Week-3

1. Example

Consider the following dataset in $\mathbb{R}^2$:

$$X = \begin{bmatrix} 0 & 0 & 1 & 1 & 3 & 3 & 5 & 5 \\ 0 & 1 & 0 & 1 & 1 & 2 & 1 & 2 \end{bmatrix}$$

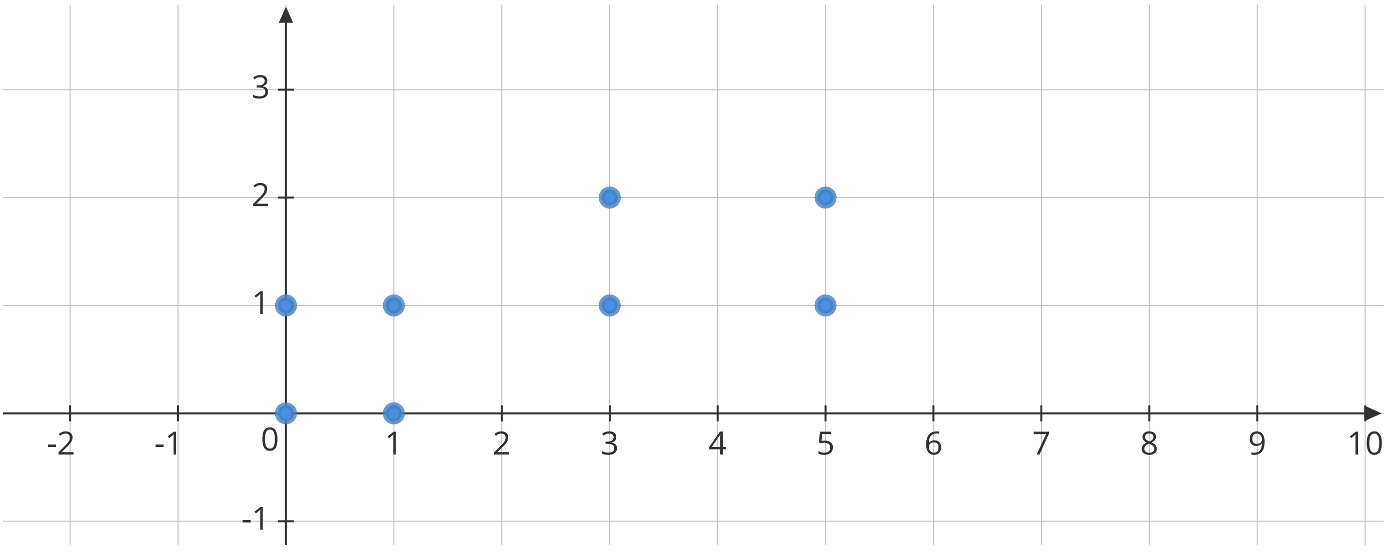
Unsupervised learning

- Representation learning
  - PCA
  - Kernel PCA
- Clustering
  - Lloyd's algorithm (k-means clustering)

1.1. Visualize the dataset

$$X = \begin{bmatrix} 0 & 0 & 1 & 1 & 3 & 3 & 5 & 5 \\ 0 & 1 & 0 & 1 & 1 & 2 & 1 & 2 \end{bmatrix}$$

Shape of the dataset is $d \times n$, where $d = 2$ and $n = 8$.



$k = 2$ is a good choice for this problem.

1.2. How many cluster assignments are possible with k means and n data-points?

$$k \times \cdots \times k = k^n$$

A sample cluster assignment:

$$z = \begin{bmatrix} 1 & 1 & 2 & 2 & 1 & 1 & 2 & 2 \end{bmatrix}$$

1.3. Run k-means with $k = 2$ and $z_0 = \begin{bmatrix} 1 & 2 & 2 & 2 & 2 & 2 & 2 & 2 \end{bmatrix}$. Plot the Voronoi regions. To which cluster does $(2, 2)$ belong? Find the value of the objective function at the end.

$$X = \begin{bmatrix} 0 & 0 & 1 & 1 & 3 & 3 & 5 & 5 \\ 0 & 1 & 0 & 1 & 1 & 2 & 1 & 2 \end{bmatrix}$$

