vectors.

$$\underbrace{\bar{x}(1), \bar{x}(2), \ldots, \bar{x}(M)}_{\text{Unlabeled Dataset.}}$$

Divides data set
K clusters.

K-Means Algorithm

- K −means is an **iterative algorithm**
- Consider the ***dataset*** of n −dimensional vectors

$$\bar{\mathbf{x}}(1), \bar{\mathbf{x}}(2), \dots, \bar{\mathbf{x}}(M)$$

Clusters

- Organize the data into K ***clusters***

K clusters.

$\mathcal{C}_1$

$\mathcal{C}_2$

$\mathcal{C}_3$

$\mathcal{C}_K$ ......

Clusters

- Organize the data into K **clusters**

$$\mathcal{C}_1, \mathcal{C}_2, \dots, \mathcal{C}_K$$

Belongs to a ~~single~~ cluster.

$\bar{x} \in \mathcal{C}_i$: denotes i^{th} cluster.

Set $\{ \dots \}$.

Clusters

- The **centroids** for the clusters are

$$\underbrace{\bar{\mu}_1, \bar{\mu}_2, \ldots, \bar{\mu}_k}_{\text{Centroids..}}$$

$\bar{\mu}_1 \times$ centroid

centroid of cluster 2

$\bar{\mu}_2 \times$

Clusters

- The **centroids** for the clusters are

$$\bar{\boldsymbol{\mu}}_1, \bar{\boldsymbol{\mu}}_2, \dots, \bar{\boldsymbol{\mu}}_K$$

Cluster assignment

- Let $\alpha_i(j)$ denote the ***cluster assignment indicator***

$\alpha_2(1) = 1 \Rightarrow \bar{x}(1) \in \mathcal{C}_2$.

$$\alpha_i(j) = \begin{cases} 1 & \text{if } \bar{x}(j) \in \mathcal{C}_i \\ 0 & \text{if } \bar{x}(j) \notin \mathcal{C}_i \end{cases}$$

$\alpha_i(j) = 1$ if jth point Belongs to ith cluster

Else : 0

For any j, $\alpha_i(j) = 1$ for only one i

Any point can belong to only a single cluster!

$\alpha_i(j) \in \{0, 1\}$ Discrete Variable.

$$\sum_{i=1}^{K} \alpha_i(j) = 1$$

Cluster assignment

- Let $\alpha_i(j)$ denote the ***cluster assignment indicator***

j^{th} point belongs to i^{th} cluster

$$\alpha_i(j) = \begin{cases} 1 & \bar{\mathbf{x}}(j) \in C_i \\ 0 & \bar{\mathbf{x}}(j) \notin C_i \end{cases}$$

j^{th} point does NOT belong to i^{th} cluster

Cost function

- The K −means **cost-function** to minimize is given as

Sum of squares of distances of all points to respective centroids.

$$\min \sum_{j=1}^{M} \sum_{i=1}^{K} \alpha_i(j) \|\bar{x}(j) - \bar{\mu}_i\|^2$$

Square of distance of jth point to its centroid.

Cost function

- The K —means **cost-function** to minimize is given as

$$\underbrace{\min \sum_{i=1}^{K} \sum_{j=1}^{M} \alpha_i(j) \|\bar{\mathbf{x}}(j) - \bar{\boldsymbol{\mu}}_i\|^2}_{\text{K means cost.}}$$

K-Means procedure

- ***Initialize centroids*** randomly
- $\overline{\boldsymbol{\mu}}_i^{(l-1)}$ denotes **centroid** in iteration $l-1$

iteration

$\bar{\mu}_1^{(0)}, \bar{\mu}_2^{(0)}, \ldots, \bar{\mu}_K^{(0)}$

iterative

$\bar{\mu}_i^{(0)}$: Denotes i^{th} centroid in iteration 0

K-Means procedure

- ***Initialize centroids*** randomly

$$\bar{\boldsymbol{\mu}}_1^{(0)}, \bar{\boldsymbol{\mu}}_2^{(0)}, \ldots, \bar{\boldsymbol{\mu}}_K^{(0)}$$

iteration 0

- $\bar{\boldsymbol{\mu}}_i^{(l-1)}$ denotes **centroid** in iteration $l-1$

$\alpha_i^{(l)}(j)$: cluster assignment in iteration l

$\bar{\mu}_i^{(l)}$: update centroid in iteration l.

