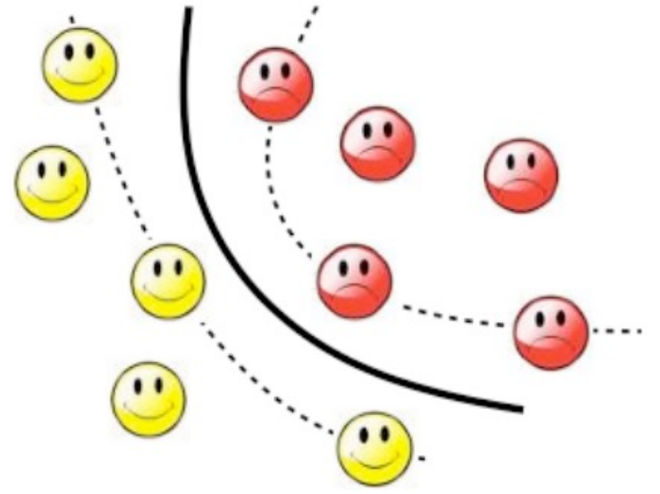
THE UNIVERSITY OF SYDNEY



Machine Learning and Data Mining

(COMP 5318)

Clustering and Expectation-Maximisation

Nguyen Hoang Tran

THE UNIVERSITY OF SYDNEY

Clustering

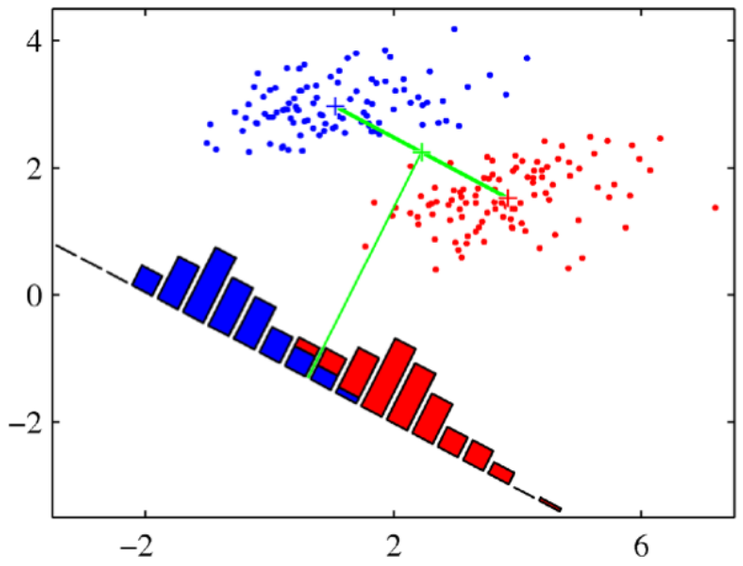
C. Bishop, *Pattern Recognition and Machine Learning*,
Chapter 9: Mixture Models and EM
Springer New York, 2006

K.P. Murphy, *Machine Learning: a Probabilistic Perspective*,
Chapters 11 and 25, Massachusetts Institute of Technology, 2006

Types of Learning

THE UNIVERSITY OF SYDNEY

Supervised Learning

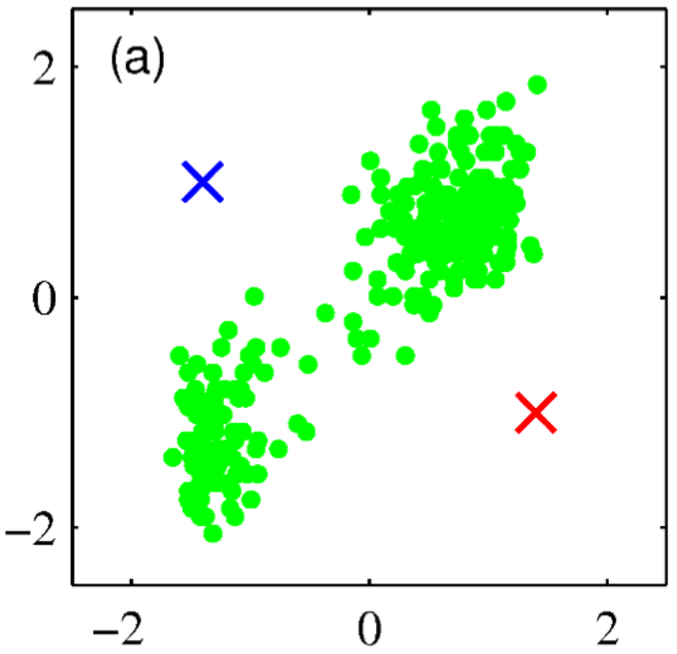


Learning input-output from examples

Regression

Classification

Unsupervised Learning

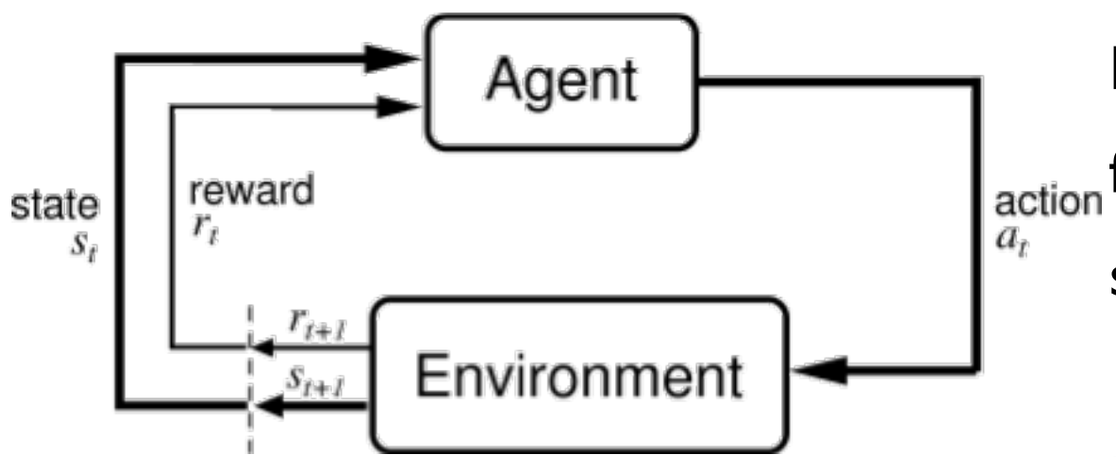


Learning underlying structure

Clustering

Density Estimation

Reinforcement Learning



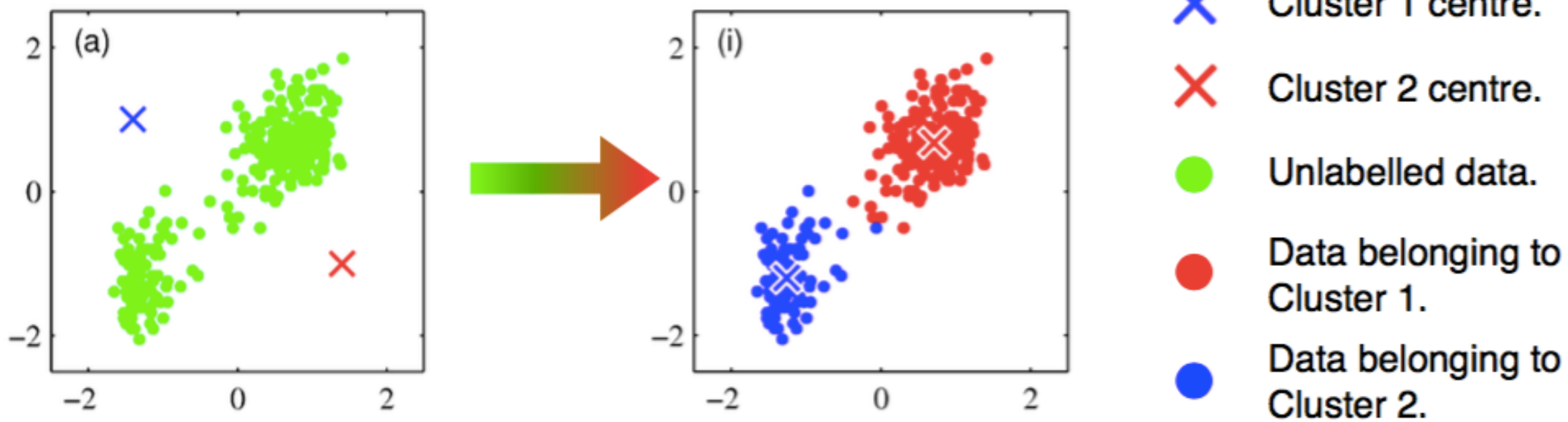
Learning policy from state-action-reward sequences.

Learning behaviour

Clustering

THE UNIVERSITY OF SYDNEY

Process of grouping similar objects together



Learn a set of clusters and assign data to a specific cluster.

Deterministic: Hard assignment to each cluster (K-means).

Probabilistic: Model assignment as a discrete latent variable.

(Mixtures of Gaussians, Dirichlet Process)

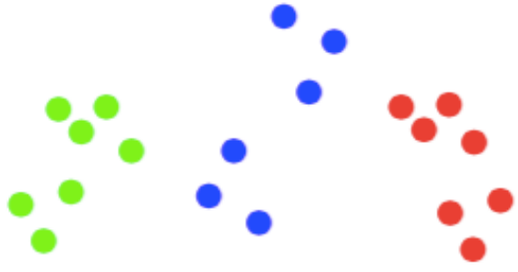
Clustering

How many clusters?

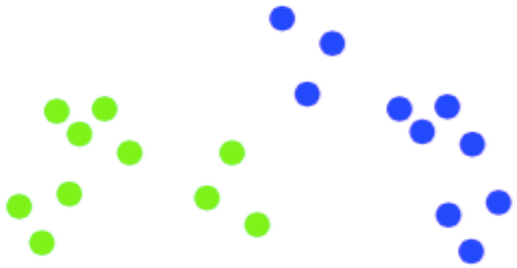
Clustering

How many clusters?

One



Three



Two

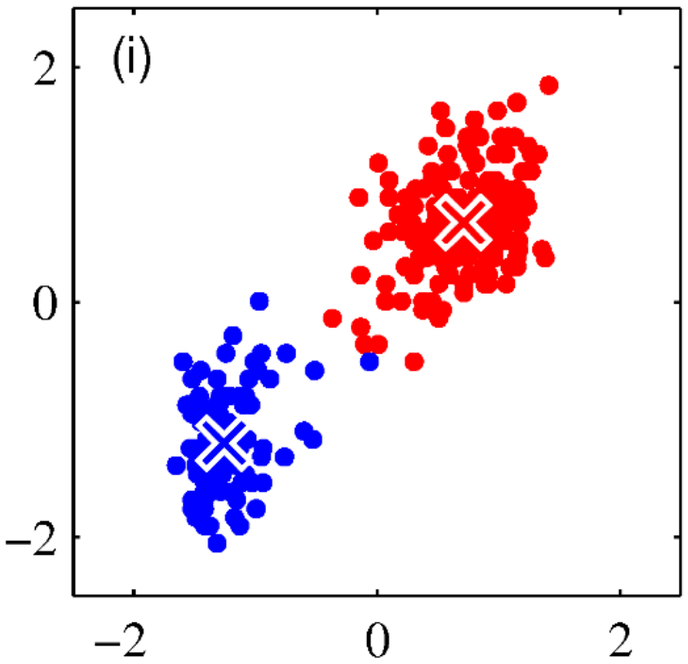
Six

Presence of ambiguous solutions.

Types of Clustering

THE UNIVERSITY OF SYDNEY

Partition Clustering

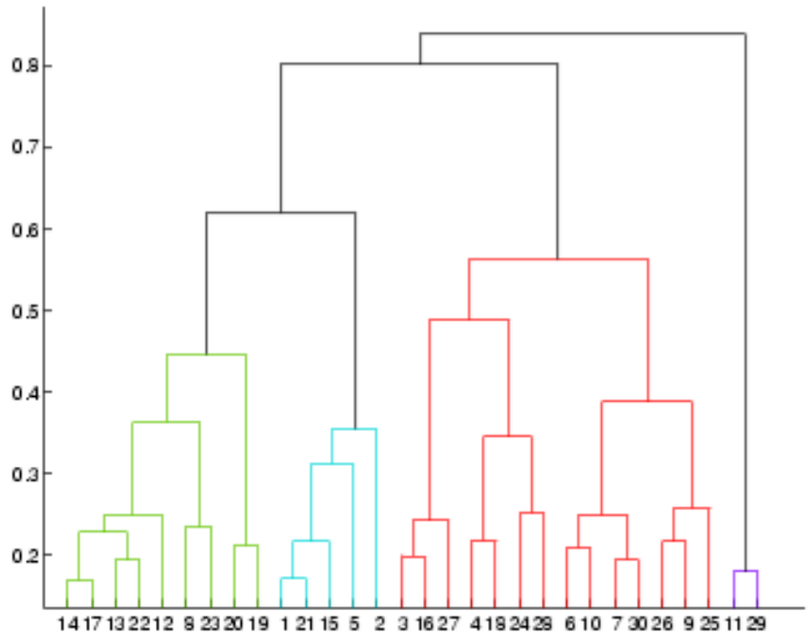


Partition the objects into disjoint sets. Faster to create.

Sensible to initial conditions.

Model selection for K.

Hierarchical Clustering



Nested tree of partitions.

Slower to create.

Often more useful.

Do not require knowing the number of clusters.

Clustering

THE UNIVERSITY OF SYDNEY

Dataset $\mathcal{D} = \{\mathbf{x}_1, ..., \mathbf{x}_N\}$ with N Observations

Each data point is D dimension

Goal: Partition dataset into K clusters. (For now, assume K is given)

$\boldsymbol{\mu}_k = (\mu_1, ..., \mu_D)$: Centroid for each cluster $k \in 1, ..., K$

Binary indicator variables

$$r_{nk} = \begin{cases} 1, & \text{if datapoint } n \text{ belongs to cluster } k \\ 0, & \sim \end{cases}$$

If $\mathbf{x}_n$ is assigned to cluster k, then $r_{nk} = 1 \ \wedge \ r_{nj} = 0 \ \forall j \neq k$

K-Means

THE UNIVERSITY OF SYDNEY

Objective function:

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \|\mathbf{x}_n - \boldsymbol{\mu}_k\|^2$$

Represents the sum of the squares of the distances of each datapoint to its assigned centroid vector.

Goal: Find $\{\boldsymbol{\mu}_k\}$ and $\{r_{nk}\}$ that minimise J.

$$\{r_{nk}, \boldsymbol{\mu}_k\}^\star = \underset{\{r_{nk}, \boldsymbol{\mu}_k\}}{\operatorname{argmin}} \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \|\mathbf{x}_n - \boldsymbol{\mu}_k\|^2$$

K-Means

THE UNIVERSITY OF SYDNEY

Iterative solution to minimise J:

1. Data Preprocessing
2. Initialise $\{\boldsymbol{\mu}_k\}$
3. Repeat 4 and 5 until convergence or Max Iterations
4. Minimise J w.r.t. $\{r_{nk}\}$ keeping $\{\boldsymbol{\mu}_k\}$
5. Minimise J w.r.t. $\{\boldsymbol{\mu}_k\}$ keeping $\{r_{nk}\}$

K-Means

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \|\mathbf{x}_n - \boldsymbol{\mu}_k\|^2$$

Line 4: Optimise w.r.t r_{nk}

Each data point is independent, so we can optimise for each n separately:

$$r_{nk} = \begin{cases} 1 & \text{if } k = \underset{j}{\operatorname{argmin}} \|\mathbf{x}_n - \boldsymbol{\mu}_j\|^2 \\ 0 & \text{Otherwise} \end{cases}$$

Assign each data point to its closest centroid.