Step-0: Initialization

$$z_0 = \begin{bmatrix} 1 & 2 & 2 & 2 & 2 & 2 & 2 & 2 \end{bmatrix}^T$$

$$\mu_1^0 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$\mu_2^0 = \frac{1}{7}\begin{bmatrix} 18 \\ 8 \end{bmatrix}$$

$$= \frac{2}{7}\begin{bmatrix} 9 \\ 4 \end{bmatrix}$$

$$= \begin{bmatrix} 2.57 \\ 1.14 \end{bmatrix}$$

Step-1: First iteration of k-means

Step-1.1: Compute the cluster assignments (Computing z)

x_i	z_t	$\|\|x_i - (0,0)\|\|^2$	$\|\|x_i - (2.57, 1.14)\|\|^2$	z_{t+1}
$(0,0)$	1	smaller		1
$(0,1)$	2	$\|\|(0,1) - (0,0)\|\|^2 = 1$ smaller	$\|\|(0,1) - (2.57, 1.14)\|\|^2 = 2.57^2 + 0.14^2$	1
$(1,0)$	2	smaller		1
$(1,1)$	2	$\|\|(1,1) - (0,0)\|\|^2 = 2$ smaller	$1.57^2 + 0.14^2 > 2$	1
$(3,1)$	2	$3^2 + 1^2 = 10$	$0.43^2 + 0.14^2$ smaller	2
$(3,2)$	2	$3^2 + 2^2 = 13$	$0.43^2 + 0.86^2$ smaller	2
$(5,1)$	2		smaller	2
$(5,2)$	2		smaller	2

$$z_1 = \begin{bmatrix} 1 & 1 & 1 & 1 & 2 & 2 & 2 & 2 \end{bmatrix}$$

Step-1.2: Compute the cluster means (Computing μ)

$$X = \begin{bmatrix} 0 & 0 & 1 & 1 & 3 & 3 & 5 & 5 \\ 0 & 1 & 0 & 1 & 1 & 2 & 1 & 2 \end{bmatrix}$$

$$\mu_1 = \frac{1}{4}\begin{bmatrix} 2 \\ 2 \end{bmatrix}$$

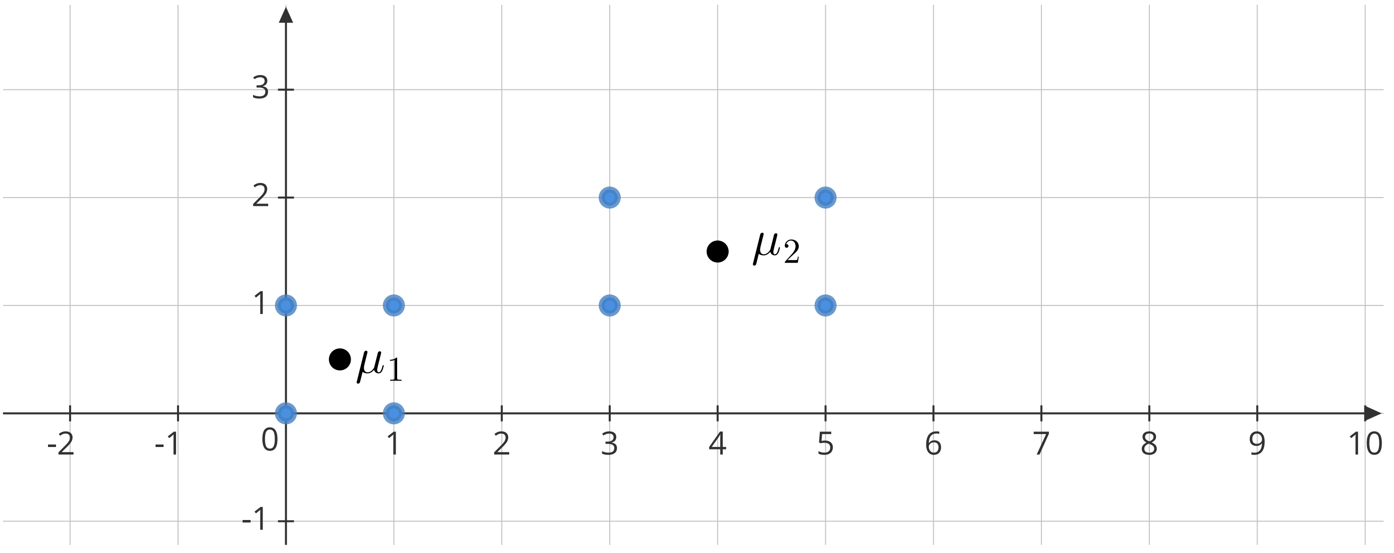
$$= \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix}$$