K-Means procedure

- In **iteration** l, for each point $\bar{\mathbf{x}}(j)$, perform

$$\min . \underbrace{\sum_{i=1}^{K} \alpha_i(j) \left\| \bar{x}(j) - \bar{\mu}_i^{(l-1)} \right\|^2}_{\text{minimize}}$$

Find $\alpha_i(j)$

K-Means procedure

Assign $\bar{x}(j)$ to cluster $\tilde{i}$ with closest centroid.

- In **iteration** l, for each point $\bar{\mathbf{x}}(j)$, perform

$$\min \sum_{i=1}^{K} \alpha_i(j) \left\| \bar{\mathbf{x}}(j) - \bar{\boldsymbol{\mu}}_i^{(l-1)} \right\|^2$$

$\alpha_i(j) = 1$ for only one value of i

Cluster determination

- This is ***minimized*** when $\alpha_{\tilde{i}}^{(l)}(j) = 1$, where

$$\underbrace{\tilde{i} = \arg\min_{i} \left\| \bar{x}(j) - \bar{\mu}_i^{(l-1)} \right\|^2}_{\text{centroid for which square of distance is minimum.}}$$

Cluster determination

- This is ***minimized*** when $\alpha_{\tilde{\imath}}^{(l)}(j) = 1$, where

$$\tilde{\imath} = \arg\min \left\| \bar{\mathbf{x}}(j) - \bar{\boldsymbol{\mu}}_i^{(l-1)} \right\|^2$$

Cluster determination

- i.e. assign $\bar{\mathbf{x}}(j)$ to the *closest* *centroid* $\overline{\boldsymbol{\mu}}_{\tilde{i}}^{(l-1)}$

$$\alpha_i^{(l)}(j) = \begin{cases} 1 & \text{if } i = \tilde{i} \\ 0 & \text{if } i \neq \tilde{i} \end{cases}$$

Cluster determination

DO this for each j
$j = 1, 2, \ldots, M$.

- i.e. assign $\bar{\mathbf{x}}(j)$ to the closest centroid $\bar{\boldsymbol{\mu}}_{\tilde{i}}^{(l-1)}$

iteration l

$$\alpha_i^{(l)}(j) = \begin{cases} 1 & i = \tilde{i} \\ 0 & i \neq \tilde{i} \end{cases}$$

cluster assignment indicator in iteration l.

Centroid determination

in iteration ℓ.

- Next ***determine the centroids*** for the given clusters
- For this, in ***each cluster*** i, minimize

minimize cost function

$$\sum_{j=1}^{M} \alpha_i^{(\ell)}(j) \left\| \bar{x}(j) - \bar{\mu}_i \right\|^2$$

Determine $\bar{\mu}$

Centroid determination

- Next ***determine the centroids*** for the given clusters
- For this, in ***each cluster*** i, minimize

$$\sum_{j=1}^{M} \alpha_i^{(l)}(j)\|\bar{\mathbf{x}}(j) - \bar{\boldsymbol{\mu}}_i\|^2$$

Centroid determination

- This can be ***expanded*** as

$$\|\bar{a}\|^2 = \bar{a}^T \bar{a}$$

$$\bar{a}^T \bar{b} = \bar{b}^T \bar{a} = \sum_k a_k b_k$$

$$\sum_{j=1}^{M} \alpha_i^{(l)}(j) \, \|\bar{x}(j) - \bar{\mu}_i\|^2$$

$$= \sum_{j=1}^{M} \alpha_i^{(l)}(j) \cdot (\bar{x}(j) - \bar{\mu}_i)^T (\bar{x}(j) - \bar{\mu}_i)$$

cluster assignments are known

Determine $\mu(i)$ for which cost is minimum.

$$= \sum_{j=1}^{M} \alpha_i^{(l)}(j) \left(\bar{x}^T(j)\bar{x}(j) - 2\mu^T \bar{x}(j) + \bar{\mu}_i^T \bar{\mu}_i \right)$$