K-Means

THE UNIVERSITY OF SYDNEY

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \|\mathbf{x}_n - \boldsymbol{\mu}_k\|^2$$

Line 5: Optimise w.r.t $\boldsymbol{\mu}_k$

$$\frac{\partial J}{\partial \boldsymbol{\mu}_k} = 0$$

$$2 \sum_{n=1}^{N} r_{nk} (\mathbf{x}_n - \boldsymbol{\mu}_k) = 0$$

$$\sum_{n=1}^{N} r_{nk} \mathbf{x}_n = \sum_{n=1}^{N} r_{nk} \boldsymbol{\mu}_k$$

$$\frac{\sum_n r_{nk} \mathbf{x}_n}{\sum_n r_{nk}} = \boldsymbol{\mu}_k$$

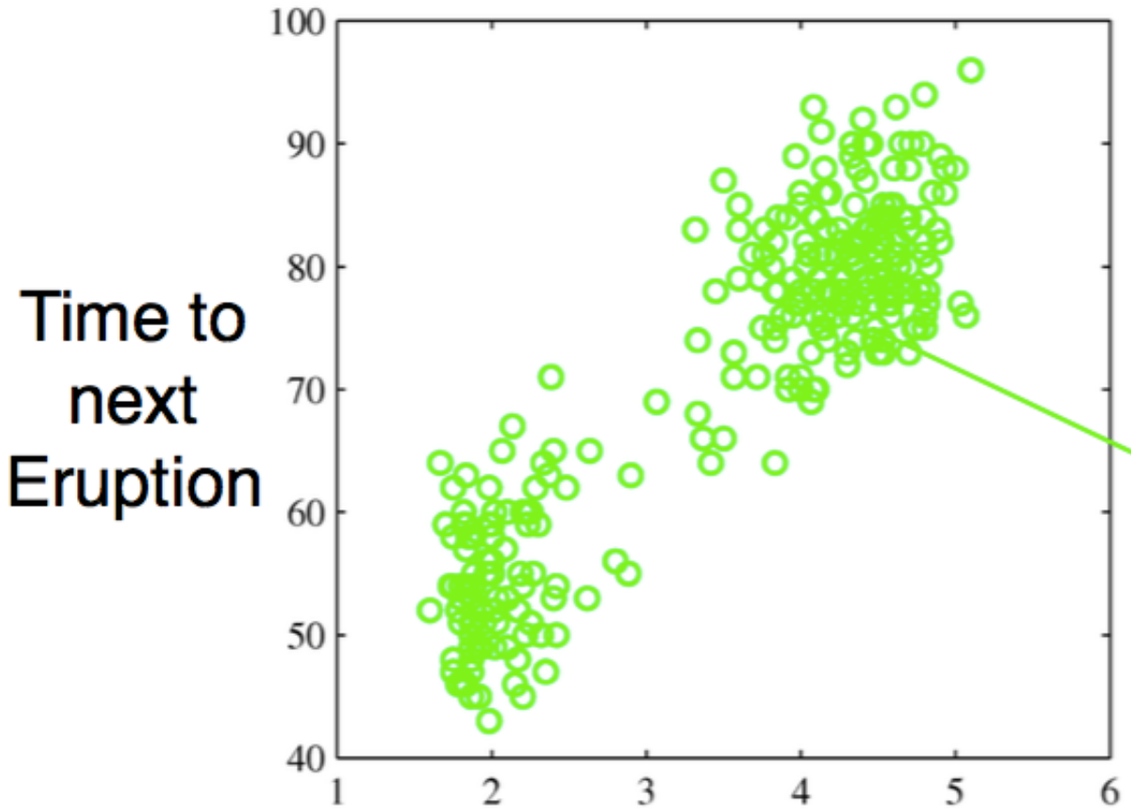
Set $\boldsymbol{\mu}_k$ equal to the mean of all data points $\mathbf{x}_n$ assigned to cluster k.

K-means

K-Means Example

THE UNIVERSITY OF SYDNEY

Hydrothermal Geyser: Old Faithful



Singe Eruption

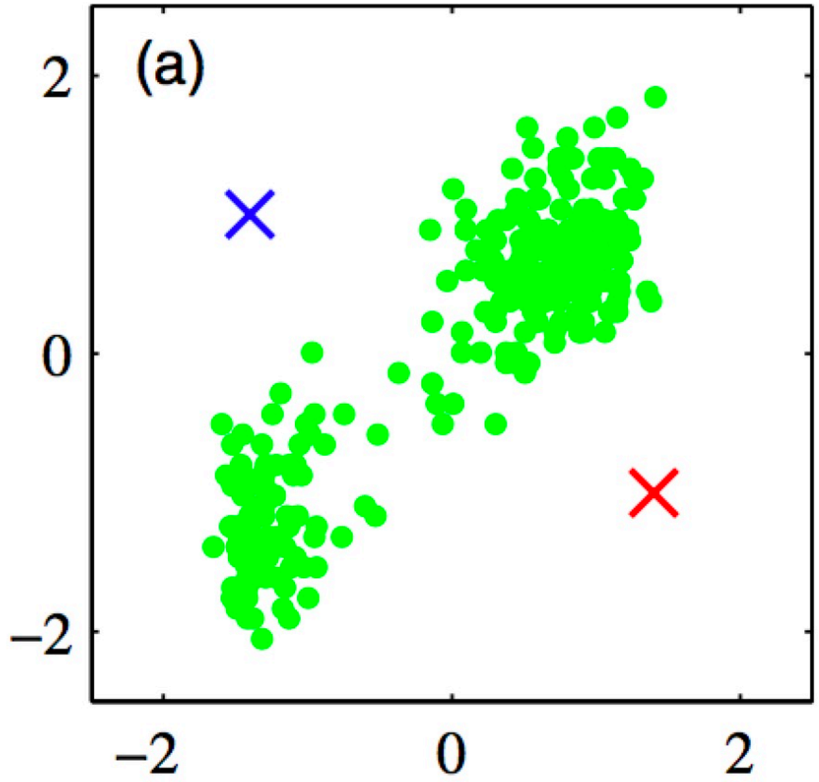
$N = 272$ Observations

K-Means Example

THE UNIVERSITY OF SYDNEY

Number of clusters: $K = 2$

1 Data Preprocessing
2 Initialise $\{\boldsymbol{\mu}_k\}$
3 Repeat until convergence or Max Iterations
4 Minimise J w.r.t. $\{r_{nk}\}$ keeping $\{\boldsymbol{\mu}_k\}$ fixed.
5 Minimise J w.r.t. $\{\boldsymbol{\mu}_k\}$ keeping $\{r_{nk}\}$ fixed.



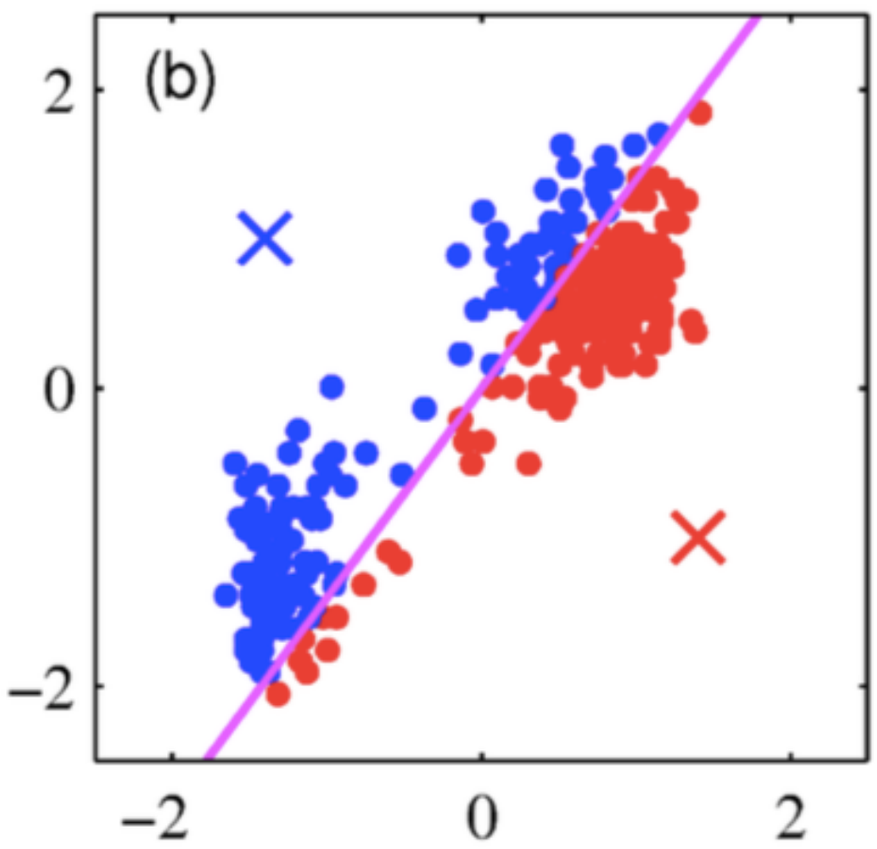
Each dimension has zero mean and unit standard deviation.

Better initialisation: Choose$\{\mu_k\}$ as average of a random subset.

K-Means Example

THE UNIVERSITY OF SYDNEY

Number of clusters: $K = 2$



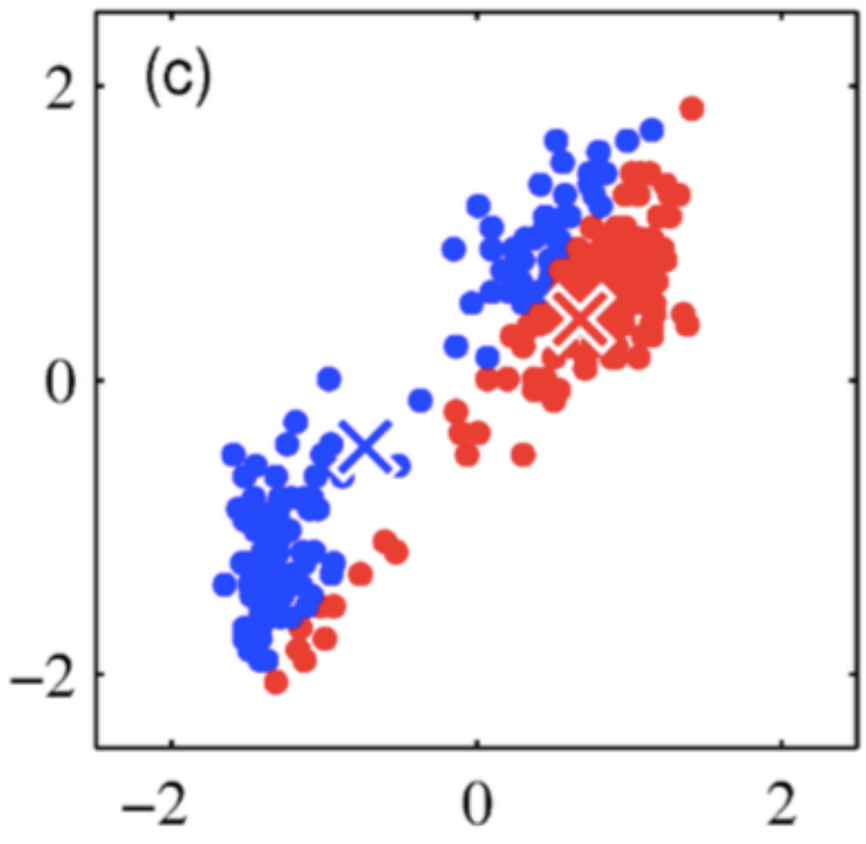
1 Data Preprocessing
2 Initialise $\{\boldsymbol{\mu}_k\}$
3 Repeat until convergence or Max Iterations
4 Minimise J w.r.t. $\{r_{nk}\}$ keeping $\{\boldsymbol{\mu}_k\}$ fixed.
5 Minimise J w.r.t. $\{\boldsymbol{\mu}_k\}$ keeping $\{r_{nk}\}$ fixed.

Each data point is assigned to the closest cluster centre.

K-Means Example

THE UNIVERSITY OF SYDNEY

Number of clusters: $K = 2$



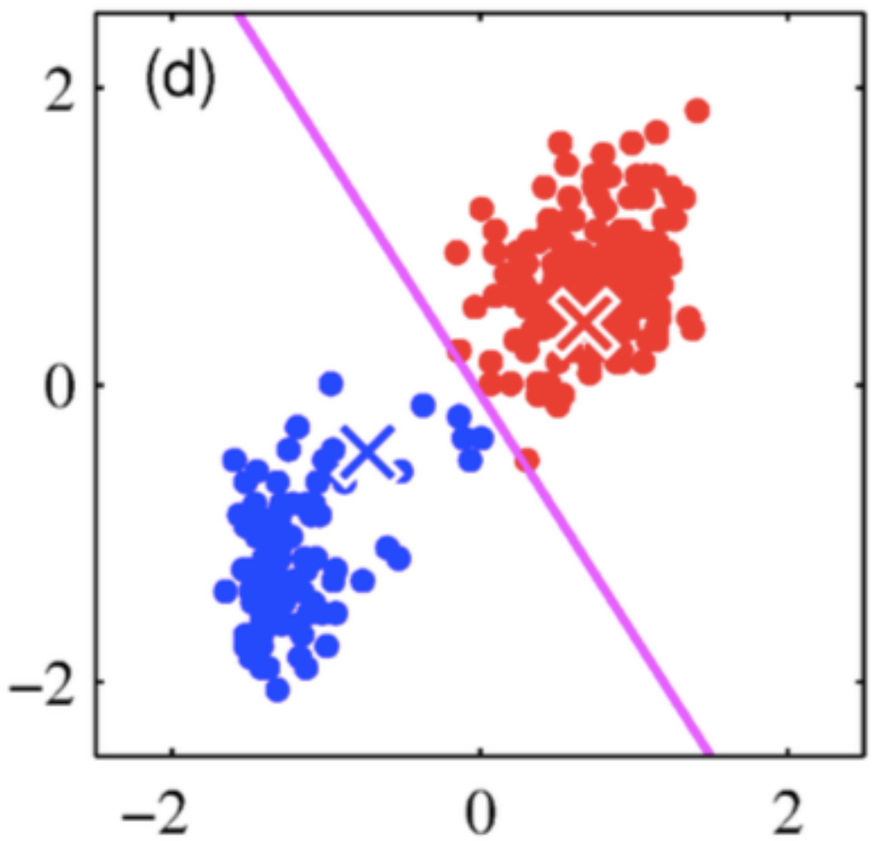
1 Data Preprocessing
2 Initialise $\{\boldsymbol{\mu}_k\}$
3 Repeat until convergence or Max Iterations
4 Minimise J w.r.t. $\{r_{nk}\}$ keeping $\{\boldsymbol{\mu}_k\}$ fixed.
5 Minimise J w.r.t. $\{\boldsymbol{\mu}_k\}$ keeping $\{r_{nk}\}$ fixed.

Re-compute each cluster centre to be the mean of the points previously assigned.

K-Means Example

THE UNIVERSITY OF SYDNEY

Number of clusters: $K = 2$



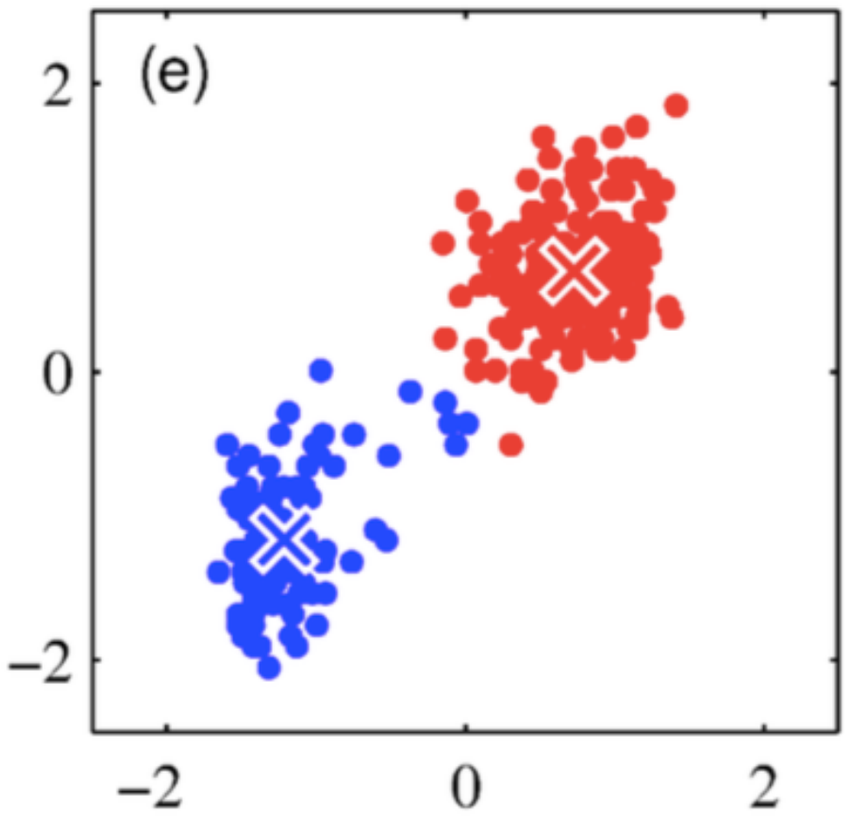
1 Data Preprocessing
2 Initialise $\{\boldsymbol{\mu}_k\}$
3 Repeat until convergence or Max Iterations
4 Minimise J w.r.t. $\{r_{nk}\}$ keeping $\{\boldsymbol{\mu}_k\}$ fixed.
5 Minimise J w.r.t. $\{\boldsymbol{\mu}_k\}$ keeping $\{r_{nk}\}$ fixed.