$$\mu_2 = \frac{1}{4}\begin{bmatrix} 16 \\ 6 \end{bmatrix}$$

$$= \begin{bmatrix} 4 \\ 1.5 \end{bmatrix}$$

At the of the first iteration, $\mu_1 = (0.5, 0.5)$, $\mu_2 = (4, 1.5)$.

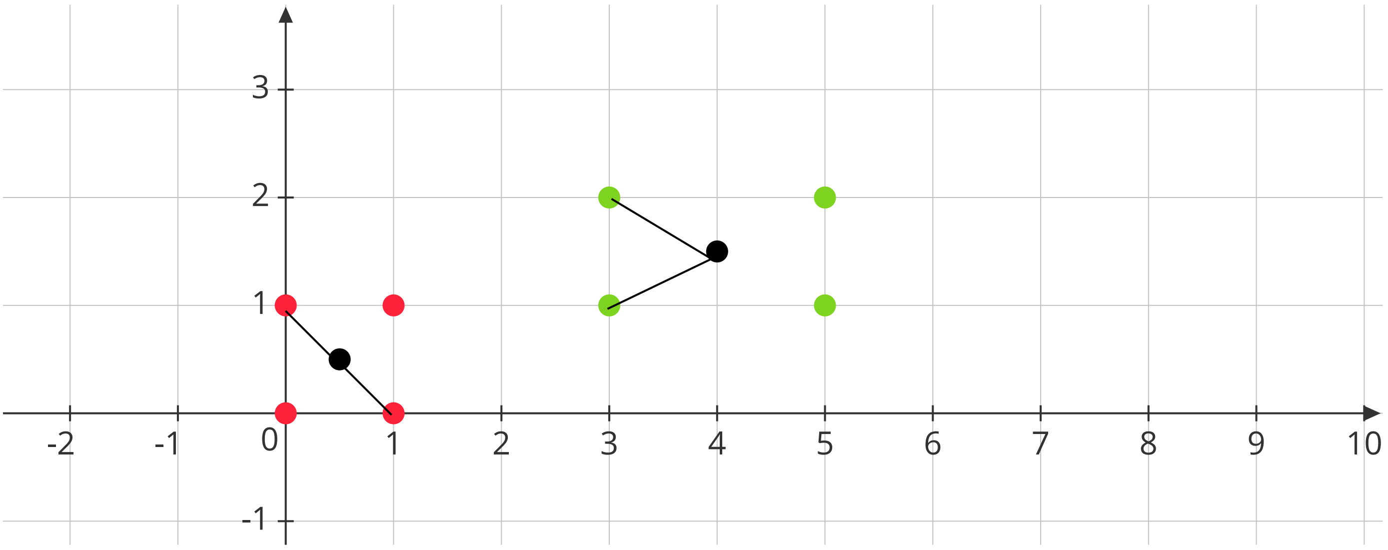
Step-2: Compute the cluster assignments



x_i	z_t	z_{t+1}
$(0, 0)$	1	1
$(0, 1)$	1	1
$(1, 0)$	1	1
$(1, 1)$	1	1
$(3, 1)$	2	2
$(3, 2)$	2	2
$(5, 1)$	2	2
$(5, 2)$	2	2

We see that $z_1 = z_2$. This means that we have converged.