Centroid determination

- This can be **expanded** as

$$\sum_{j=1}^{M} \alpha_i^{(l)}(j)\|\bar{\mathbf{x}}(j)-\bar{\boldsymbol{\mu}}_i\|^2$$

Take gradient and set equal to zero.

$$= \sum_{j=1}^{M} \alpha_i^{(l)}(j)\left(\bar{\mathbf{x}}^T(j)\bar{\mathbf{x}}(j)+\bar{\boldsymbol{\mu}}_i^T\bar{\boldsymbol{\mu}}_i-2\bar{\mathbf{x}}^T(j)\bar{\boldsymbol{\mu}}_i\right)$$

$$\nabla_{\bar{\mu}} f = \begin{bmatrix} \partial f/\partial \mu_1 \\ \partial f/\partial \mu_2 \\ \vdots \end{bmatrix}$$

Centroid determination

- Taking the ***gradient*** and ***setting to zero*** yields

TO minimize

$$\sum_{i=1}^{M} \alpha_i^{(l)}(j)\left(0 + 2\bar{\mu}_i - 2\bar{x}(j)\right) = 0$$

$$\Rightarrow \bar{\mu}_i^{(l)} = \frac{\sum_{i=1}^{M} \alpha_i^{(l)}(j)\,\bar{x}(j)}{\sum_{i=1}^{M} \alpha_i^{(l)}(j)} = \frac{\text{Sum of all points in cluster } i}{\text{Total number of points in cluster } i}$$

Centroid determination

- Taking the ***gradient*** and ***setting to zero*** yields

Centroid of cluster i in iteration l.

$$\bar{\boldsymbol{\mu}}_i^{(l)} = \frac{\sum_{j=1}^{M} \alpha_i^{(l)}(j)\bar{\mathbf{x}}(j)}{\sum_{j=1}^{M} \alpha_i^{(l)}(j)}$$

$\bar{x}(j) \in C_i$, point j lies in cluster i

$$= \frac{\sum_{j:\bar{\mathbf{x}}(j)\in C_i} \bar{\mathbf{x}}(j)}{\sum_{j:\bar{\mathbf{x}}(j)\in C_i} 1}$$

Mean of all points in cluster i

Centroid determination

- i.e. **average** of all points assigned to cluster i in iteration l

Stopping criterion

- **Stopping criterion**: Stop when clusters are **stable**
  - i.e., when cluster assignments **do NOT change**

Centroids also do NOT change.

