University of Nottingham
UK | CHINA | MALAYSIA

COMP3055
Machine Learning

Topic 8 – Data Clustering

Ying Weng
2024 Autumn

Supervised VS Unsupervised Learning

- Supervised learning
  - Learns a function that maps an input to an output based on example input-output pairs.
  - Training data is labeled.
- Unsupervised learning
  - Learns from test data that has not been labeled.
  - Learns relationships between elements in a data set and classify the raw data without "help.”
  - Typical application includes data clustering.

Motivating Problems

- A true colour image – 24bits/pixel, R – 8 bits, G – 8 bits, B – 8 bits

16777216 possible colours

- A gif image - 8bits/pixel

256 possible colours

Motivating Problems

(r, g, b)

For each pixel in the original image

Find the closest colour in the Colour Table

r_0	g_0	b_0
r_1	g_1	b_1
⋮		
r_{255}	g_{255}	b_{255}

Colour Table

Record the index of that colour (for storage or transmission)

To reconstruct the image, place the indexed colour from the Colour Table at the corresponding spatial location

Motivating Problems

(r, g, b)

For each pixel in the original image

Find the closest colour in the Colour Table

r_0	g_0	b_0
r_1	g_1	b_1
⋮		
r_{255}	g_{255}	b_{255}

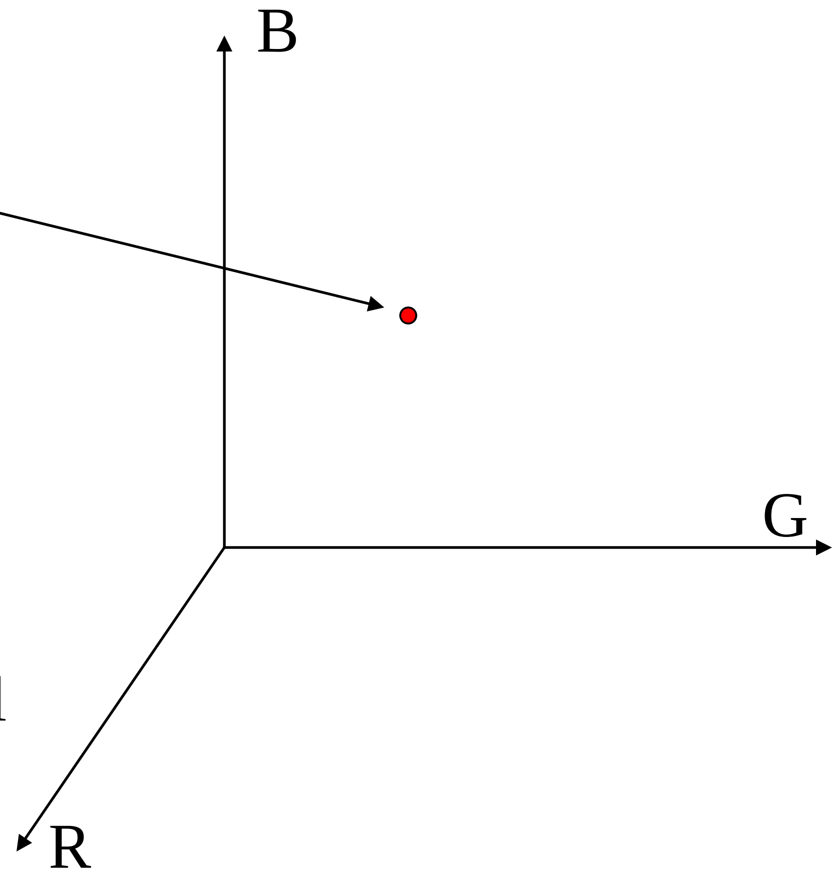
Colour Table

How to choose the colours in the table?