Step-2: Compute the means



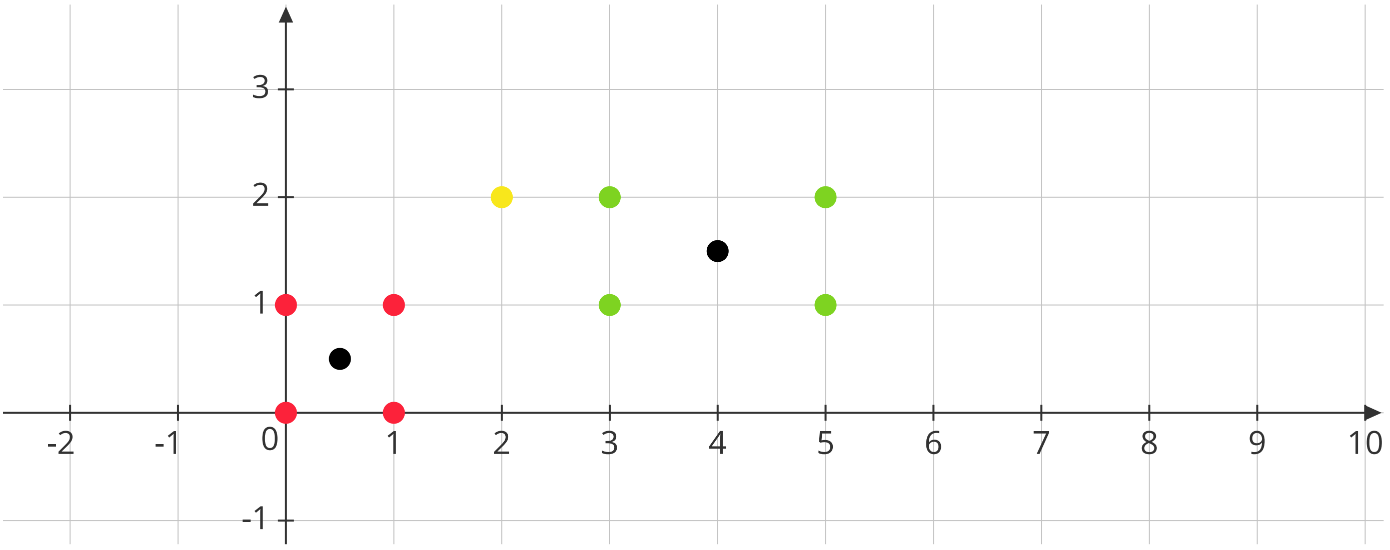
The value of the objective function:

$$f(D) = \sum_{i=1}^{n} ||x_i - \mu_{z_i}||^2$$

$$
\begin{aligned}
f(D) &= 0.5 \times 4 + 1.25 \times 4 \\
&= 2 + 5 \\
&= \boxed{7}
\end{aligned}
$$

$f(D)$ captures intra-cluster distances (within-cluster distances) and not inter-cluster (between two clusters) distances.