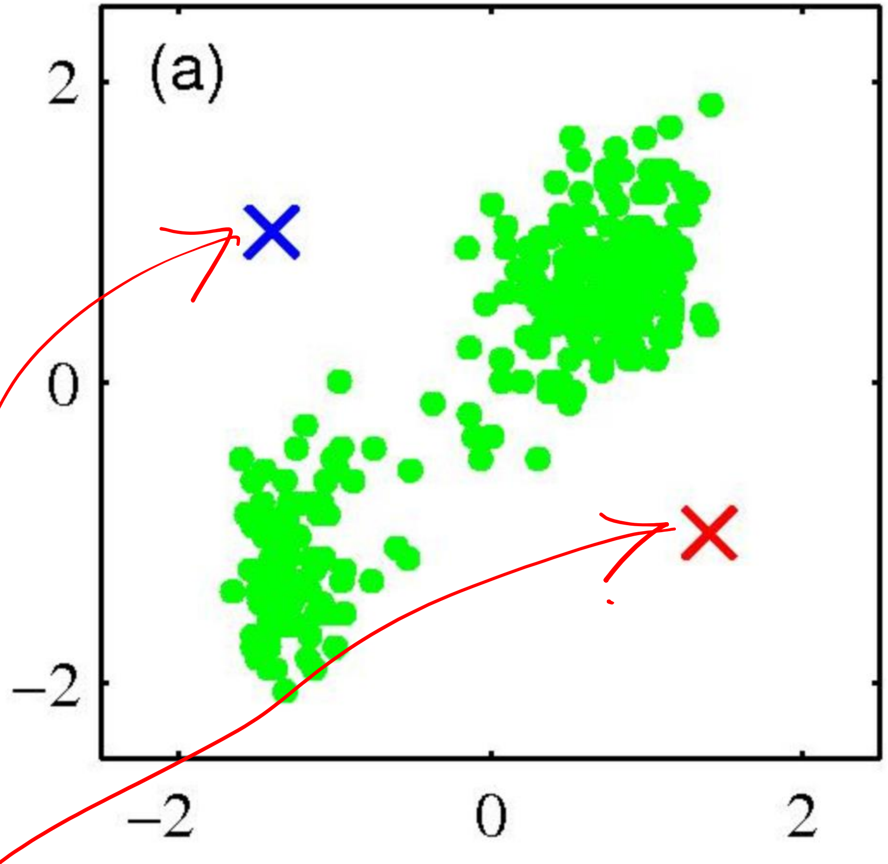
K-Means Example

clusters

- Pick K ***random points*** as cluster ***centroids***
- Shown here for $K = 2$

iteration 0

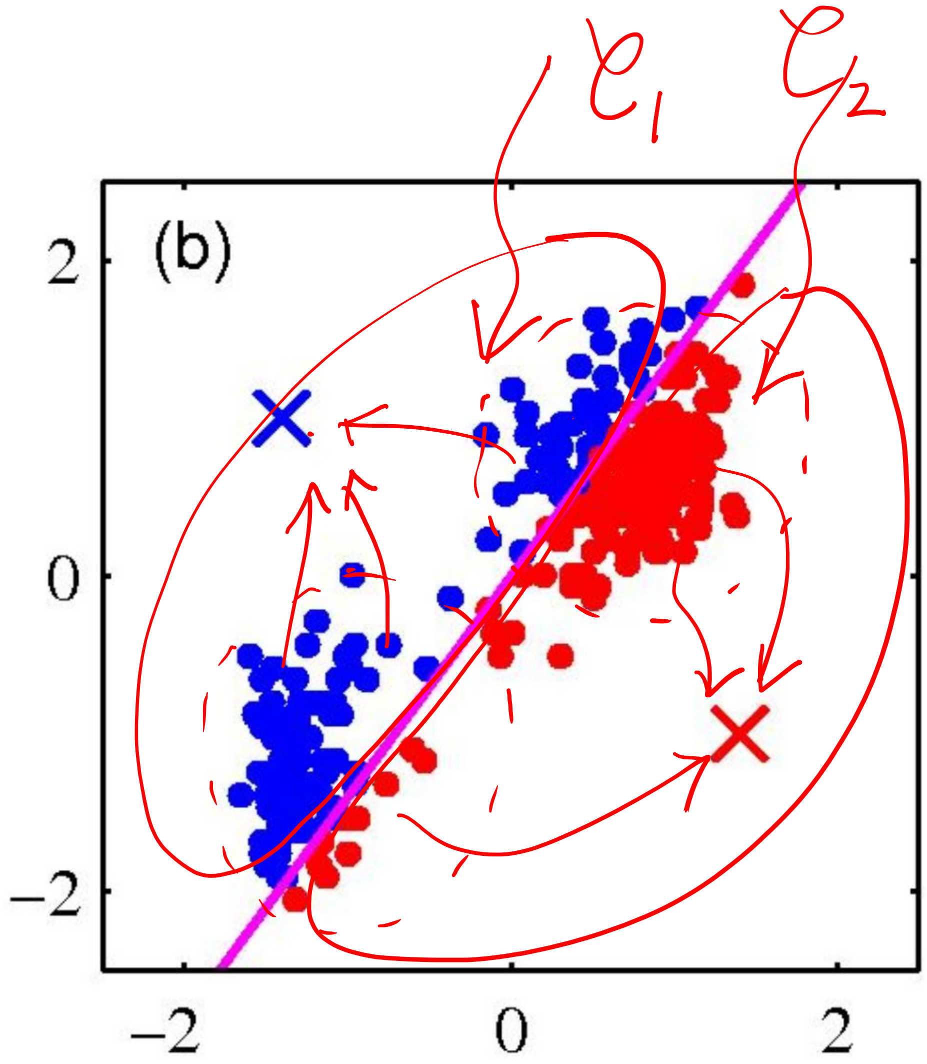
$\mu_1^{(0)}$



$\mu_2^{(0)}$