Each data point is assigned to the closest cluster centre.

K-Means Example

THE UNIVERSITY OF SYDNEY

Number of clusters: $K = 2$



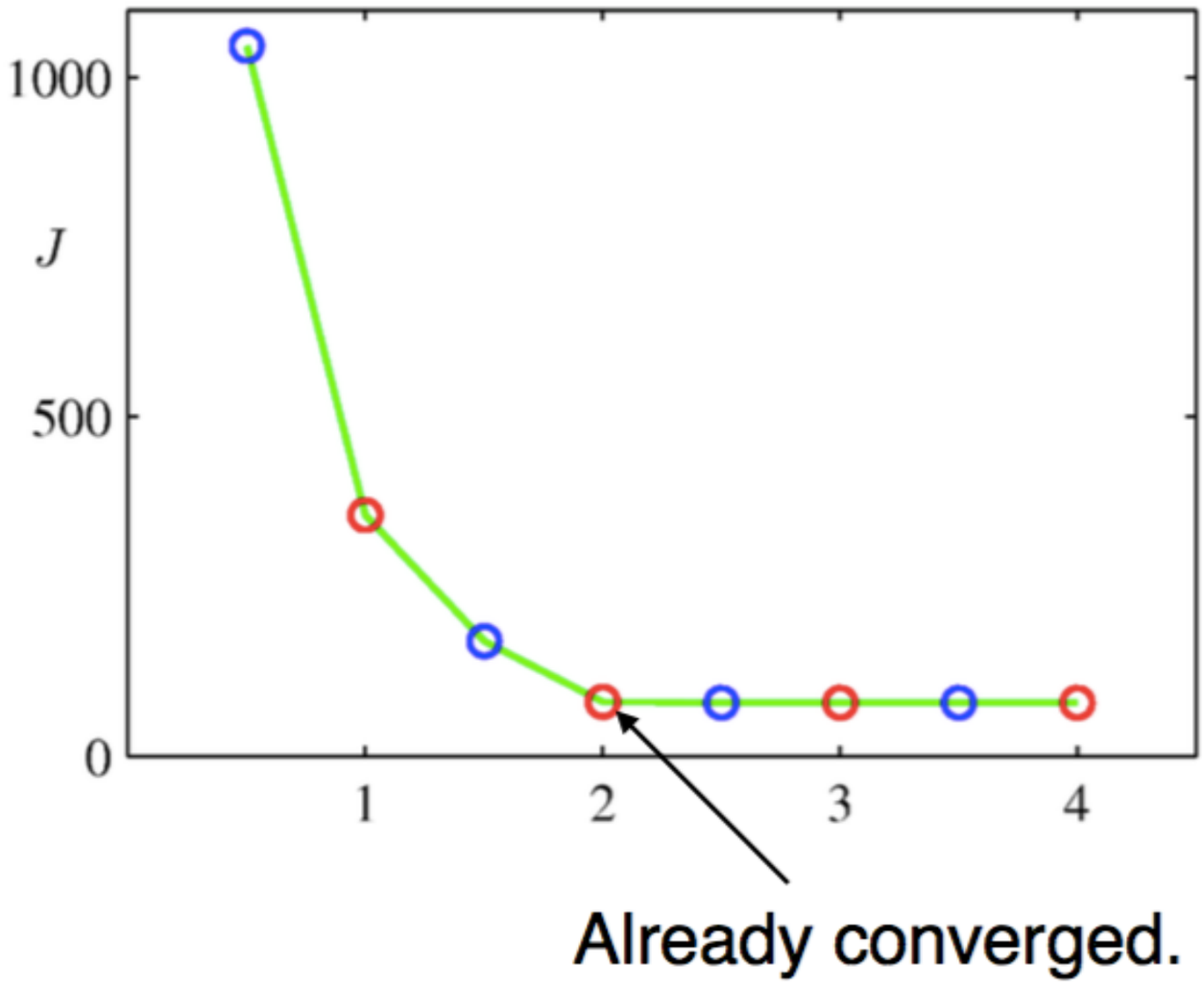
1 Data Preprocessing
2 Initialise $\{\boldsymbol{\mu}_k\}$
3 Repeat until convergence or Max Iterations
4 Minimise J w.r.t. $\{r_{nk}\}$ keeping $\{\boldsymbol{\mu}_k\}$ fixed.
5 Minimise J w.r.t. $\{\boldsymbol{\mu}_k\}$ keeping $\{r_{nk}\}$ fixed.

Re-compute each cluster centre to be the mean of the points previously assigned.

K-Means Example

THE UNIVERSITY OF SYDNEY

Plot of the cost function for each iteration.



K-Means Example 2

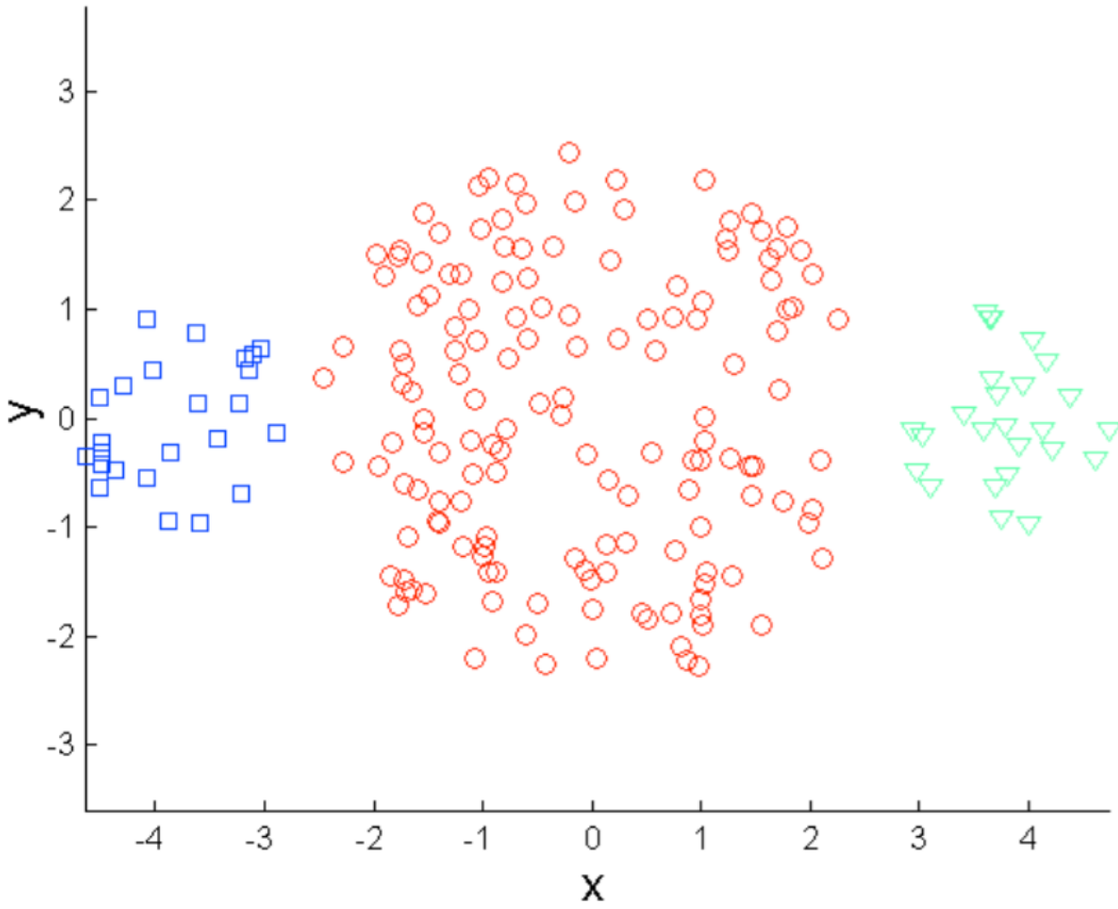
THE UNIVERSITY OF SYDNEY

Image segmentation and compression.

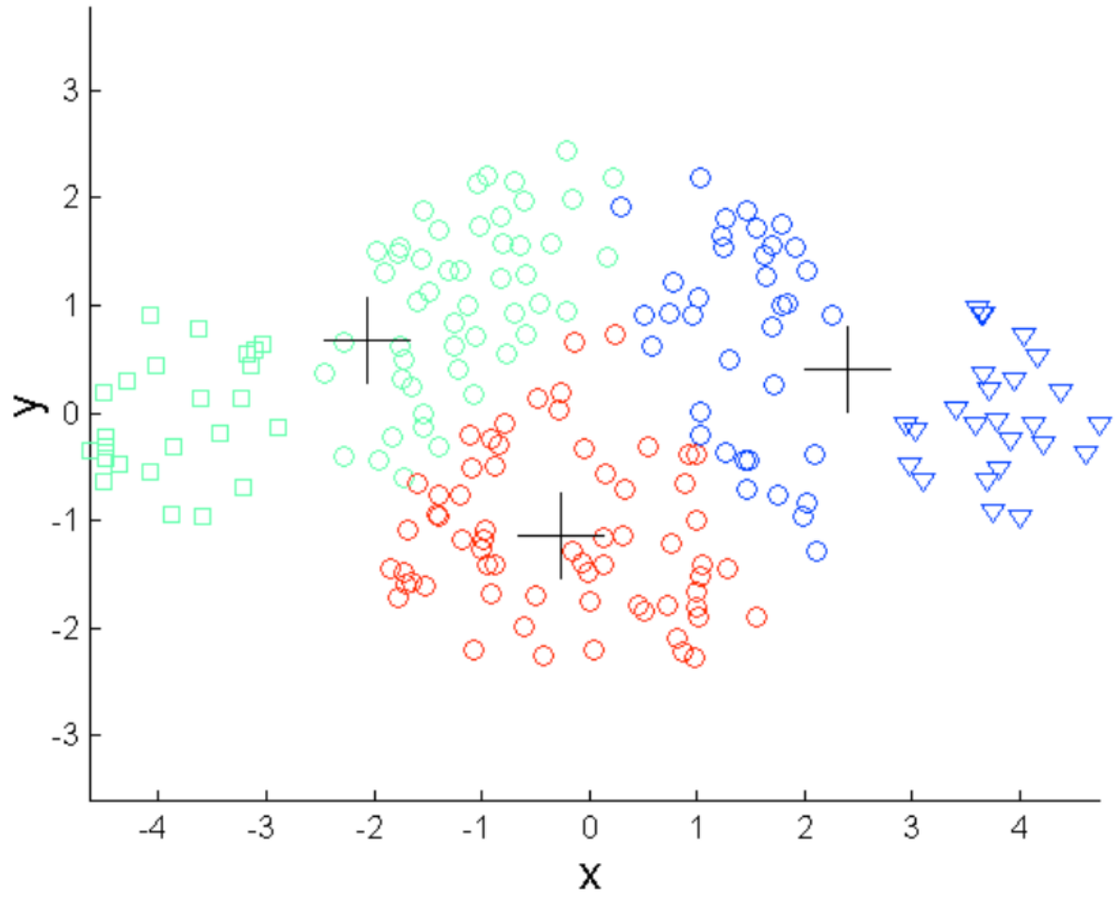
K-Means Limitations

THE UNIVERSITY OF SYDNEY

Differing Sizes:



Original Points

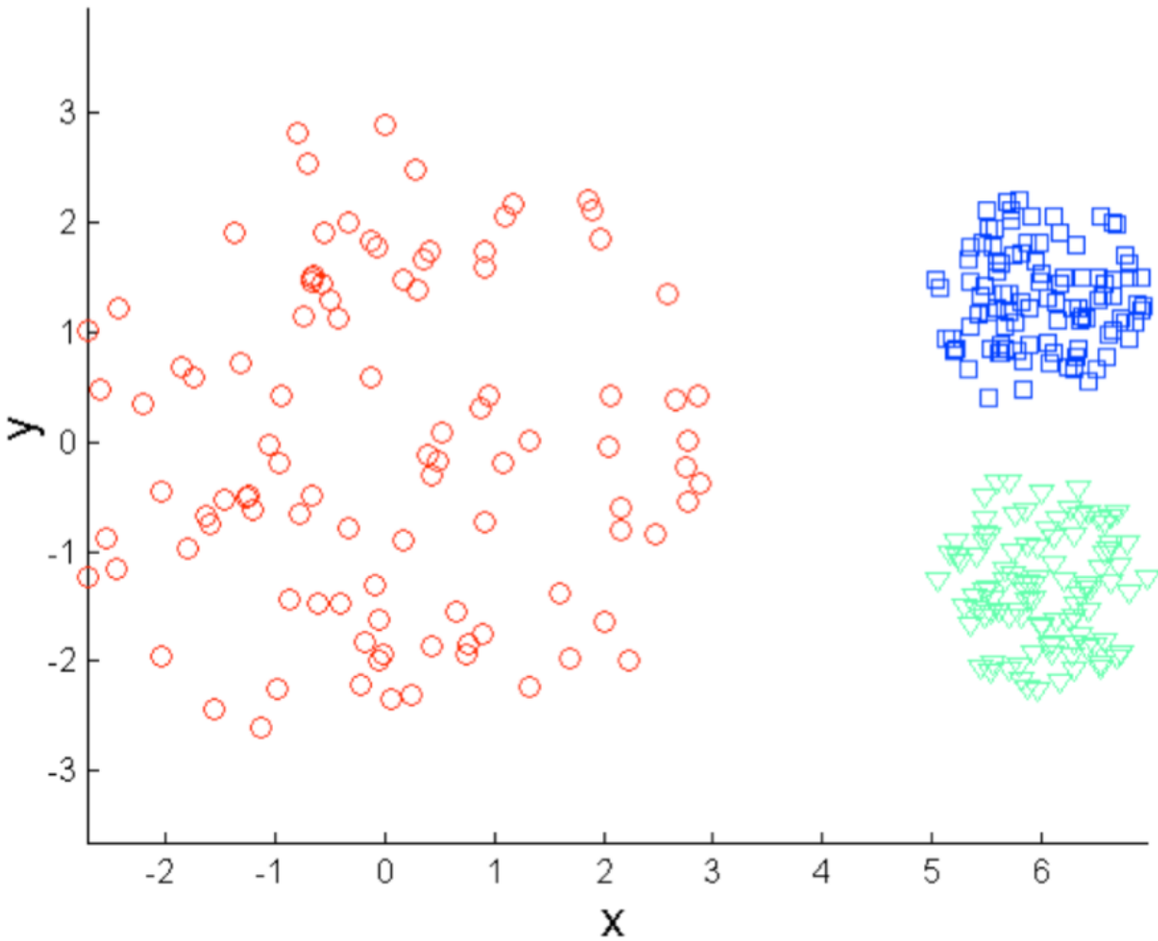


K-Means 3 Clusters

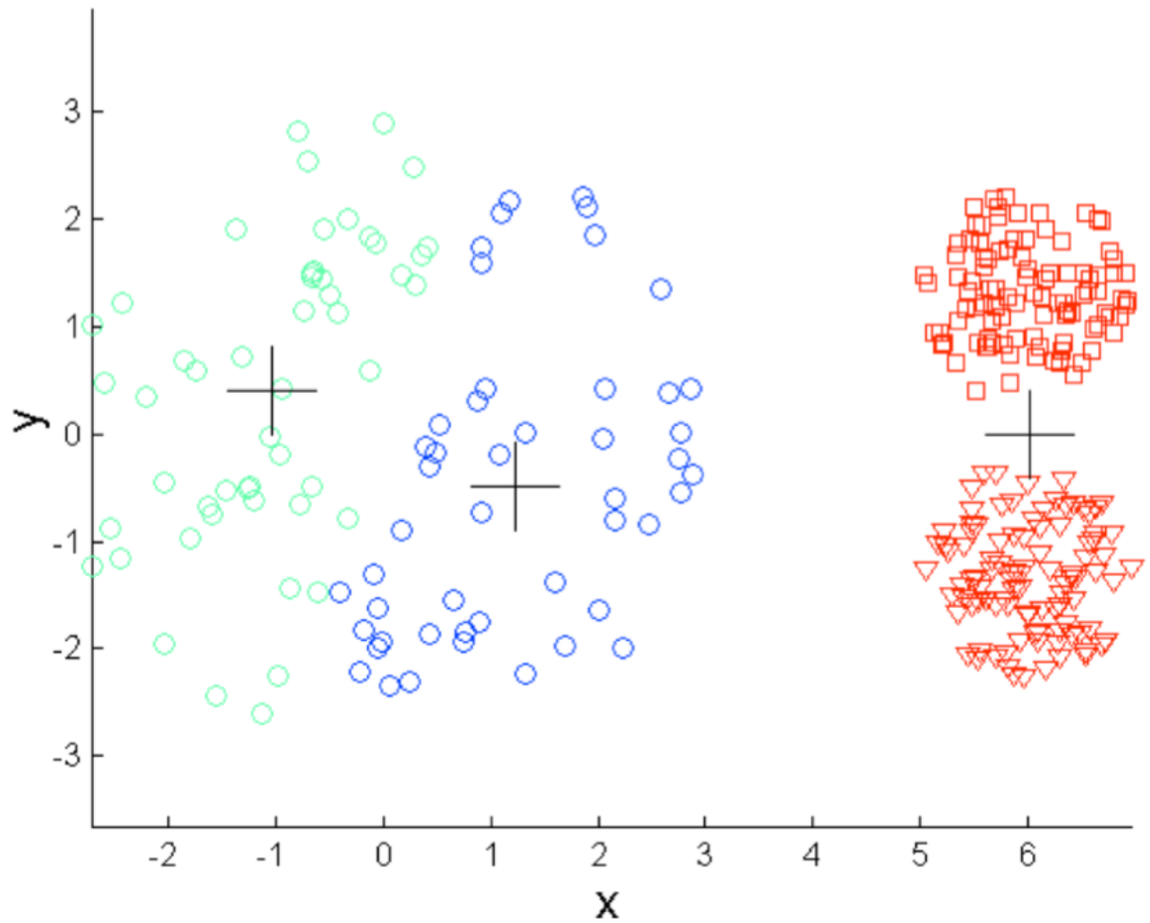
K-Means Limitations

THE UNIVERSITY OF SYDNEY

Differing Density:



Original Points

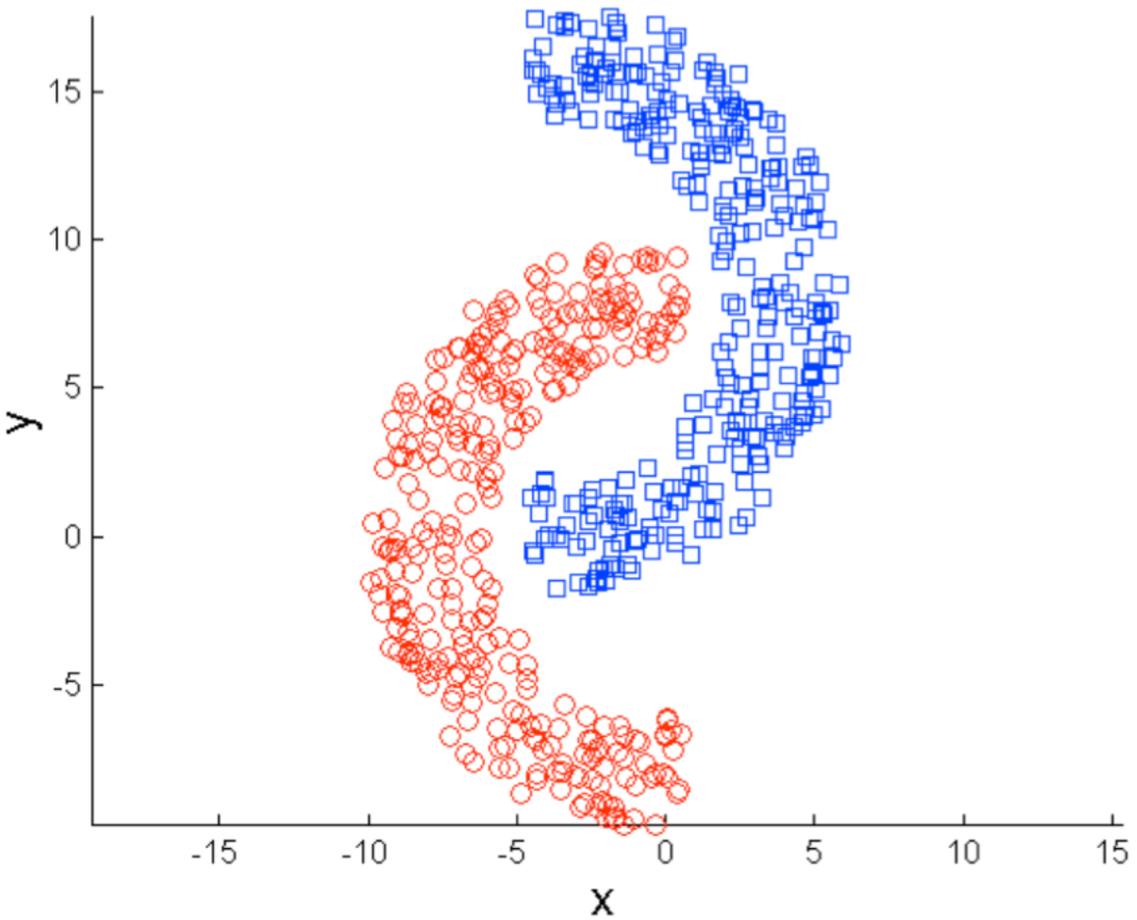


K-Means 3 Clusters

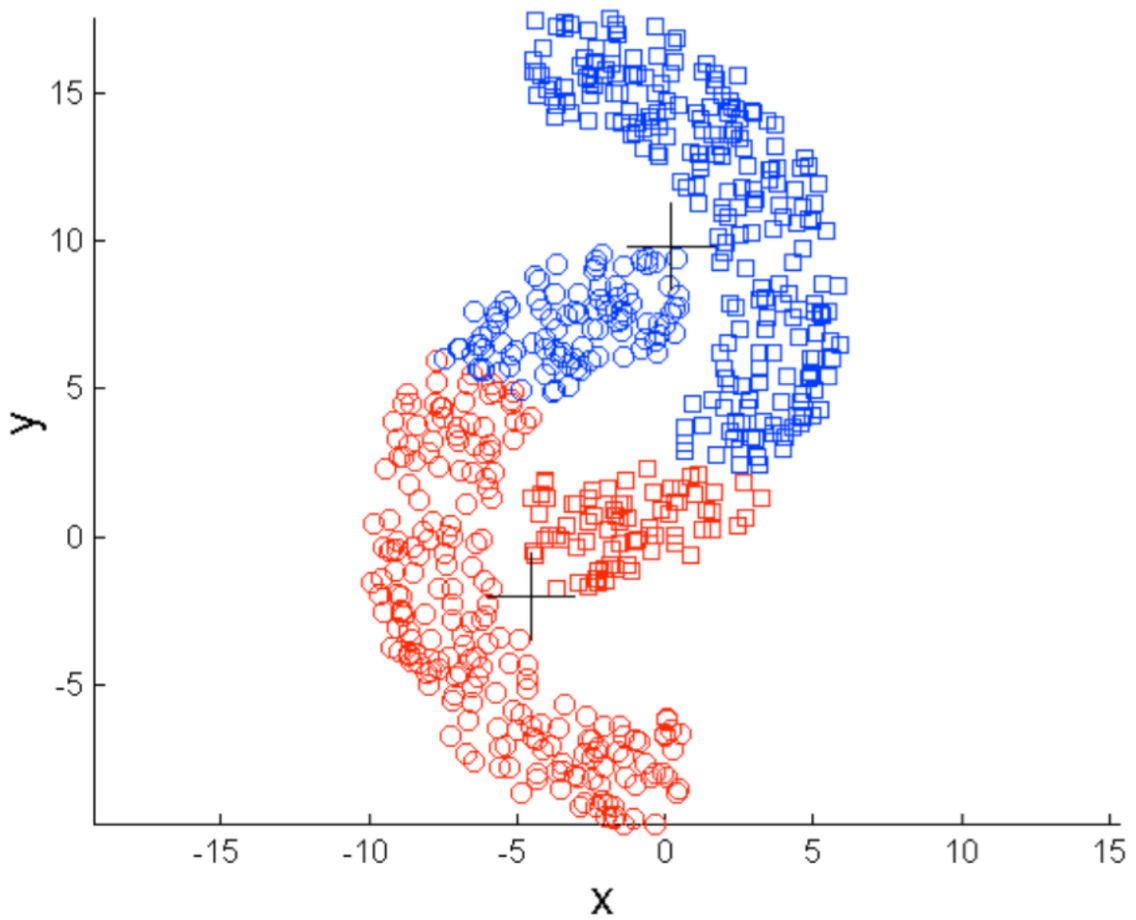
K-Means Limitations

THE UNIVERSITY OF SYDNEY

Non-Globular Shapes:



Original Points

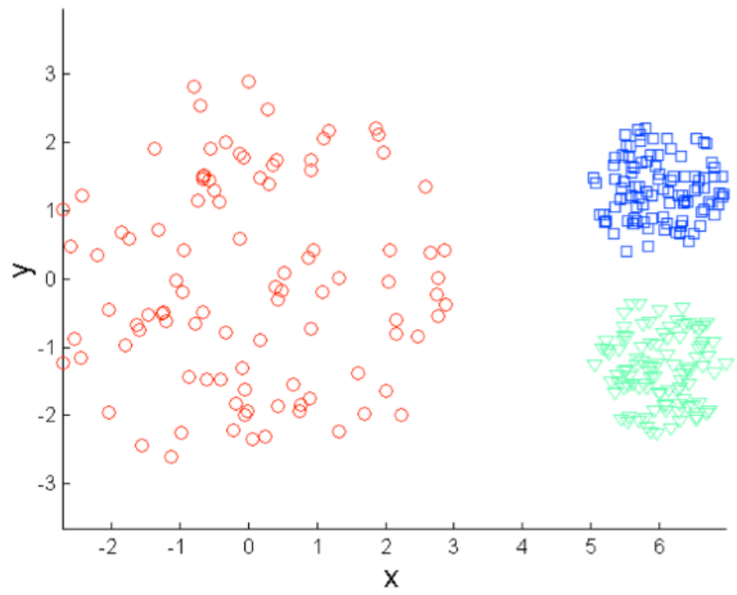


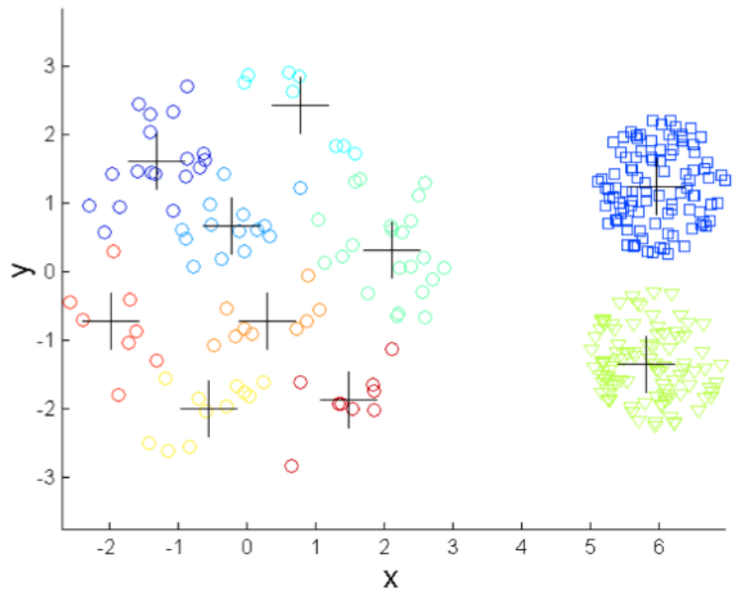
K-Means 2 Clusters

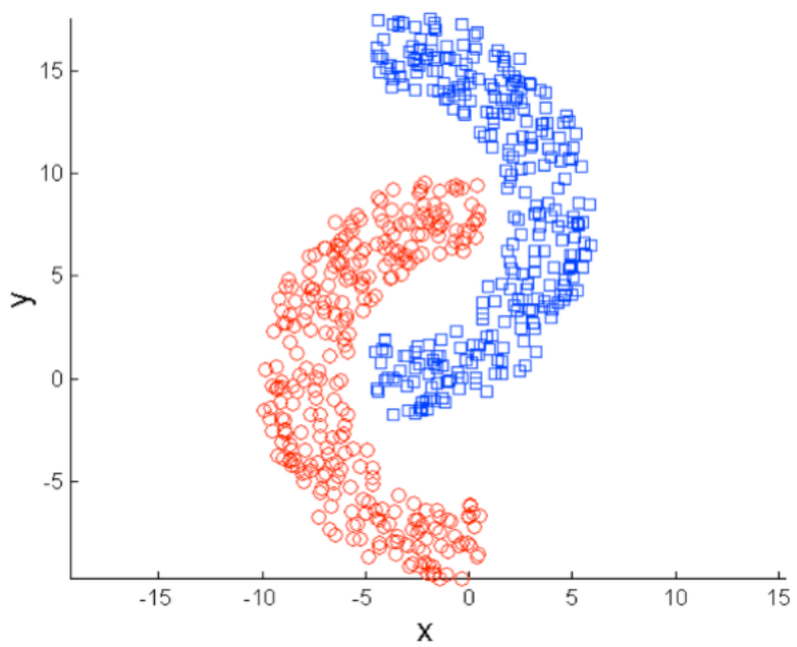
Overcome K-Means Limitations

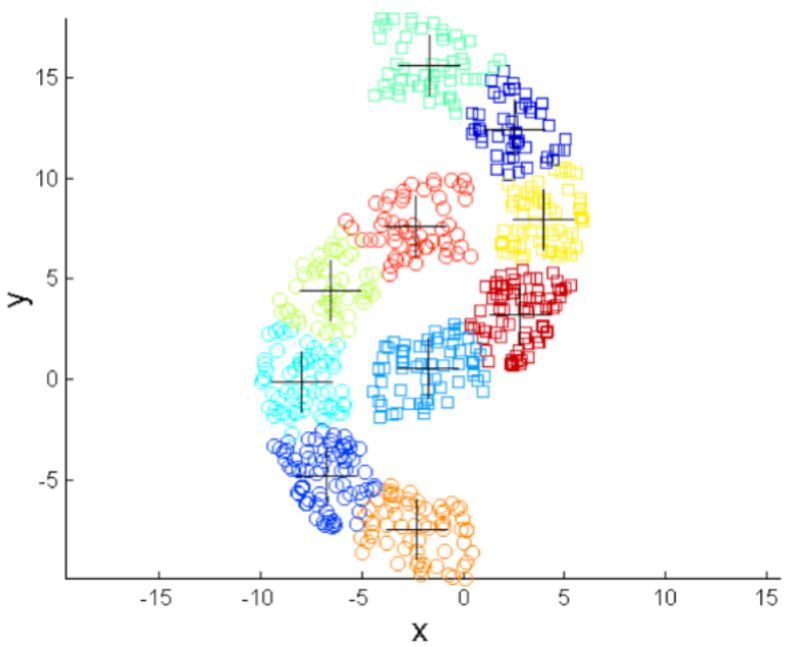
THE UNIVERSITY OF SYDNEY

Use large number of clusters.