<u>Voronoi regions</u>



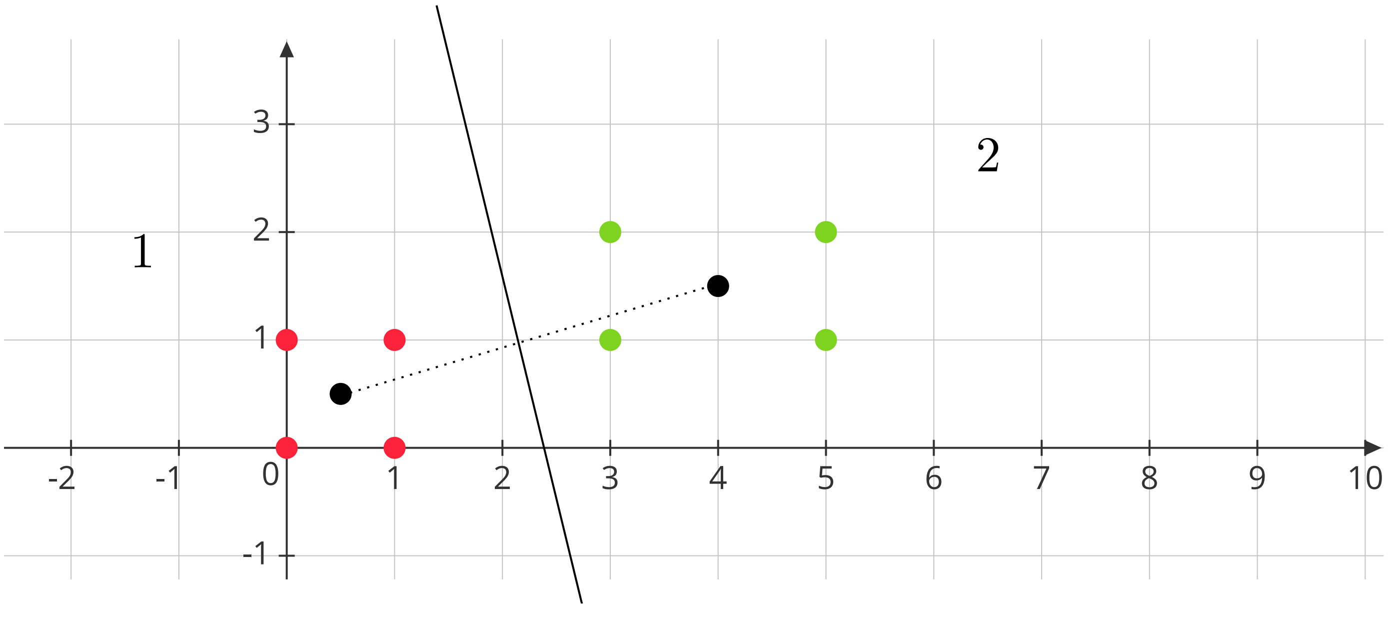
Distance squared from cluster 1

$$d_1^2 = (2 - 0.5)^2 + (2 - 0.5)^2 = 4.5$$

Distance squared from cluster 2

$$d_2^2 = (2 - 4)^2 + (2 - 1.5)^2 = 4.25$$

The test-point $(2, 2)$ belongs to cluster number 2.



The Voronoi regions are half-planes.

1.4. Run k-means with $k = 2$ and $z = \begin{bmatrix} 1 & 2 & 1 & 2 & 1 & 2 & 1 & 2 \end{bmatrix}$. Plot the Voronoi regions. To which cluster does $(2, 2)$ belong? Find the value of the objective function at the end.

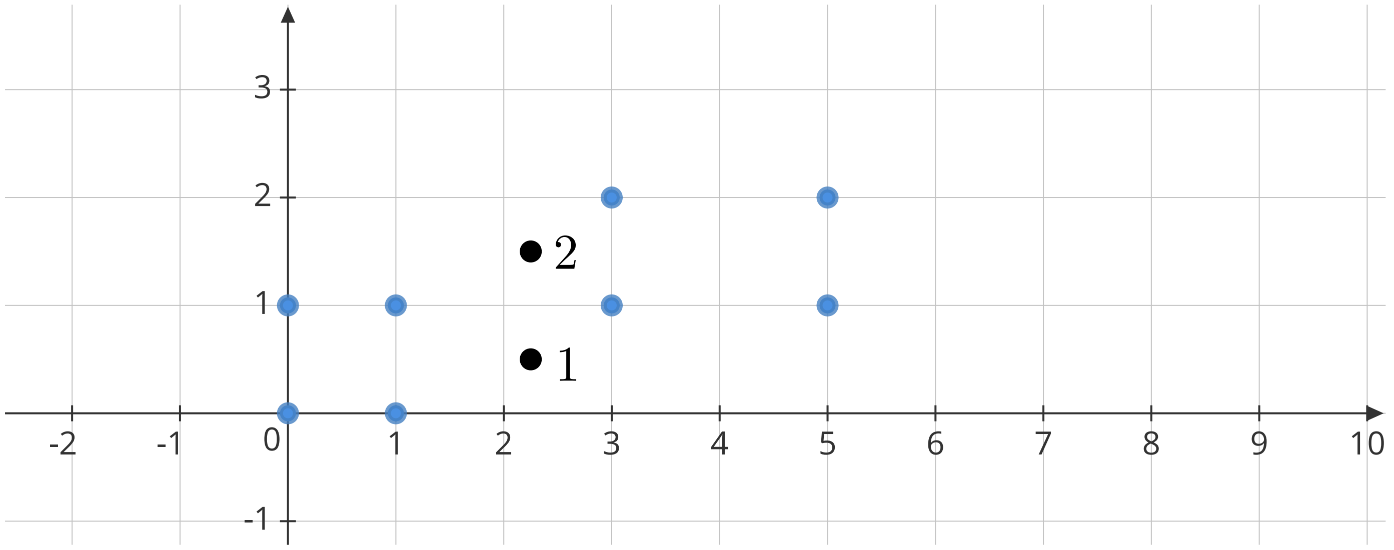
Step-0

$$X = \begin{bmatrix} 0 & 0 & 1 & 1 & 3 & 3 & 5 & 5 \\ 0 & 1 & 0 & 1 & 1 & 2 & 1 & 2 \end{bmatrix}$$