... the
indexed colour from the Colour Table at the corresponding spatial location

Motivating Problems

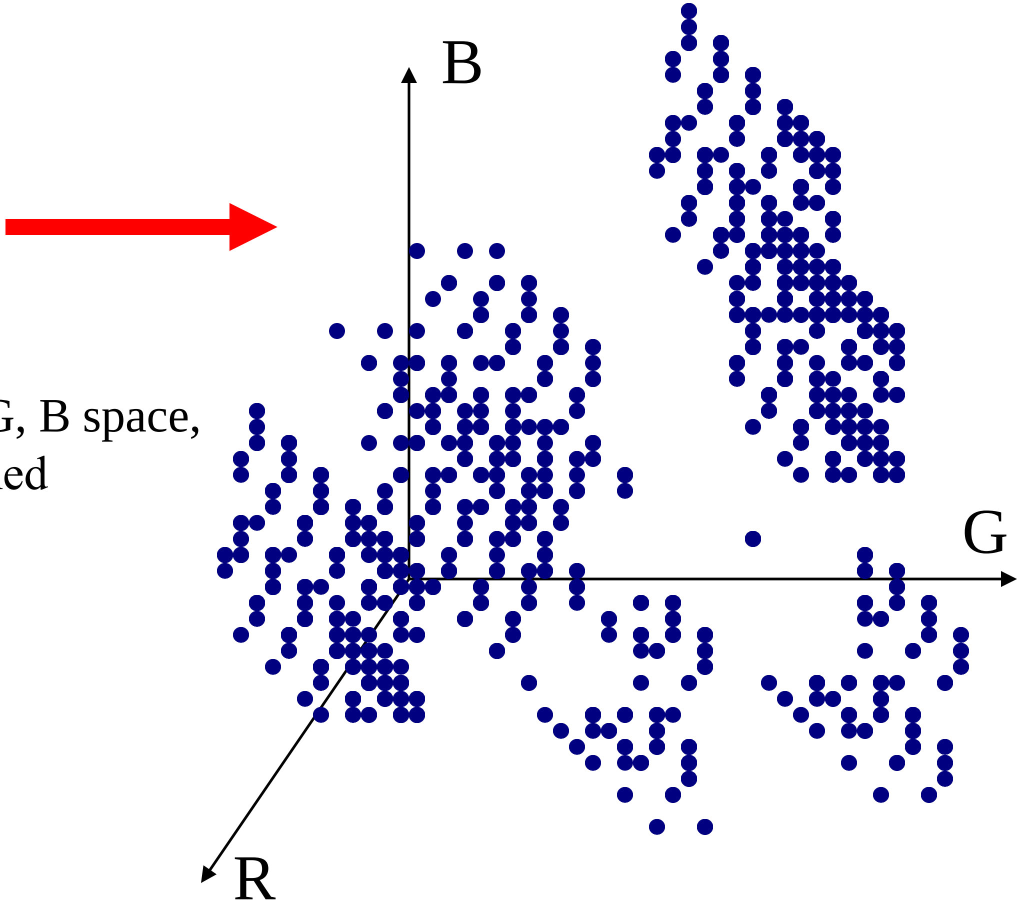
(r, g, b)