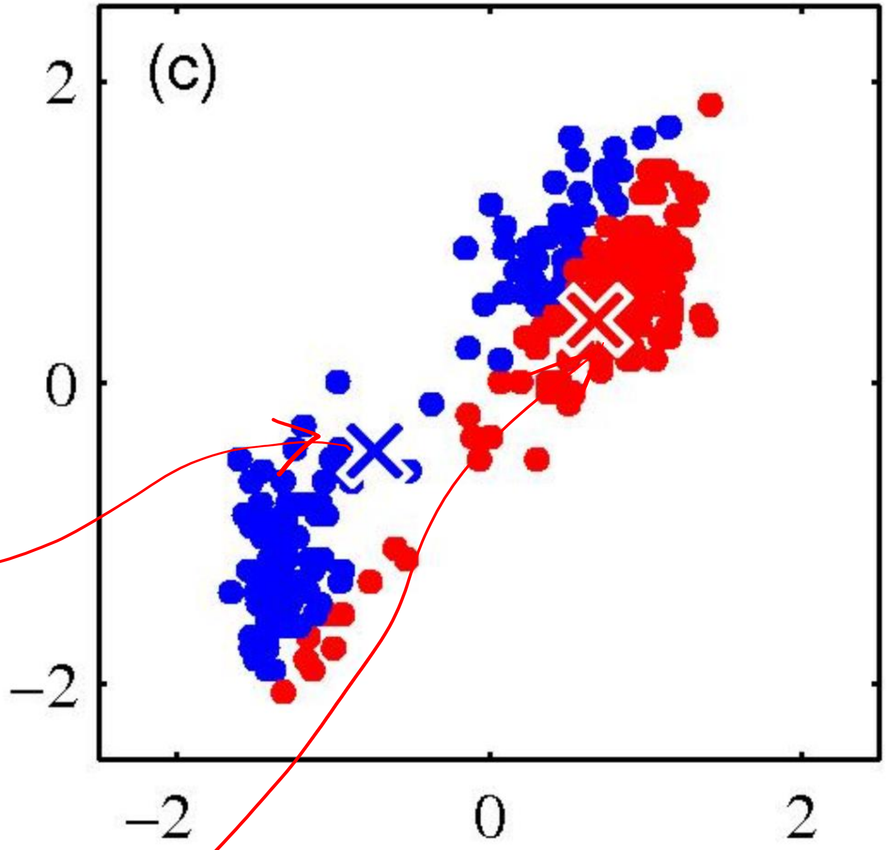
Iteration 1

- Assign data points to ***closest centroid***



Iteration 1

- Change each ***centroid*** to the **average** of the assigned points



Centroid of cluster 2 in iteration 1