$$z_0 = \begin{bmatrix} 1 & 2 & 1 & 2 & 1 & 2 & 1 & 2 \end{bmatrix}^T$$

$$\mu_1 = \frac{1}{4}\begin{bmatrix} 9 \\ 2 \end{bmatrix} = \begin{bmatrix} 2.25 \\ 0.5 \end{bmatrix}$$

$$\mu_2 = \frac{1}{4}\begin{bmatrix} 9 \\ 6 \end{bmatrix} = \begin{bmatrix} 2.25 \\ 1.5 \end{bmatrix}$$



Step-1: Compute the cluster assignments

If there are ties ($d_1 = d_2$), keep as it is.

Here the subscript corresponds to the data-point:

$$
\begin{aligned}
z_1 &= 1 \\
z_2 &= 2 \\
z_3 &= 1 \\
z_4 &= 2 \\
z_5 &= 1 \\
z_6 &= 2 \\
z_7 &= 1 \\
z_8 &= 2
\end{aligned}
$$

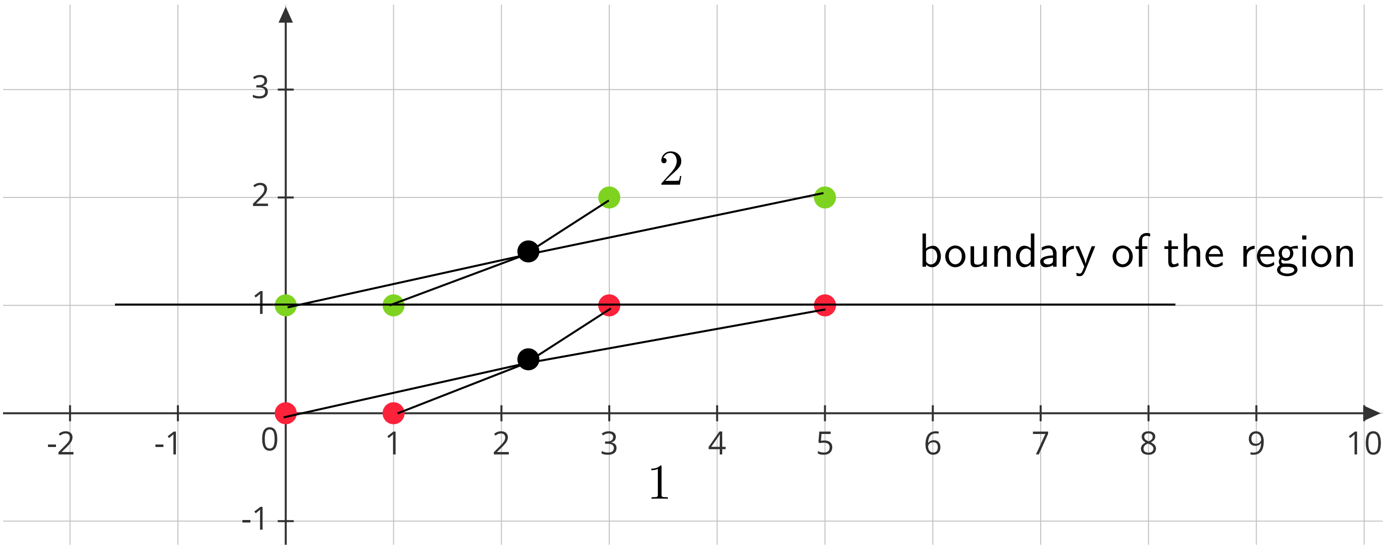
Initialization:

$$z_0 = \begin{bmatrix} 1 & 2 & 1 & 2 & 1 & 2 & 1 & 2 \end{bmatrix}^T$$

After one iteration:

$$z_1 = \begin{bmatrix} 1 & 2 & 1 & 2 & 1 & 2 & 1 & 2 \end{bmatrix}^T$$

Stop.