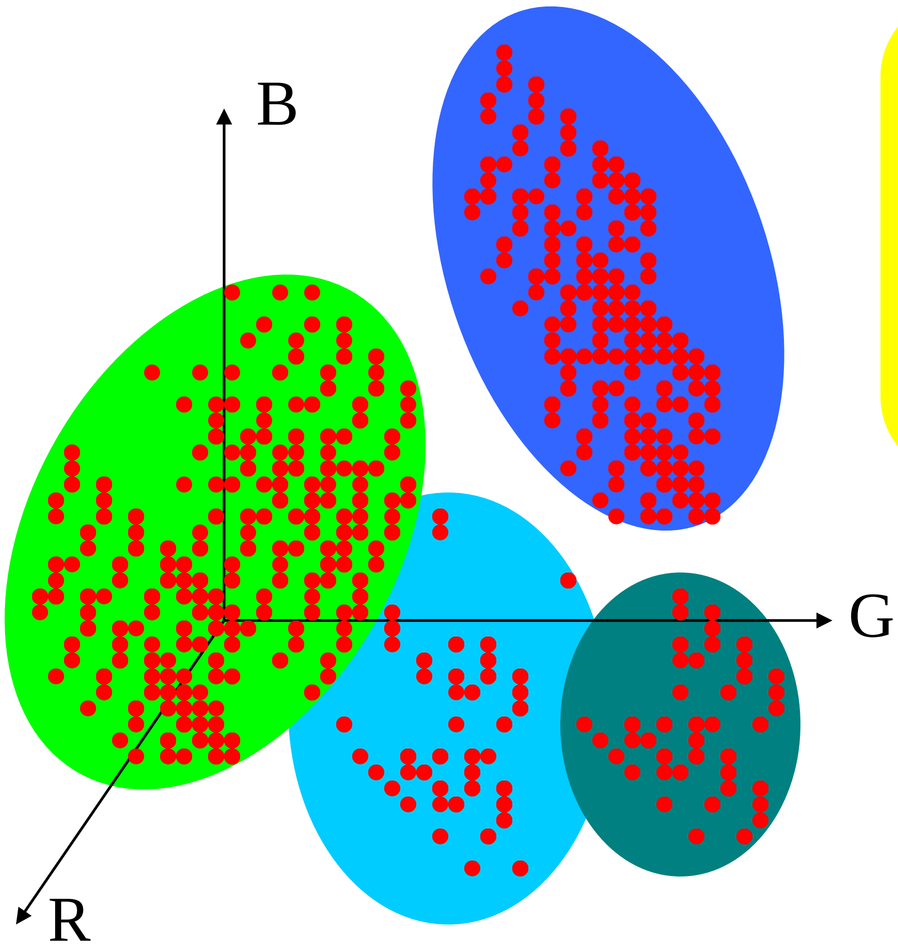
A pixel corresponding to
A point in the 3 dimensional
R, G, B space