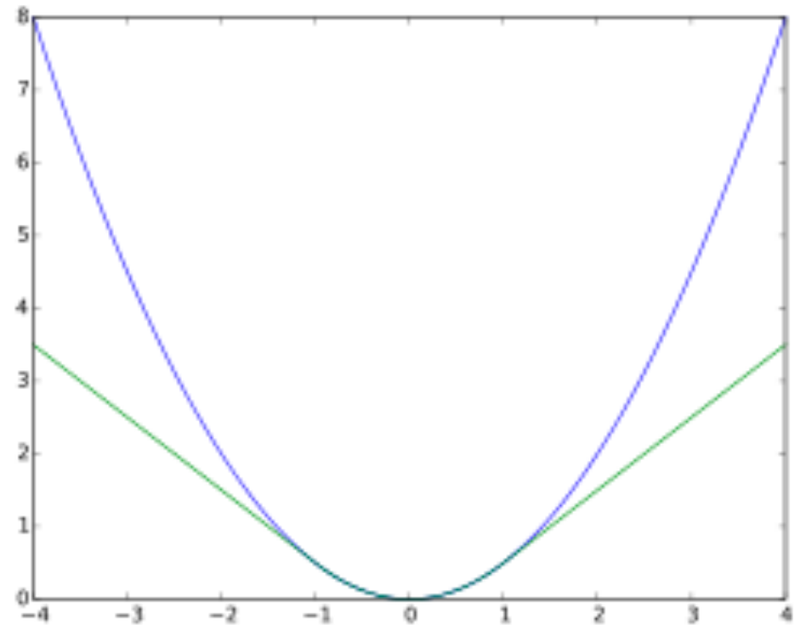


K-Means Enhancement

THE UNIVERSITY OF SYDNEY

Generalise distance function:

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \mathcal{V}(\mathbf{x}_n, \boldsymbol{\mu}_k)$$



Robustness to outliers.

$$\mathcal{V}(\mathbf{x}_n, \mu_k) = \begin{cases} 1/2\|\mathbf{x}_n - \mu_k\|_2^2, & \text{if } \|\mathbf{x}_n - \mu_k\| \leq \delta \\ \delta\|\mathbf{x}_n - \mu_k\|_1 - 1/2\delta^2, & \text{otherwise} \end{cases}$$

THE UNIVERSITY OF SYDNEY

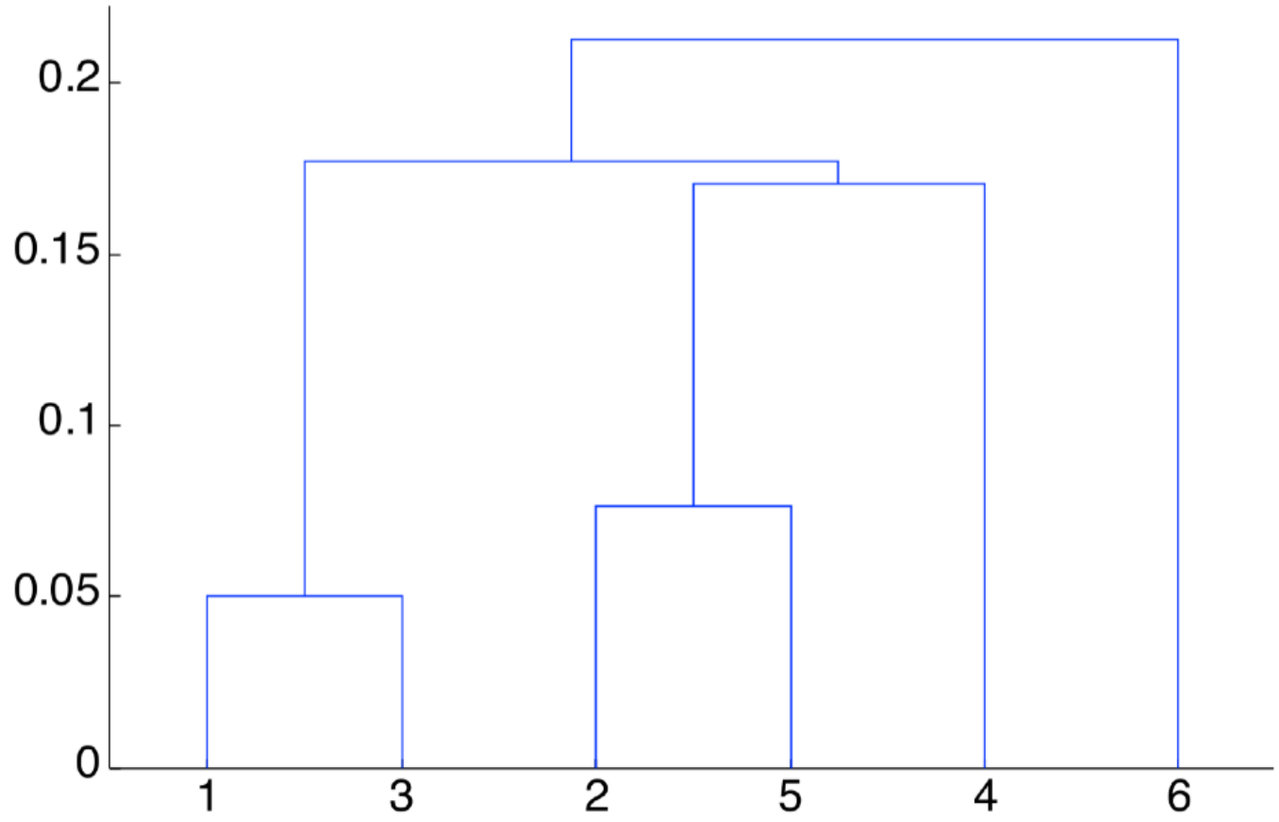
Hierarchical Clustering

Hierarchical Clustering

THE UNIVERSITY OF SYDNEY

Nested set of clusters organised as a hierarchical tree.

Any number of clusters can be obtained by 'cutting' the dendrogram.



Dendrogram

Uses a similarity matrix.

Hierarchical Agglomerative Clustering

THE UNIVERSITY OF SYDNEY

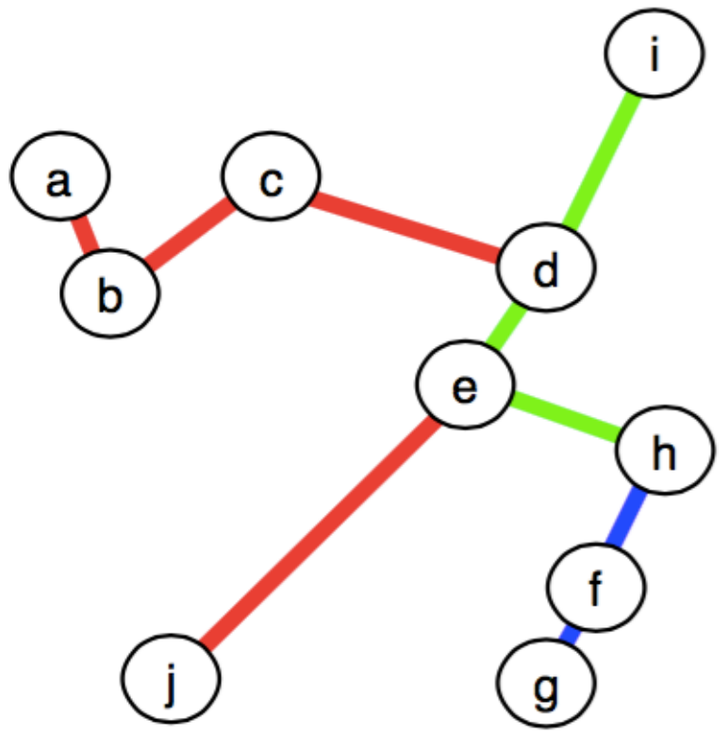
Simple clustering algorithm.
Uses a inter cluster similarity measure.

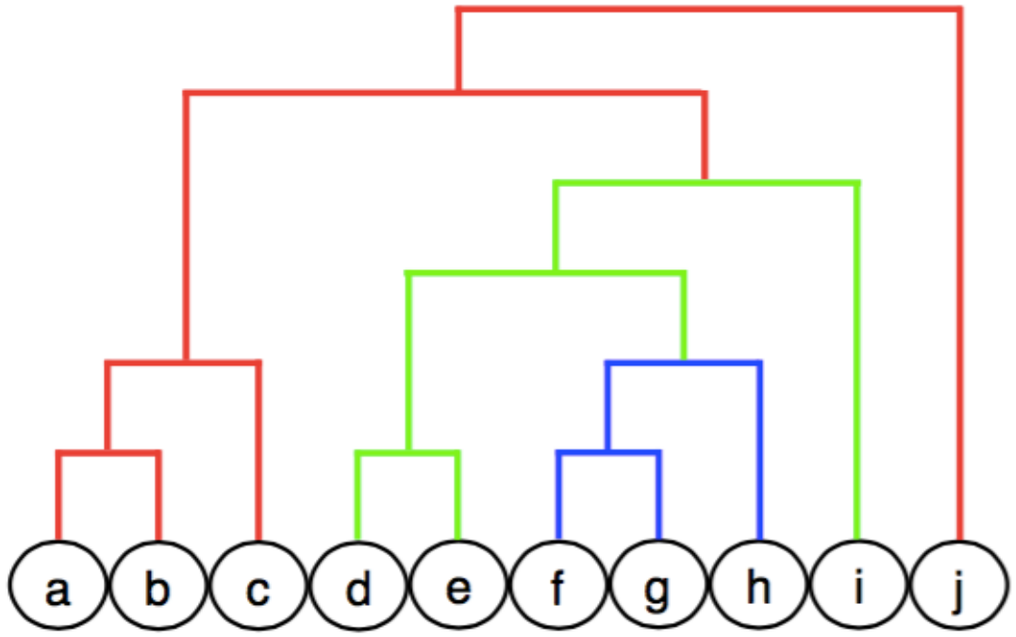
1. Initialise: Every data point is a cluster.
2. Repeat until one cluster remains.
3. Compute distances between all clusters.
4. Merge closest clusters.
5. Update dendrogram.

Hierarchical Agglomerative Clustering

THE UNIVERSITY OF SYDNEY

Example:



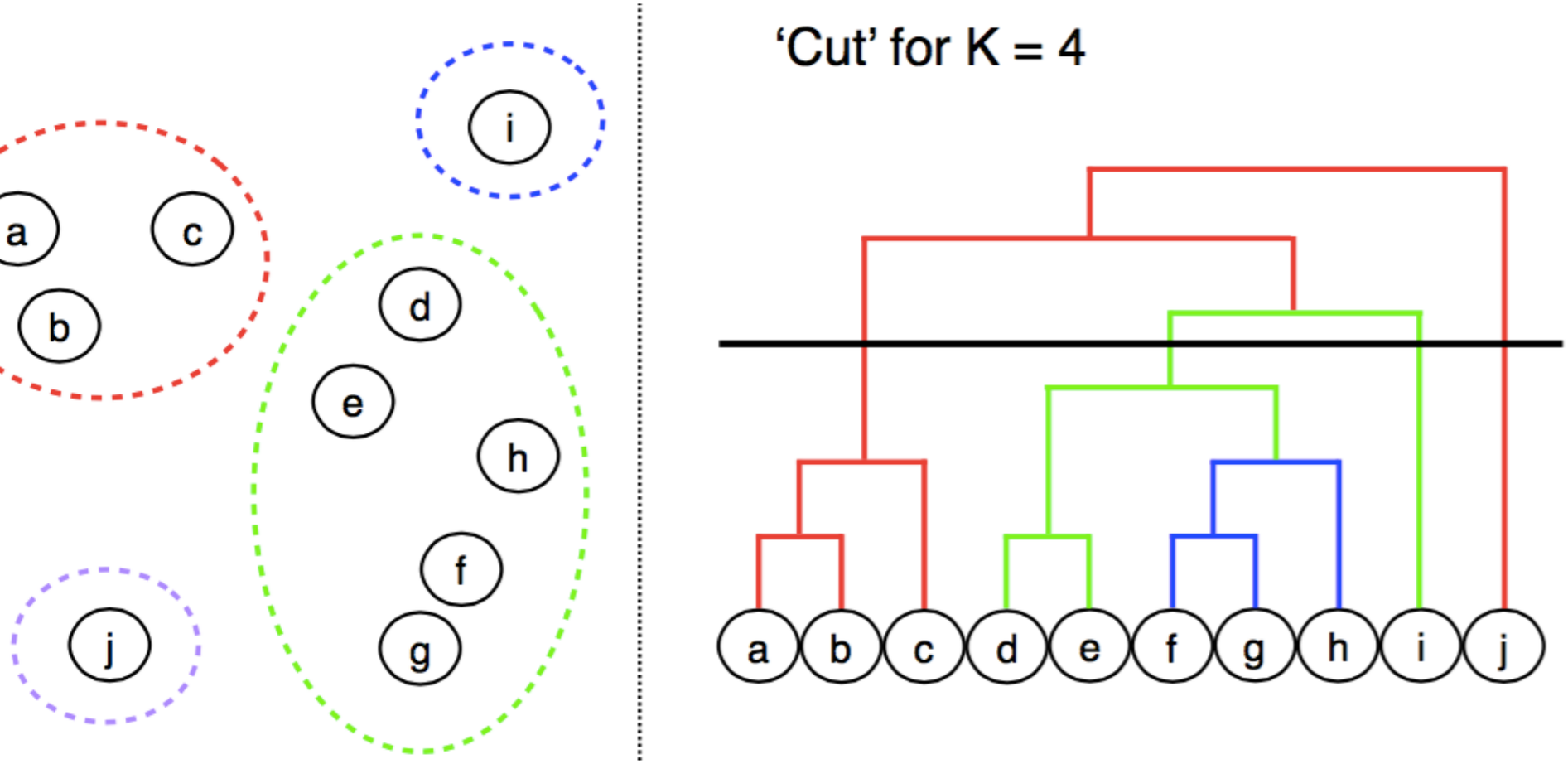


Dendrogram

Hierarchical Agglomerative Clustering

THE UNIVERSITY OF SYDNEY

Example:



Dendrogram

Inter Cluster Similarity

THE UNIVERSITY OF SYDNEY

Nearest Neighbour

$$D_{\min}(C_i, C_j) = \min_{\mathbf{x} \in C_i, \mathbf{y} \in C_j} \|\mathbf{x} - \mathbf{y}\|^2$$