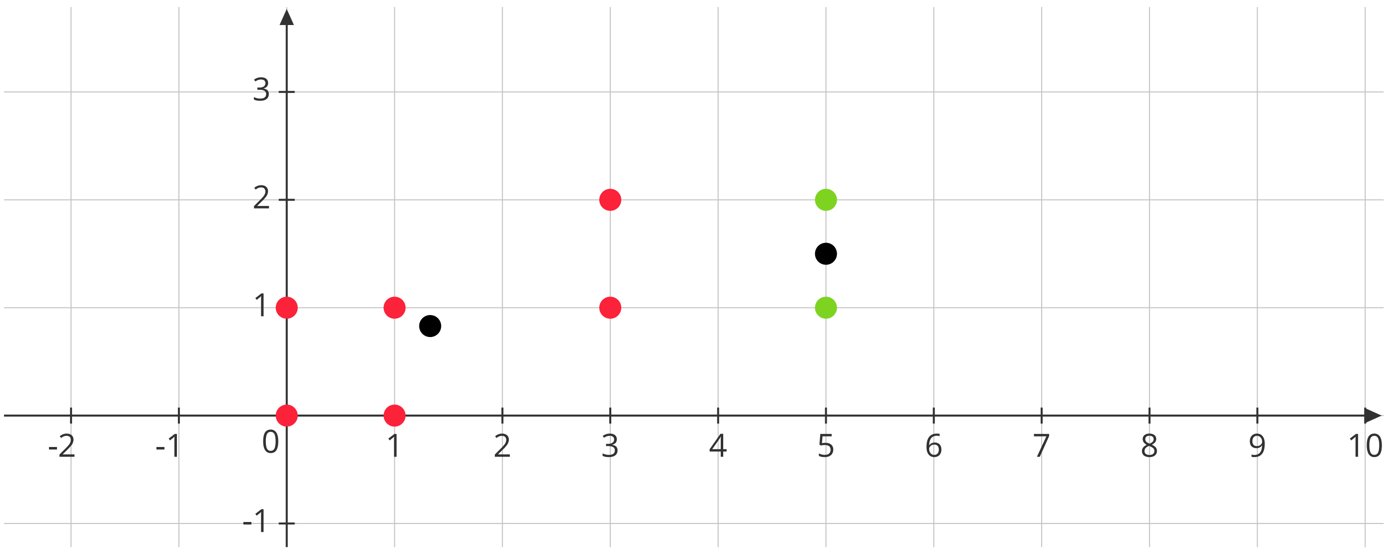


Main takeaway: the final clusters are dependent on the initialization.

$$f(D) > 7$$

This is worse than the previous init.

1.5. Run k-means with $k = 2$ and $z = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 2 & 2 \end{bmatrix}$. Plot the Voronoi regions. To which cluster does $(2, 2)$ belong? Find the value of the objective function at the end.



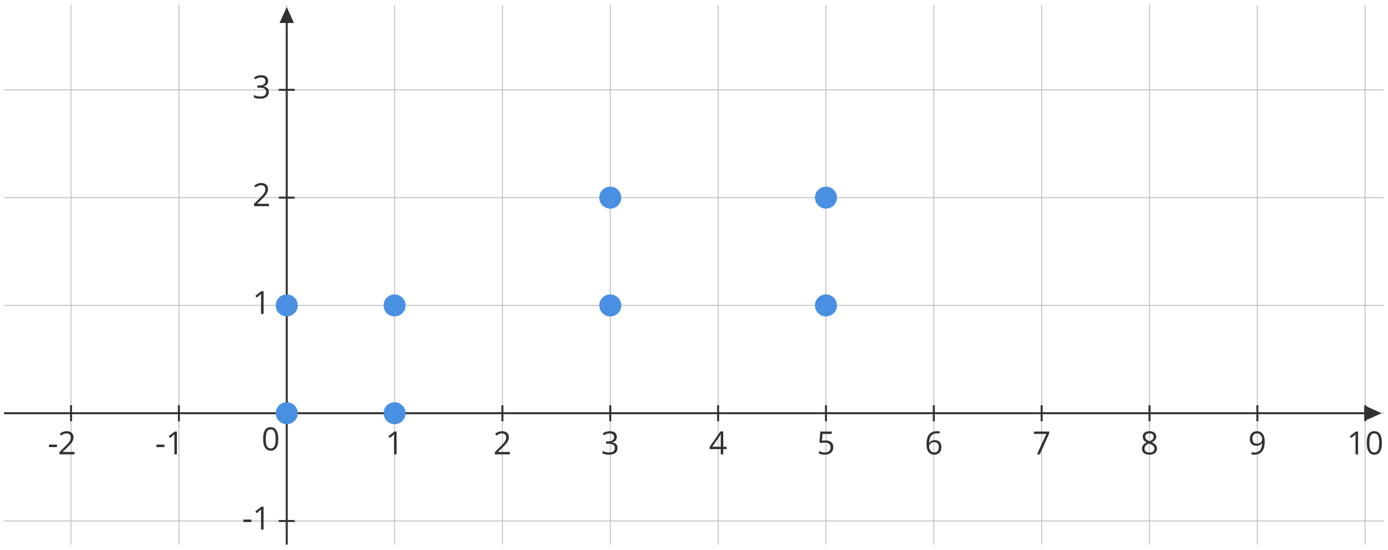
1.6. Run k-means with $k = 2$, but by initializing the first mean as $\mu_1 = \begin{bmatrix} 100 \\ 100 \end{bmatrix}$ and $\mu_2 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$. What do you observe?