Motivating Problems

Map all pixels into the R, G, B space, "clouds" of pixels are formed

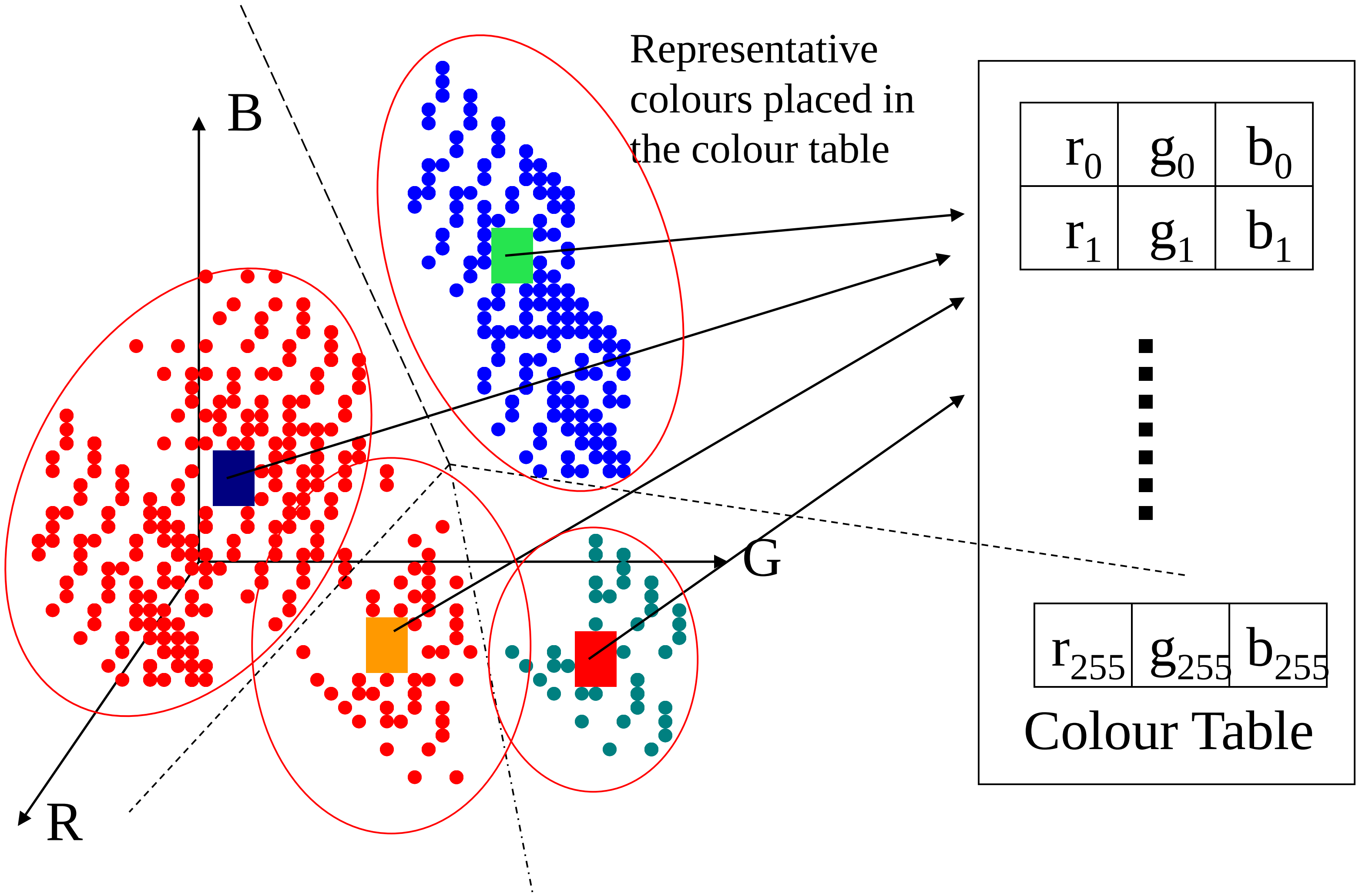


Motivating Problems



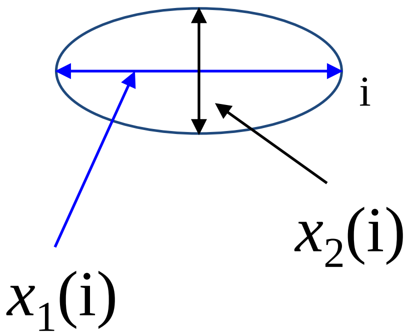
Group pixels that are close to each other, and replace them by one single colour