Furthest Neighbour

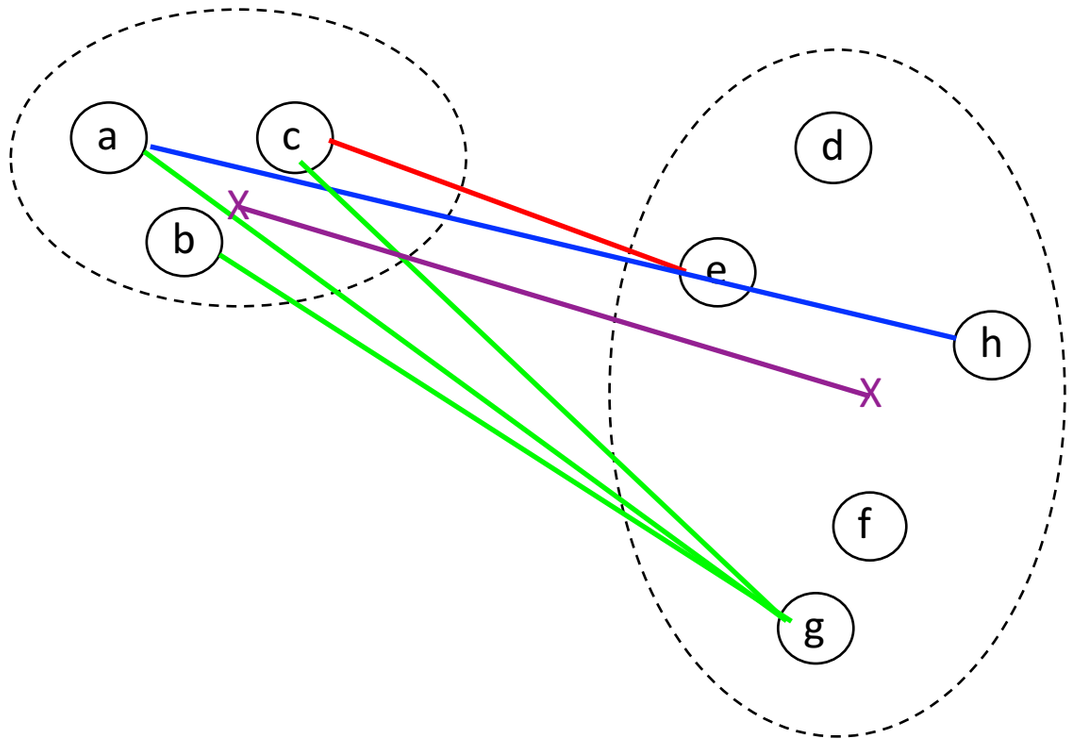
$$D_{\max}(C_i, C_j) = \max_{\mathbf{x} \in C_i, \mathbf{y} \in C_j} \|\mathbf{x} - \mathbf{y}\|^2$$

Group Average

$$D_{\text{avg}}(C_i, C_j) = \frac{1}{|C_i||C_j|} \sum_{\mathbf{x} \in C_i} \sum_{\mathbf{y} \in C_j} \|x - y\|^2$$

Centroid Distance

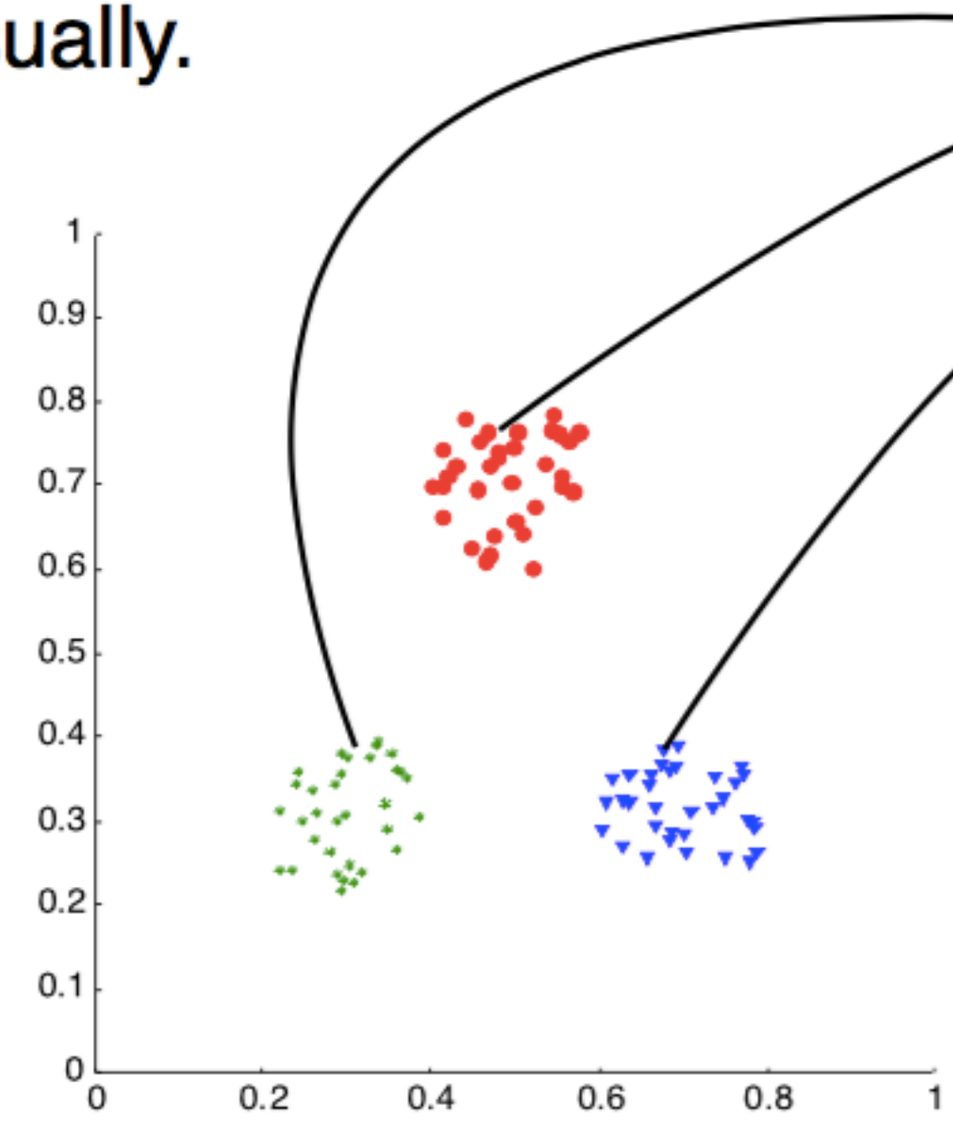
$$D_{\text{means}}(C_i, C_j) = \|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|$$

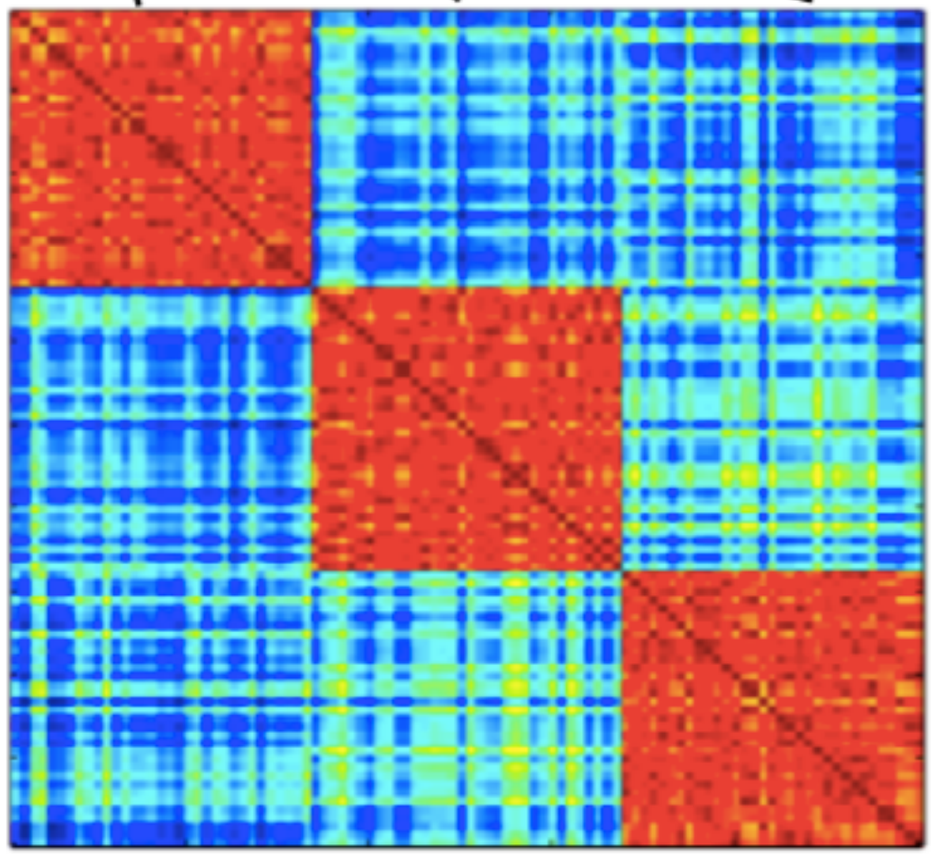


Cluster Validation

THE UNIVERSITY OF SYDNEY

Similarity matrix with respect to cluster labels and inspect visually.

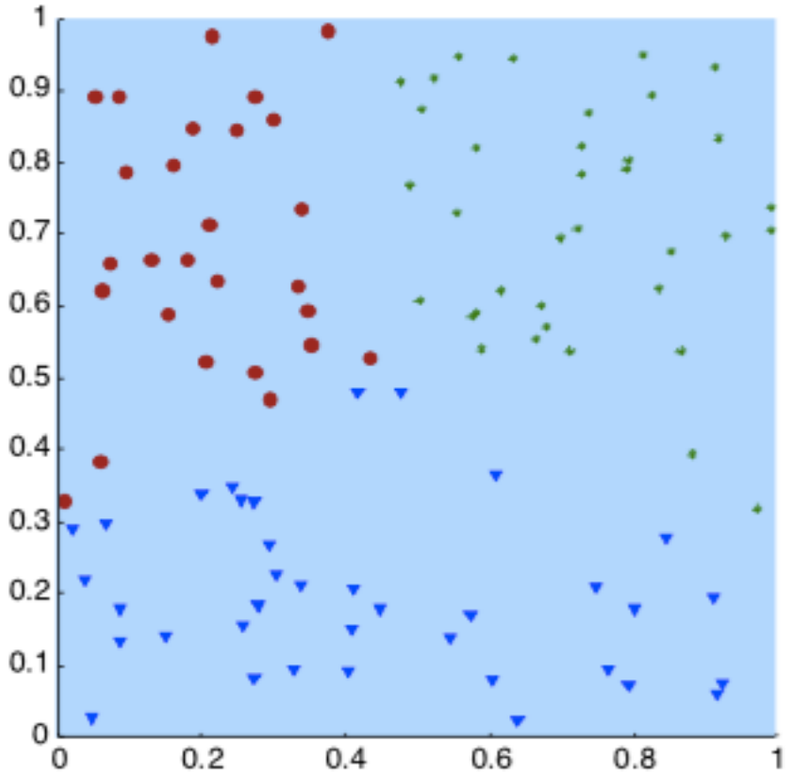


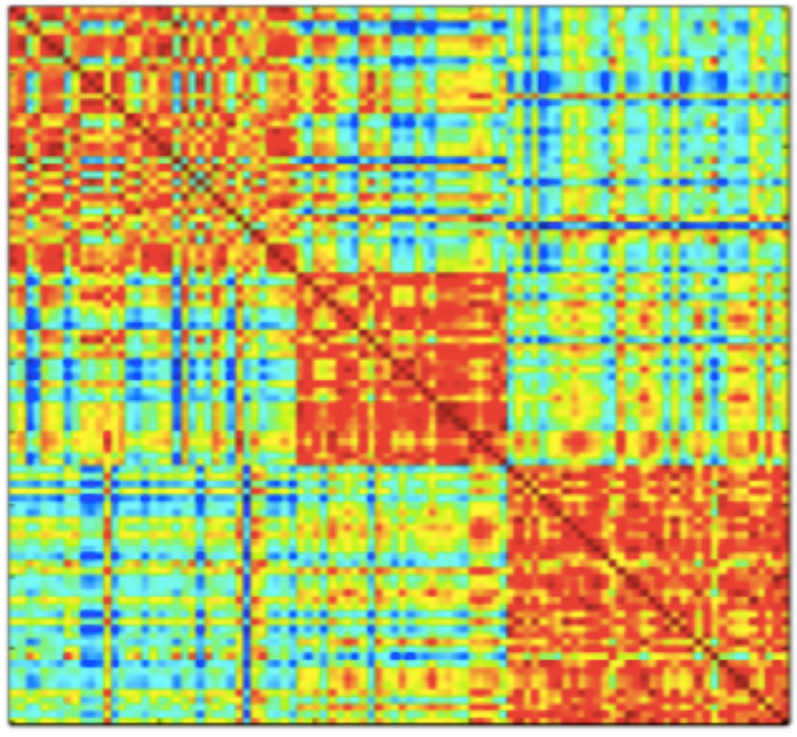


Cluster Validation

THE UNIVERSITY OF SYDNEY

Random data clusters are not well defined.





THE UNIVERSITY OF SYDNEY

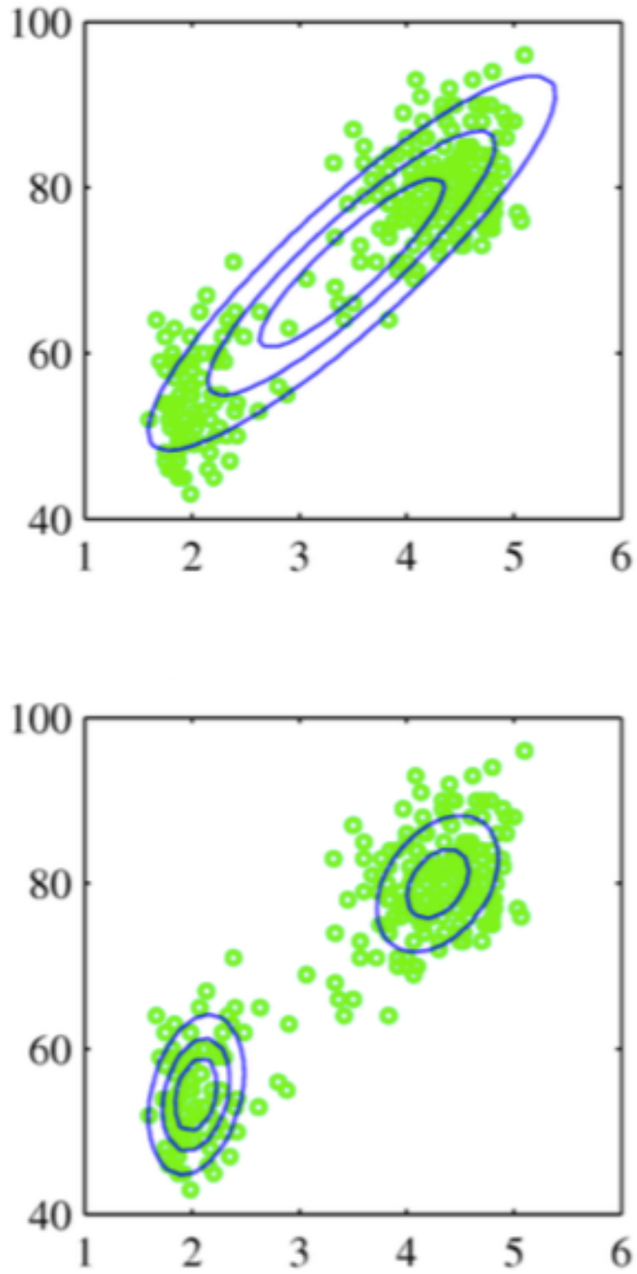
Probabilistic Approach to Clustering

Mixture of Gaussians

Gaussian mixture distribution with K components.

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

$$0 \leq \pi_k \leq 1 \qquad \sum_{k=1}^{K} \pi_k = 1$$



Mixture models provide a probabilistic framework for clustering.

Mixture of Gaussians

THE UNIVERSITY OF SYDNEY

Let us introduce a latent random variable

$$\mathbf{z} = \{z_k\}_{k \in 1,\ldots,K} \qquad z_k \in \{0, 1\} \qquad \sum_{k=1}^{K} z_k = 1$$

$\mathbf{z}$ has K possible states.

$$p(z_k = 1) = \pi_k \qquad p(\mathbf{z}) = \prod_{k=1}^{K} \pi_k^{z_k}$$

$$p(\mathbf{x}|z_k = 1) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \qquad p(\mathbf{x}|\mathbf{z}) = \prod_{k=1}^{K} \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)^{z_k}$$

$$p(\mathbf{x}) = \sum_{\mathbf{z}} p(\mathbf{x}|\mathbf{z})p(\mathbf{z}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

Mixture of Gaussians

Let us apply Bayes theorem and infer the value of the latent variable.

$$p(\mathbf{z}|\mathbf{x}) = \frac{p(\mathbf{x}|\mathbf{z})p(\mathbf{z})}{p(\mathbf{x})}$$

$$\gamma(z_k) \equiv p(z_k = 1|\mathbf{x}) = \frac{p(z_k = 1)p(\mathbf{x}|z_k = 1)}{p(\mathbf{x})}$$

$$= \frac{\pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}$$

$\gamma(z_k)$ is called *responsibility* that component k takes for explaining $\mathbf{x}$.