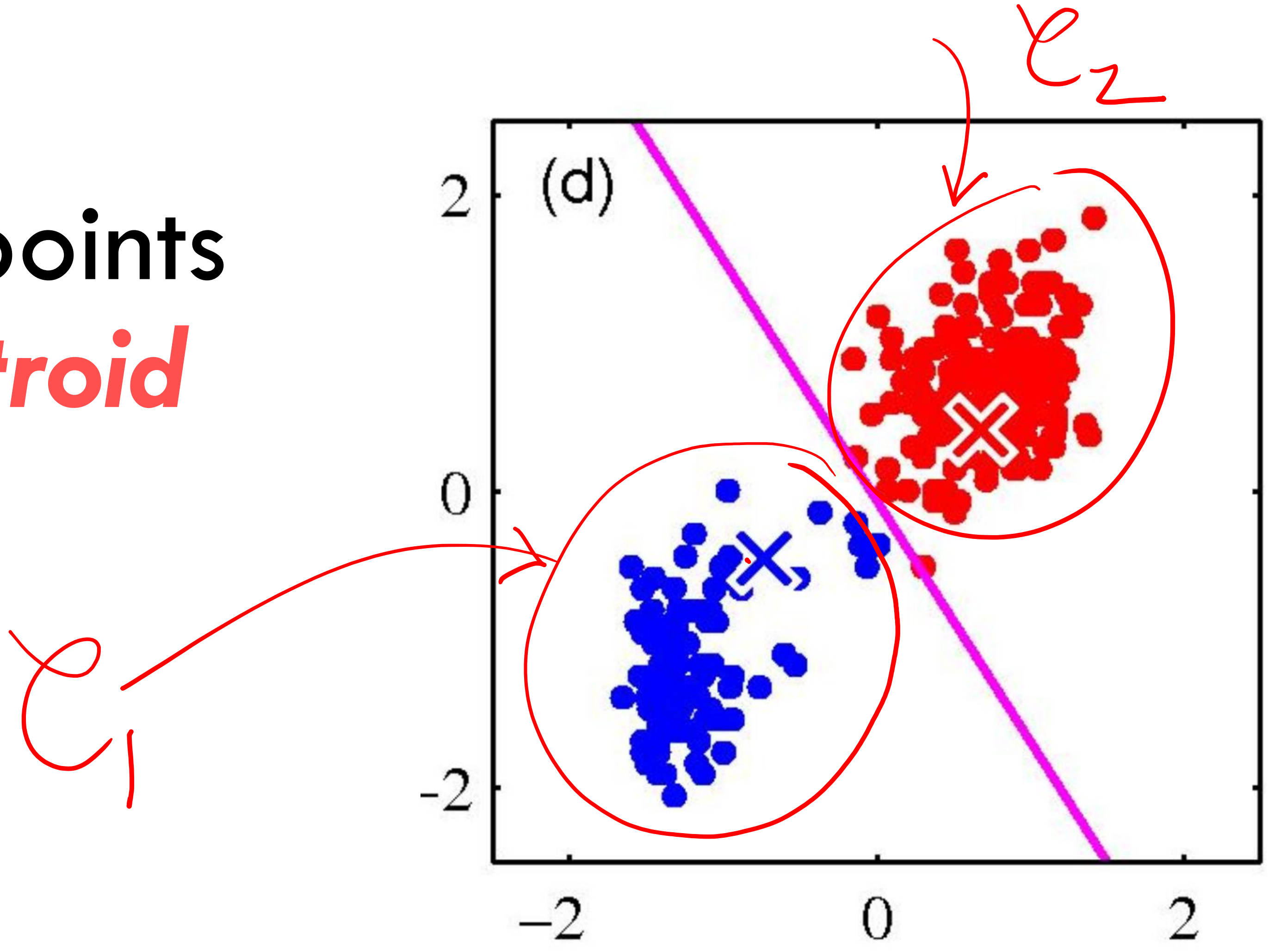
Centroid of cluster 1 in iteration 1

$\bar{\mu}_1^{(1)}$

$\bar{\mu}_2^{(1)}$

Iteration 2

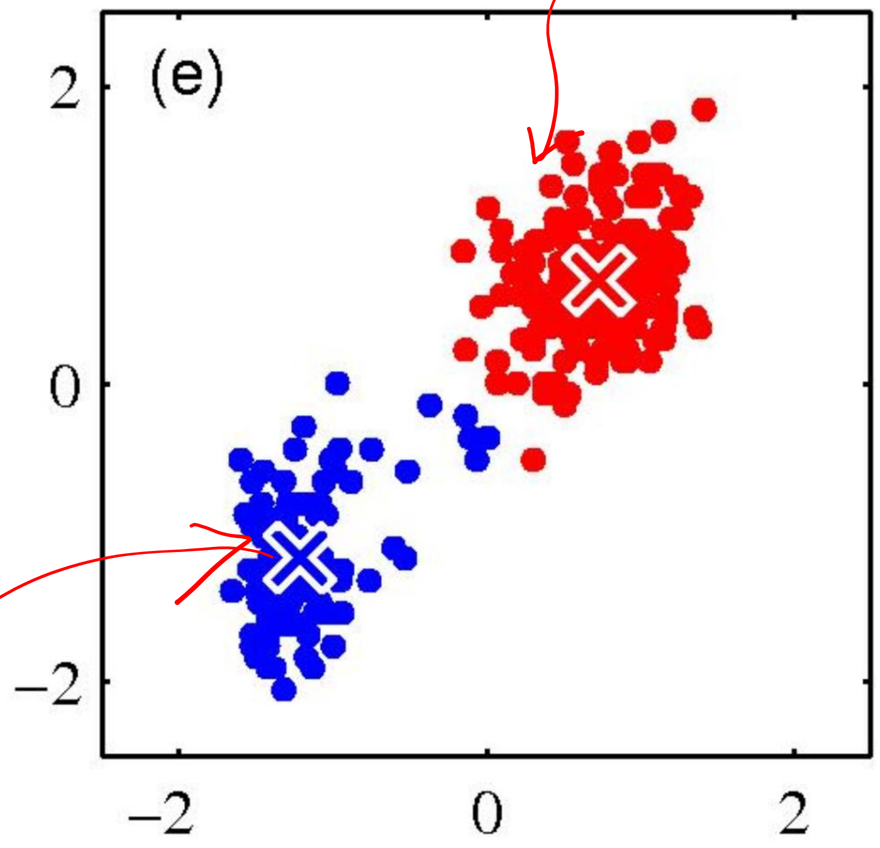
- Assign data points to ***closest centroid***