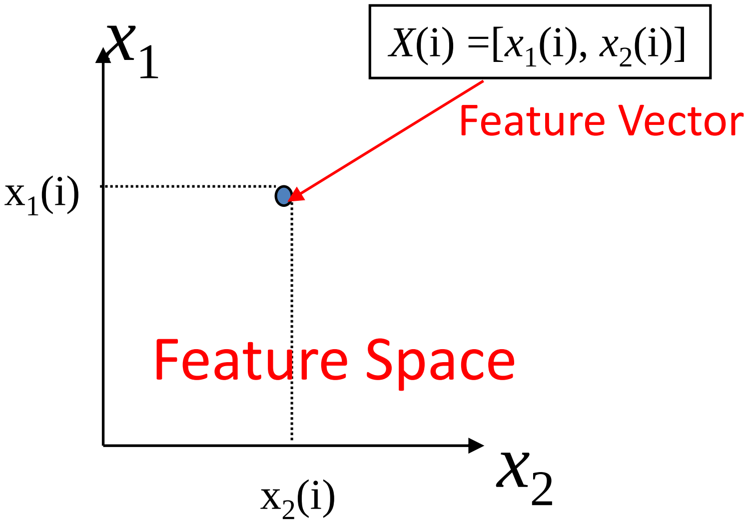
Motivating Problems

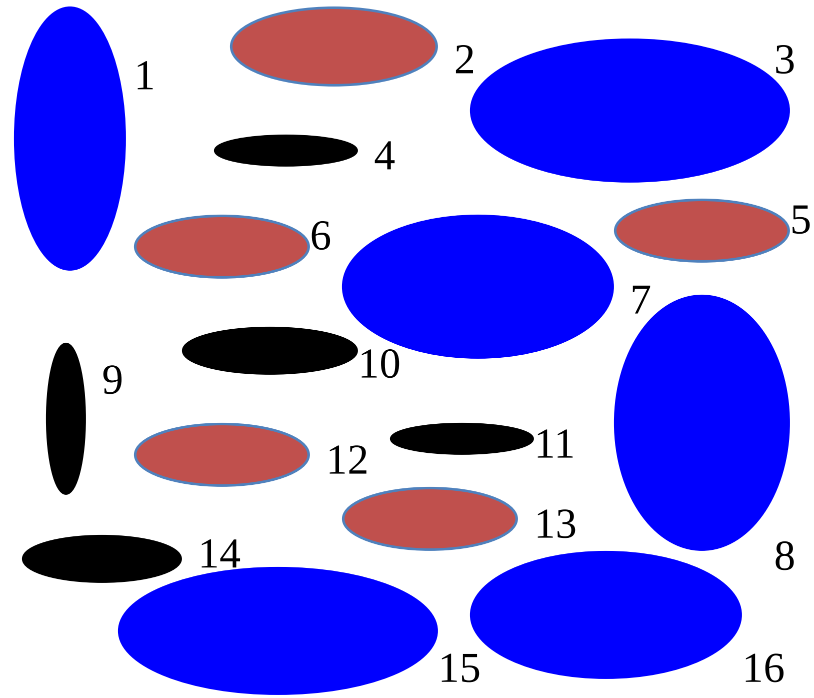


Motivating Example

- Classify objects (Oranges, Potatoes) into large, middle, small sizes



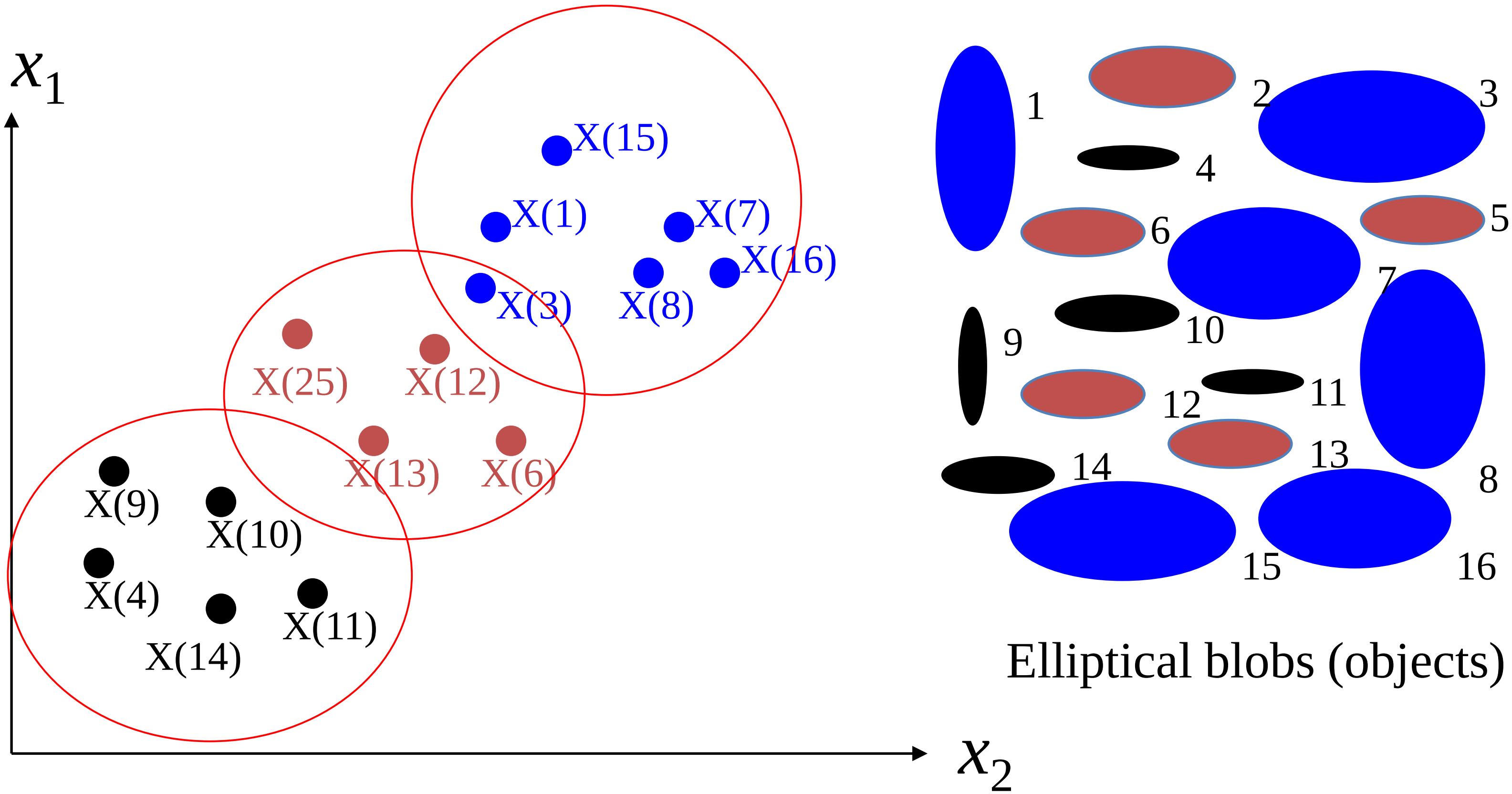




Elliptical blobs (objects)

Motivating Example

- From Objects to Feature Vectors to Points in the Feature Space



Elliptical blobs (objects)

Motivation of Clustering

- Patterns within a valid cluster are *more similar to each other* than they are to a pattern belonging to a different cluster.

- In clustering, the problem is to group a given collection of *unlabeled patterns* into meaningful clusters. Clustering is data *driven method*, the clusters are obtained solely from the data.