π_k is the prior probability of component k.
$\gamma(z_k)$ is the posterior probability after $\mathbf{x}$ is observed.

THE UNIVERSITY OF SYDNEY

Expectation Maximisation

EM for Gaussian Mixtures

Expectation Maximisation (EM)

THE UNIVERSITY OF SYDNEY

Elegant and powerful method for finding MLE or MAP solutions for models with latent variables.

Intuition: If we knew what cluster each point belonged to (i.e. the z variables), we could partition the data and find the MLE for each cluster separately.

EM for Gaussian Mixtures

Likelihood function:

$$\ln p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right\}$$

Conditions to be satisfied at maximum likelihood:

$$0 = \frac{\partial \ln p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma})}{\partial \boldsymbol{\mu}_k}$$

$$0 = -\sum_{n=1}^{N} \underbrace{\frac{\pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_j \pi_j \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}}_{\gamma(z_{nk})} \boldsymbol{\Sigma}_k (\mathbf{x}_n - \boldsymbol{\mu}_k)$$

$$\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \mathbf{x}_n \qquad N_k = \sum_{n=1}^{N} \gamma(z_{nk})$$

N_k is the effective number of points assigned to cluster k.

EM for Gaussian Mixtures

THE UNIVERSITY OF SYDNEY

$$0 = \frac{\partial \ln\, p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma})}{\partial \boldsymbol{\Sigma}_k}$$

$$\Rightarrow \quad \boldsymbol{\Sigma}_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk})(\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^{\mathrm{T}}$$

$$0 = \frac{\partial \ln\, p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma})}{\partial \pi_k}$$

$$\Rightarrow \quad \pi_k = \frac{N_k}{N}$$

EM Algorithm

THE UNIVERSITY OF SYDNEY

1 Initialise means $\boldsymbol{\mu}_k$, covariances $\boldsymbol{\Sigma}_k$ and mixing coefficients π_k.

2 E-step

$$\gamma(z_{nk}) = \frac{\pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}$$

3 M-step

$$\boldsymbol{\mu}_k^{\text{new}} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \mathbf{x}_n$$

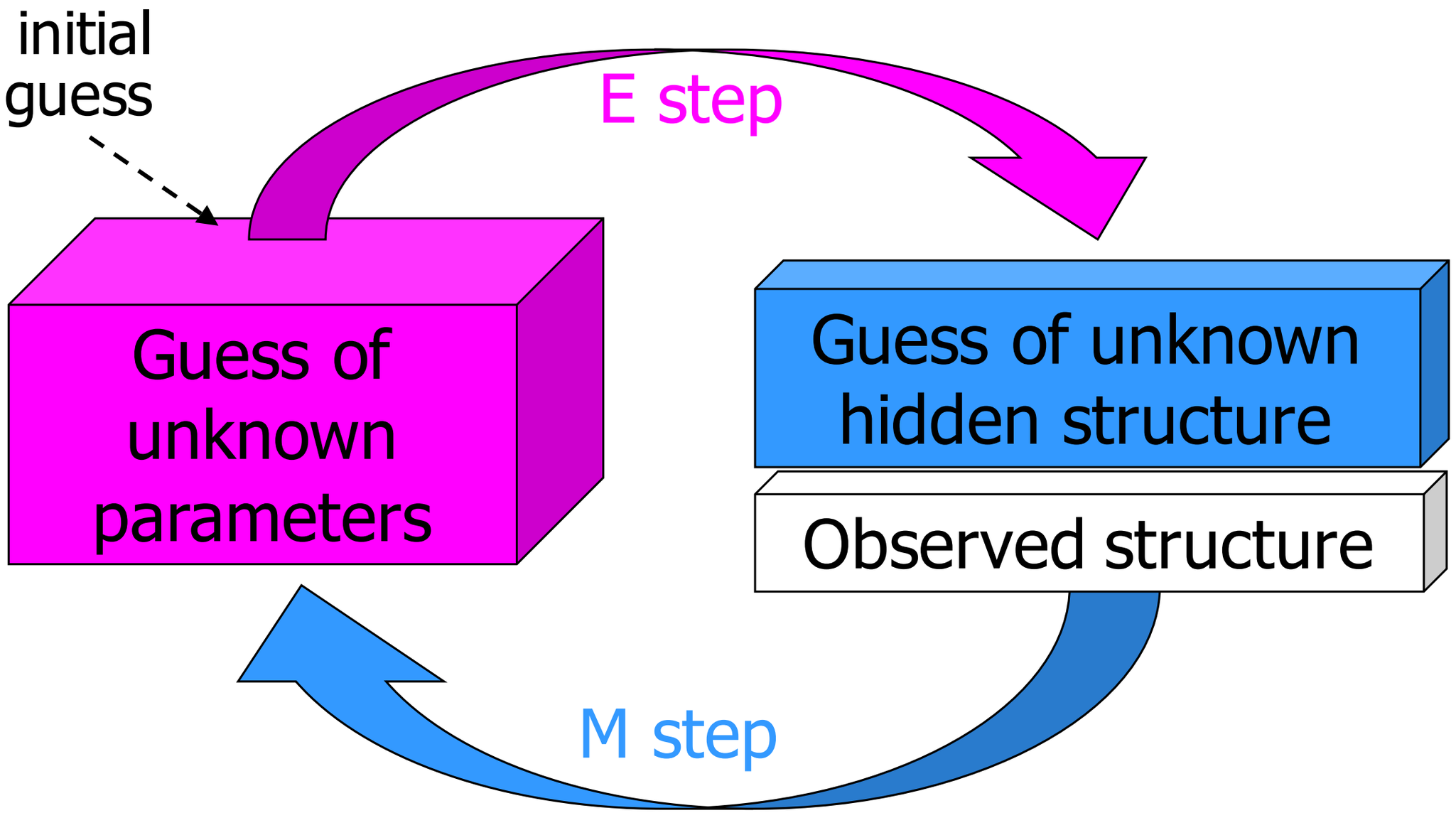
$$\boldsymbol{\Sigma}_k^{\text{new}} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \left(\mathbf{x}_n - \boldsymbol{\mu}_k^{\text{new}}\right) \left(\mathbf{x}_n - \boldsymbol{\mu}_k^{\text{new}}\right)^{\mathrm{T}}$$

$$\pi_k^{\text{new}} = \frac{N_k}{N}$$

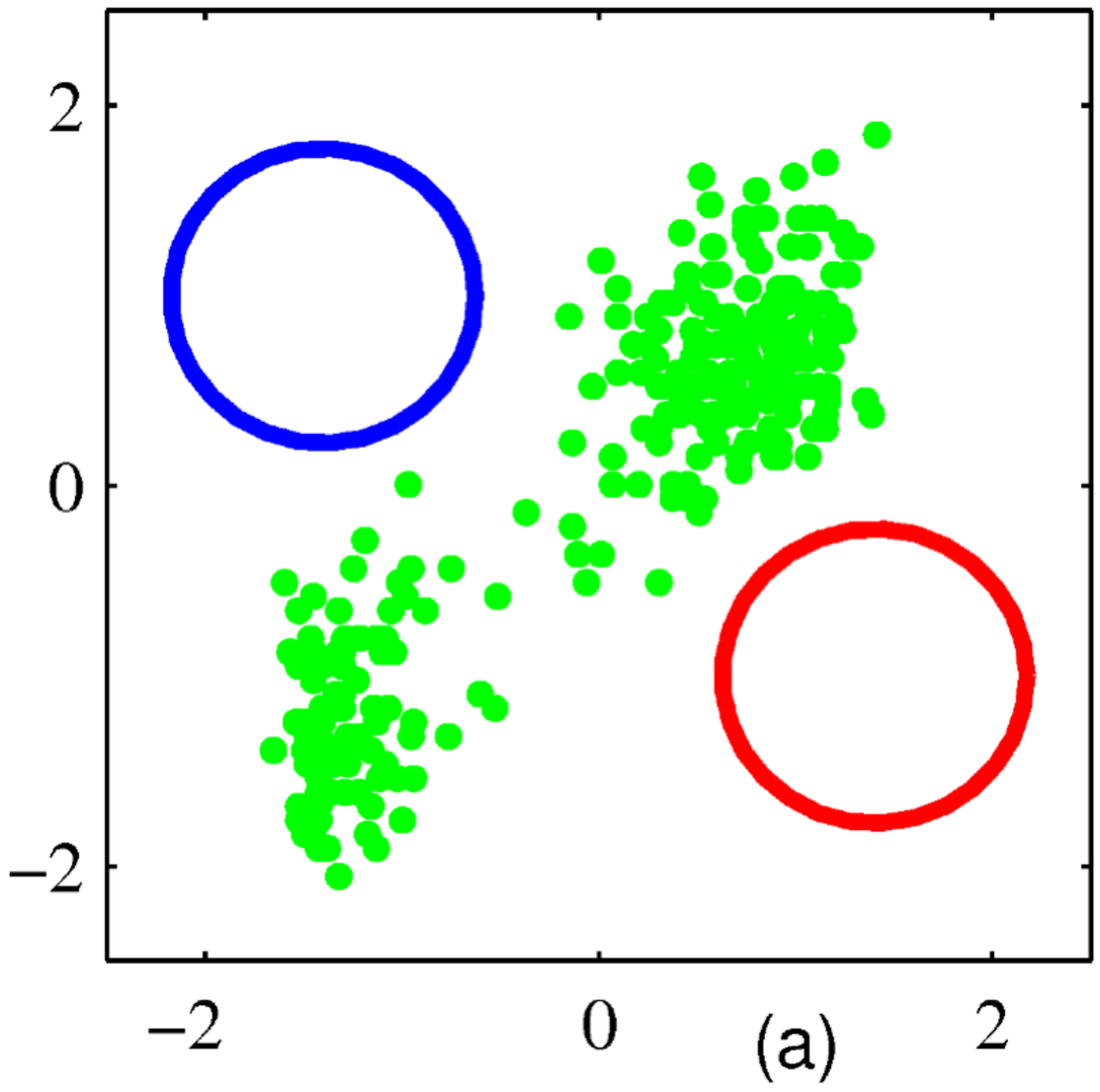
4 Eval Likelihood

$$\ln p(\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right\}$$

EM Algorithm



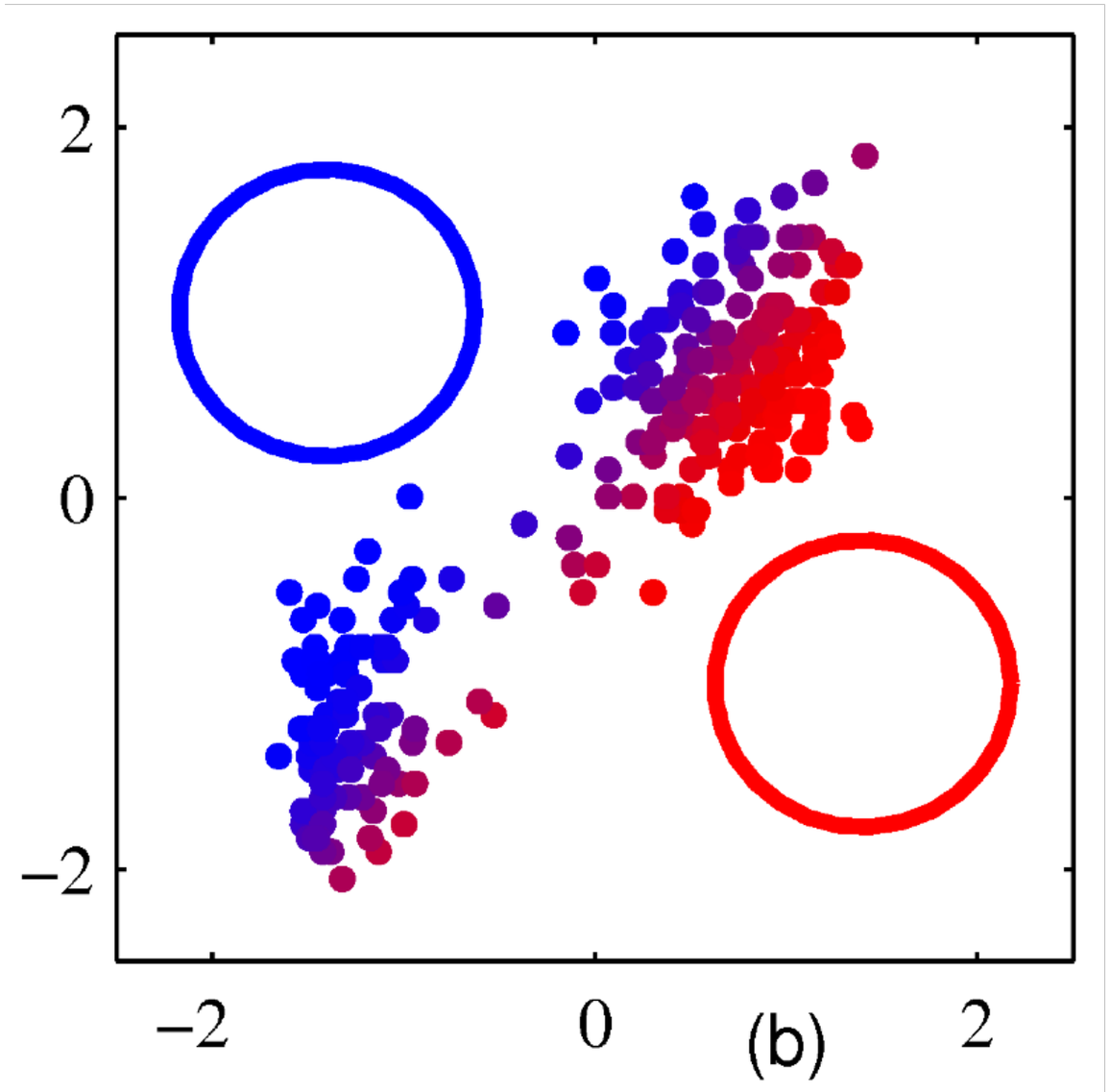
EM Example



Initial values for mean vectors (same as K-means example).

Diagonal covariance matrices (showing one std contour).