Iteration 2

- Change each ***centroid*** to the **average** of the assigned points

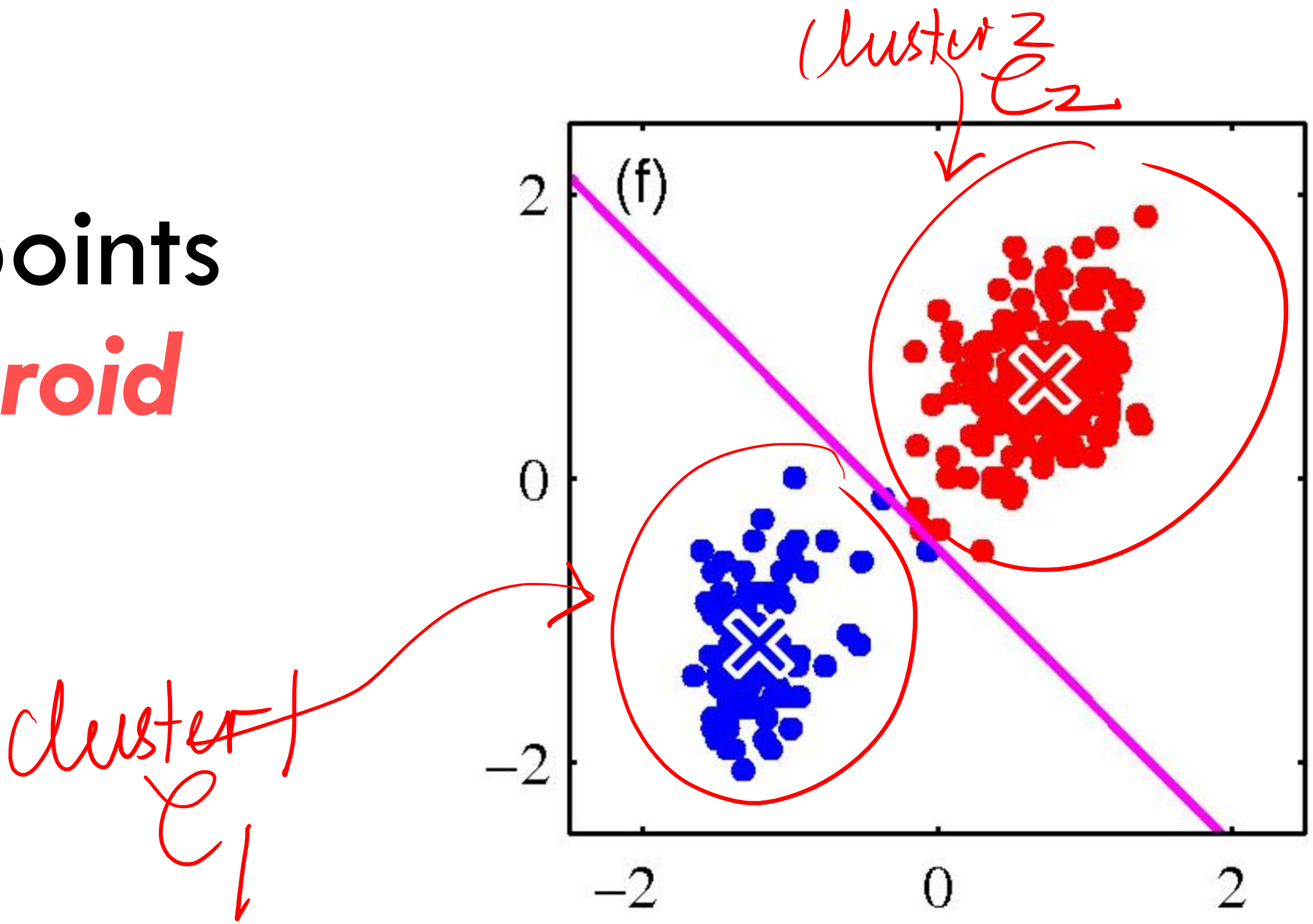
$\bar{u}_2^{(2)}$



$\bar{u}_1^{(2)}$

Iteration 3

- Assign data points to ***closest centroid***

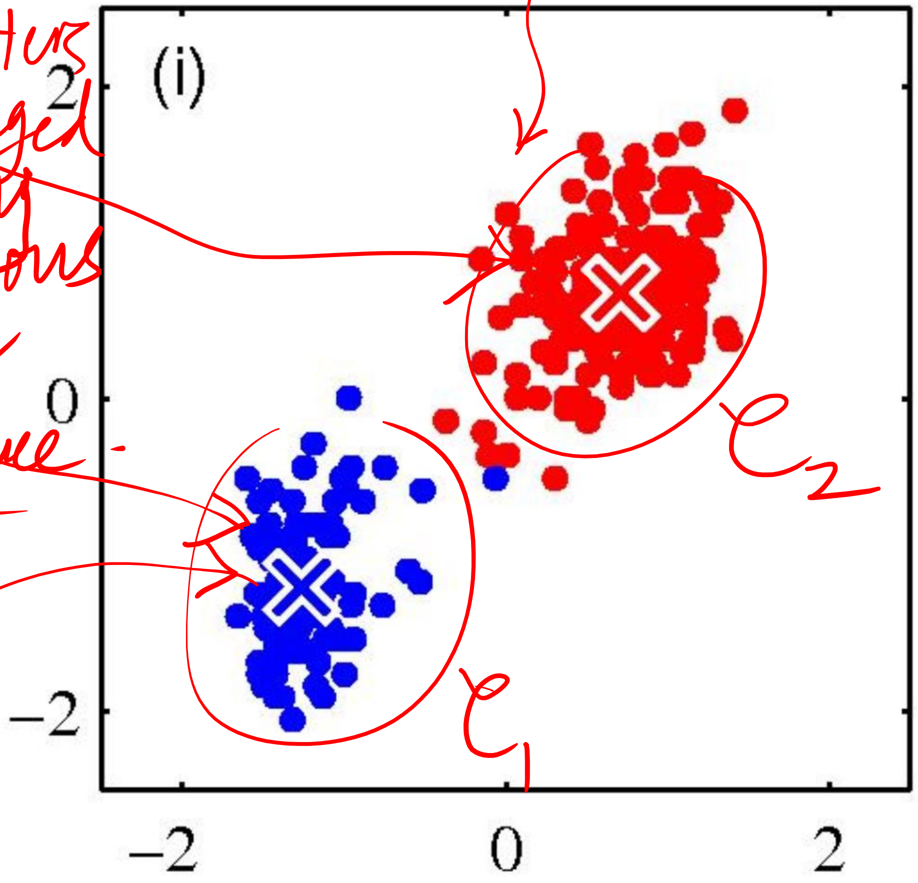


Iteration 3

- Change each **centroid** to the **average** of the assigned points

Centroids/clusters have NOT changed significantly from previous iteration ⇒ convergence.

- ***Convergence achieved!***



$\bar{\mu}_1^{(3)}$

Instructors may use this white area (14.5 cm / 25.4 cm) for the text.
Three options provided below for the font size.

Font: Avenir (Book), Size: 32, Colour: Dark Grey

Font: Avenir (Book), Size: 28, Colour: Dark Grey

Font: Avenir (Book), Size: 24, Colour: Dark Grey

Do not use the space below.