- Clustering could be used in the field of pattern-analysis, grouping, decision-making, and machine-learning situations, including data mining, document retrieval, image segmentation

K-Means

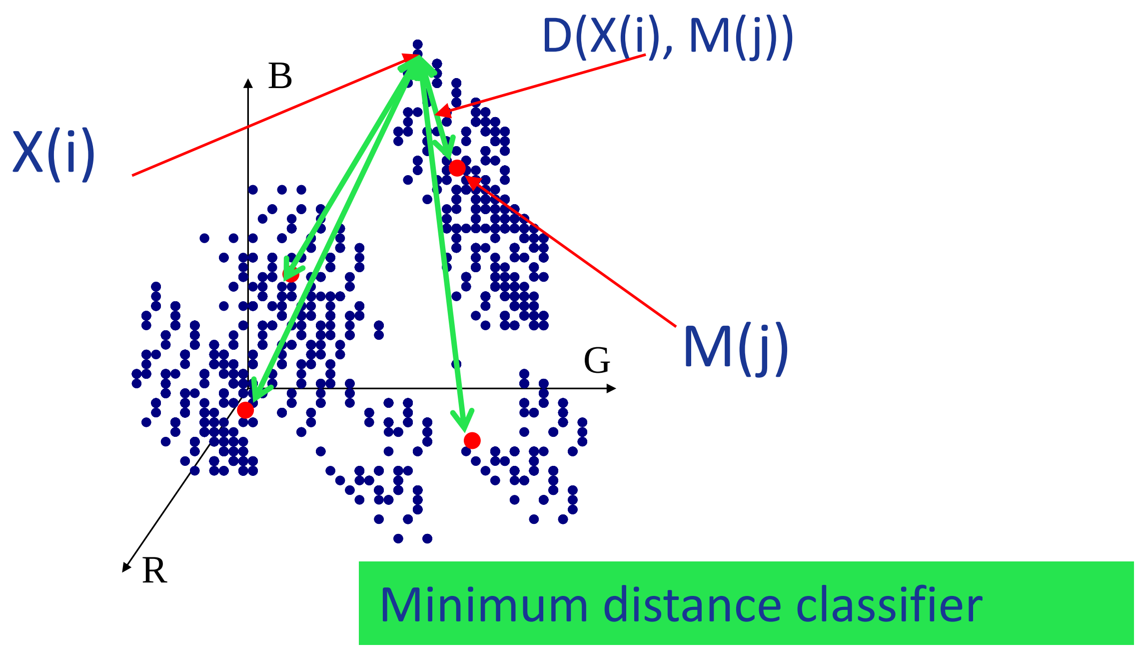
- An algorithm for partitioning (or clustering) N data points into K disjoint subsets S_j containing N_j data points
  - Define, $X(i) = [x_1(i), x_2(i), \ldots, x_n(i)]$, $i = 1, 2, \ldots N$, as N data points
  - We want to cluster these N points into K subsets, or K clusters, where K is pre-set
  - For each cluster, we define $M(j) = [m_1(j), m_2(j), \ldots, m_n(j)]$, $j=1, 2, \ldots K$, as its prototype or cluster centroids
  - Define the distance between data point X(i) and cluster prototype M(j) as