EM Example

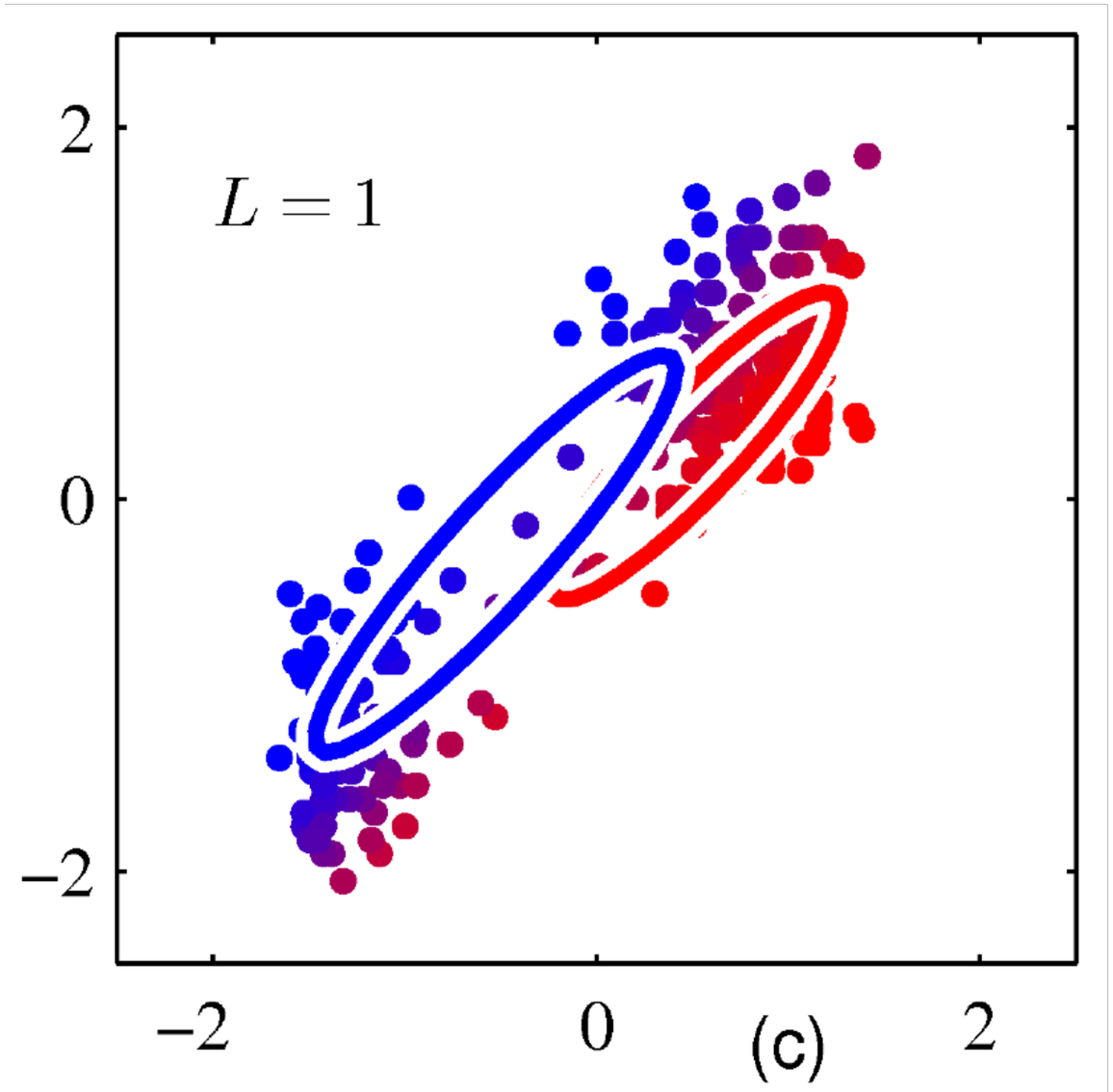


Initial E step.

$$\gamma(z_{nk}) = \frac{\pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}$$

Colour
proportional to responsibilities.

EM Example



M Step:

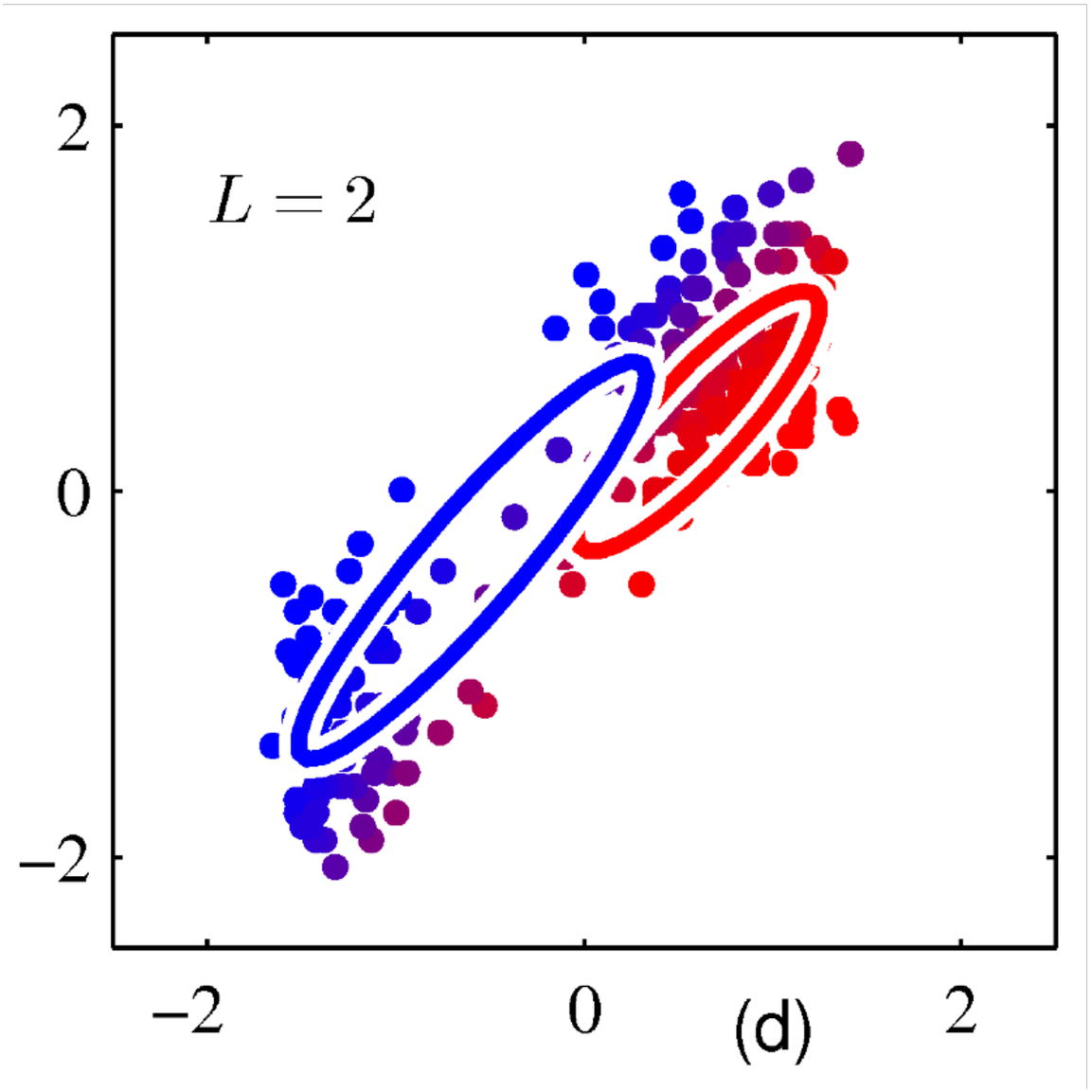
The means move towards the weighted average of dataset with respective ink colour (responsibilities).

$$\boldsymbol{\mu}_k^{\text{new}} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \mathbf{x}_n$$

The covariance matrices adapt to the covariance of the respective ink.

$$\boldsymbol{\Sigma}_k^{\text{new}} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \left(\mathbf{x}_n - \boldsymbol{\mu}_k^{\text{new}}\right) \left(\mathbf{x}_n - \boldsymbol{\mu}_k^{\text{new}}\right)^{\text{T}}$$

EM Example



M Step:

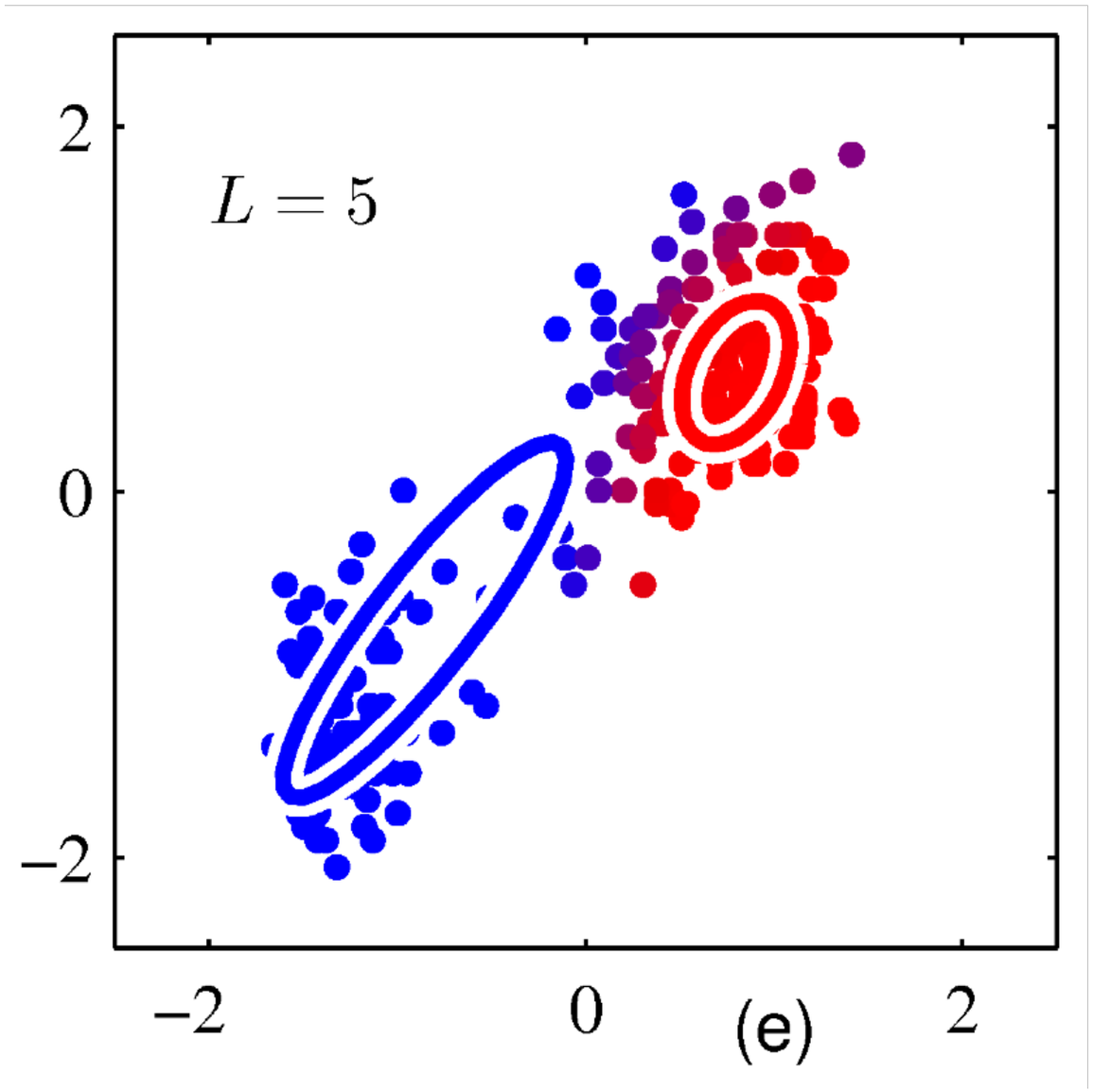
The means move towards the weighted average of dataset with respective ink colour (responsibilities).

$$\boldsymbol{\mu}_k^{\text{new}} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \mathbf{x}_n$$

The covariance matrices adapt to the covariance of the respective ink.

$$\boldsymbol{\Sigma}_k^{\text{new}} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \left(\mathbf{x}_n - \boldsymbol{\mu}_k^{\text{new}}\right) \left(\mathbf{x}_n - \boldsymbol{\mu}_k^{\text{new}}\right)^{\text{T}}$$

EM Example



M Step:

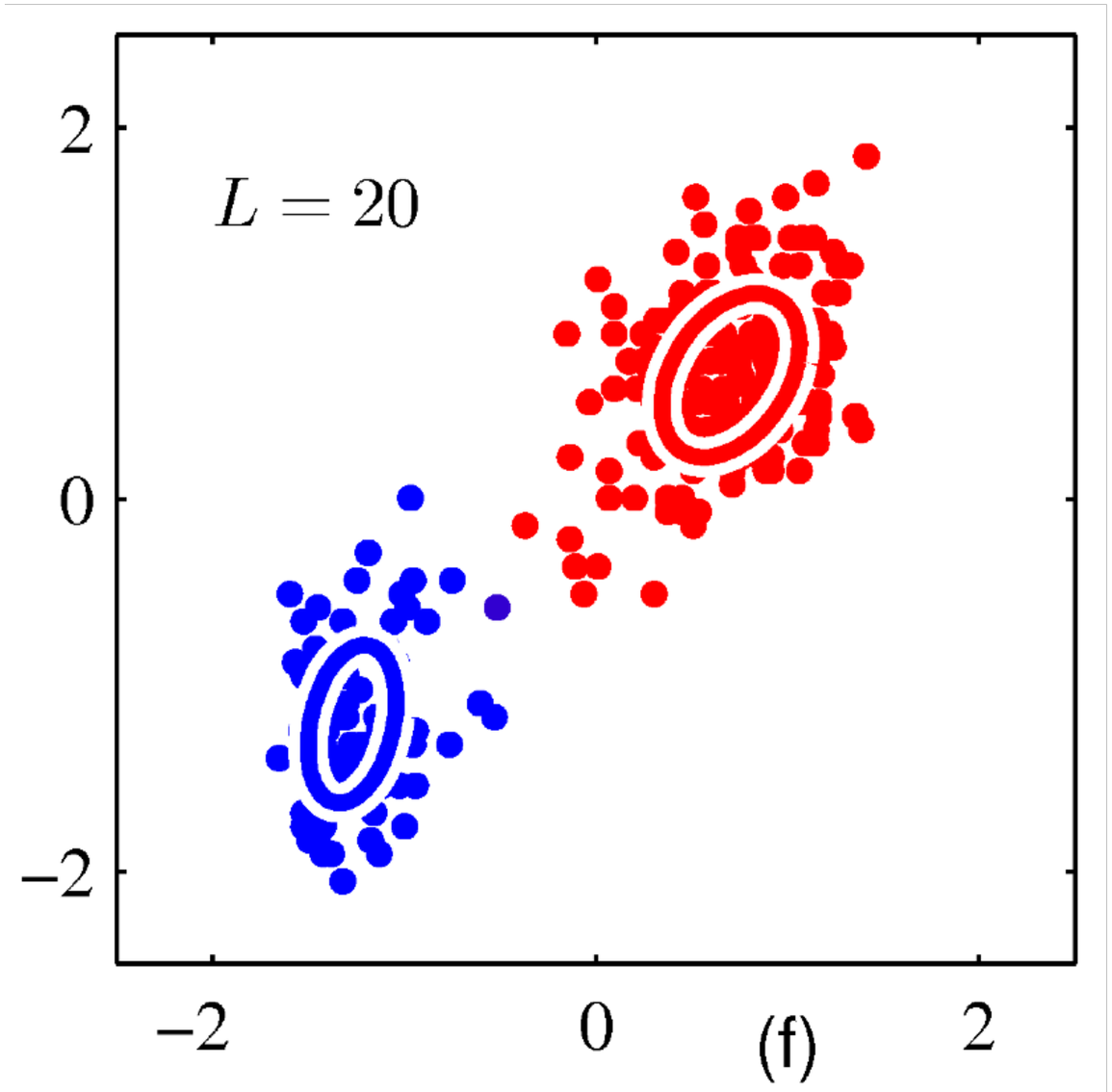
The means move towards the weighted average of dataset with respective ink colour (responsibilities).

$$\boldsymbol{\mu}_k^{\text{new}} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \mathbf{x}_n$$

The covariance matrices adapt to the covariance of the respective ink.

$$\boldsymbol{\Sigma}_k^{\text{new}} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \left(\mathbf{x}_n - \boldsymbol{\mu}_k^{\text{new}}\right) \left(\mathbf{x}_n - \boldsymbol{\mu}_k^{\text{new}}\right)^{\text{T}}$$

EM Example



M Step:

The means move towards the weighted average of dataset with respective ink colour (responsibilities).

$$\boldsymbol{\mu}_k^{\text{new}} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \mathbf{x}_n$$

The covariance matrices adapt to the covariance of the respective ink.

$$\boldsymbol{\Sigma}_k^{\text{new}} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \left(\mathbf{x}_n - \boldsymbol{\mu}_k^{\text{new}}\right) \left(\mathbf{x}_n - \boldsymbol{\mu}_k^{\text{new}}\right)^{\text{T}}$$

Clustering