$$D\big(X(i), M(j)\big) = \left\| X(i) - M(j) \right\| = \sqrt{\sum_{l=1}^{n} \big(x_l(i) - m_l(j)\big)^2}$$

K-Means

- A data point X(i) is assigned to the jth cluster, C(j), $X(i) \in C(j)$, if following condition holds

$$D\big(X(i), M(j)\big) \leq D\big(X(i), M(l)\big) \quad \text{for all } l = 1, 2, \ldots, k$$



K-Means Algorithm

Step 1

- ✪ Arbitrarily choose from the given sample set k initial cluster centres,

$M^{(0)}(j) = [m^{(0)}_1(j), m^{(0)}_2(j), ,\ldots, m^{(0)}_n(j)]$ $j = 1, 2, \ldots, K$,

e.g., the first K samples of the sample set
or can also be generated randomly

Set $t = 0$ (t is the iteration index)

K-Means Algorithm

Step 2

- Assign each of the samples X(i) = [x_1(i), x_2(i), …, x_n(i)], i = 1, 2, ….N, to one of the clusters according to the distance between the sample and the centre of the cluster:

$$
\begin{aligned}
&X(i) \in C^{(t)}(j) \\
&\text{if} \quad D\left(X(i), M^{(t)}(j)\right) \le D\left(X(i), M^{(t)}(l)\right) \\
&\text{for all } l = 1, 2, \ldots, k
\end{aligned}
$$

K-Means Algorithm

Step 3

Update the cluster centres to get

$M^{(t+1)}(j) = [m^{(t+1)}_1(j), m^{(t+1)}_2(j), ,\ldots, m^{(t+1)}_n(j)]$; $j = 1, 2, \ldots, K$

according to

$$M^{(t+1)}(j) = \frac{1}{N_j^{(t)}} \sum_{X(i) \in C^{(t)}(j)} X(i)$$

$N^{(t)}_j$ is the number of samples in $C^{(t)}_j$

K-Means Algorithm

Step 4

– Calculate the error of approximation

$$E(t)=\sum_{j=1}^{K}\sum_{X(i)\in C^{(t)}(j)}\left\|X(i)-M^{(t)}(j)\right\|$$

K-Means Algorithm

Step 5

- If the terminating criterion is met, then stop, otherwise

  Set $t = t+1$

  Go to Step 2.

K-Means Algorithm

Stopping criterions

- The K-means algorithm can be stopped based on following criterions

1. The errors do not change significantly in two consecutive epochs

   $|E(t)-E(t-1)|<\varepsilon$, where ε is some preset small value

2. No further change in the assignment of the data points to clusters in two consecutive epochs.

3. It can also stop after a fixed number of epochs regardless of the error

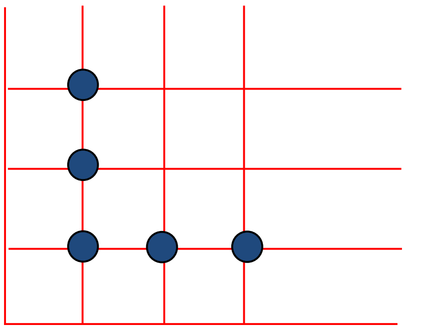
K-Means Algorithm

A worked example to see how it works exactly

Five 2-dimensional data points:

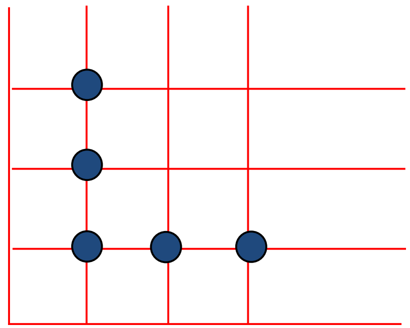
(1, 1), (2, 1), (3, 1), (1, 2), (1, 3)

Cluster them into two clusters and find the cluster centres



K-Means Algorithm

A worked example to see how it works exactly



(1) Euclidean distance to $m_1^0(1,2)$

$\sqrt{(1-1)^2+(1-2)^2}=1$

$\sqrt{(2-1)^2+(1-2)^2}=\sqrt{2}=1.41$

$\sqrt{(3-1)^2+(1-2)^2}=\sqrt{5}=2.24$

$\sqrt{(1-1)^2+(2-2)^2}=0$

$\sqrt{(1-1)^2+(3-2)^2}=1$

Euclidean distance to $m_2^0(3,1)$

$\sqrt{(1-3)^2+(1-1)^2}=2$ $\quad\therefore C_1$

$\sqrt{(2-3)^2+(1-1)^2}=1$ $\quad\therefore C_2$

$\sqrt{(3-3)^2+(1-1)^2}=0$ $\quad\therefore C_2$

$\sqrt{(1-3)^2+(2-1)^2}=\sqrt{5}=2.24$ $\quad\therefore C_1$

$\sqrt{(1-3)^2+(3-1)^2}=\sqrt{8}=2.83$ $\quad\therefore C_1$

C_1 class: $(1,1),(1,2),(1,3) \Longrightarrow m_1^{(0+1)}$:

$$\frac{1}{3}\sum_{i=1}^{3}x_i=\frac{1+1+1}{3}=1$$

$$\frac{1}{3}\sum_{i=1}^{3}y_i=\frac{1+2+3}{3}=2$$

C_2 class: $(2,1),(3,1) \Longrightarrow m_2^{(0+1)}$:

$$\frac{1}{2}\sum_{i=1}^{2}x_i=\frac{2+3}{2}=2.5$$

$$\frac{1}{2}\sum_{i=1}^{2}y_i=\frac{1+1}{2}=1$$

(2) Euclidean distance to $m_1^1(1,2)$

1

$\sqrt{2}=1.41$

$\sqrt{5}=2.24$

0

1

Euclidean distance to $m_2^1(2.5,1)$

$\sqrt{(1-2.5)^2+(1-1)^2}=1.5$ $\quad\therefore C_1$

$\sqrt{(2-2.5)^2+(1-1)^2}=0.5$ $\quad\therefore C_2$

$\sqrt{(3-2.5)^2+(1-1)^2}=0.5$ $\quad\therefore C_2$

$\sqrt{(1-2.5)^2+(2-1)^2}=\sqrt{3.25}=1.80$ $\quad\therefore C_1$

$\sqrt{(1-2.5)^2+(3-1)^2}=\sqrt{6.25}=2.5$ $\quad\therefore C_1$

$\therefore C_1$ class: $(1,1),(1,2),(1,3)$

C_2 class: $(2,1),(3,1)$

K-Means Algorithm

- What is the algorithm doing exactly?
  - It tries to find the centre vectors M(j)'s that optimize the following cost function

$$E = \sum_{j=1}^{K} \sum_{X(i) \in C(j)} \left\| X(i) - M(j) \right\|$$

K-Means Algorithm

- Some remarks
  - Is a gradient descent algorithm, trying to minimize a cost function *E*
  - In general, the algorithm does not achieve a global minimum of E over the assignments
  - Sensitive to initial choice of cluster centers. Different starting cluster centroids may lead to different solution
  - Is a popular method, many more advanced methods derived from this simple algorithm

Any Questions?