Different ways of initialization

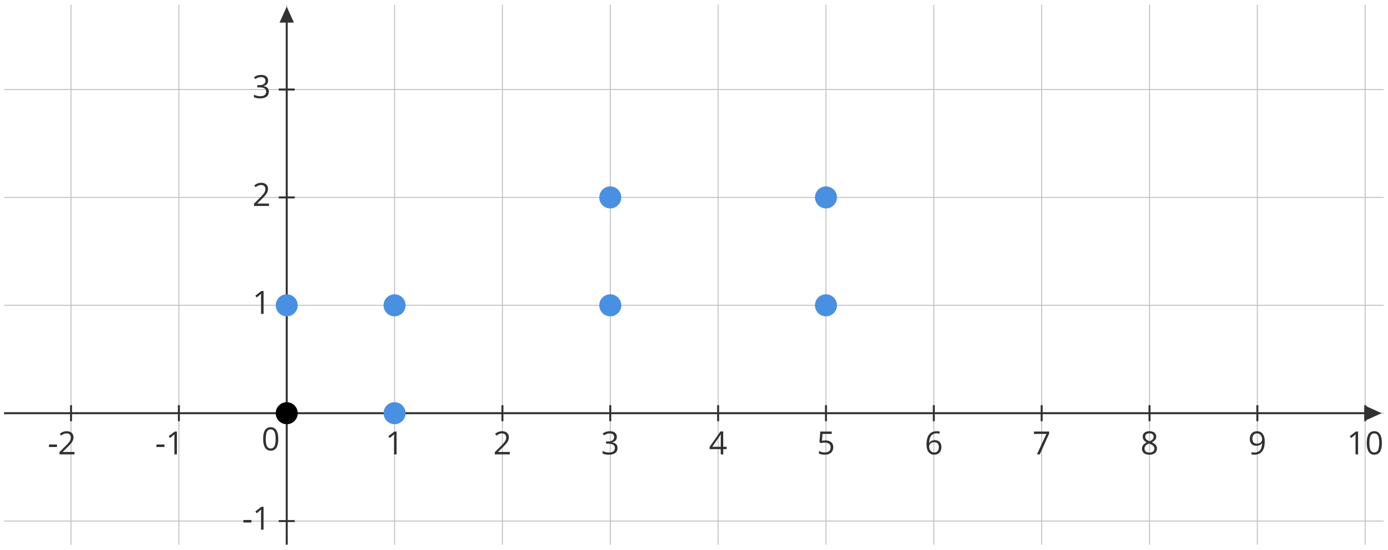
- Choose z values for all n data-points
- Choose k points in the dataset at random and make them the initial means.
- Choose some k points in $\mathbb{R}^d$ as the means.

All points will be assigned to cluster 2.

1.7. For $k = 3$, run a simulation of the K-means++ algorithm. Compute the probabilities of choosing different points as the 3 means.



Step-1: Choose a point uniformly at random

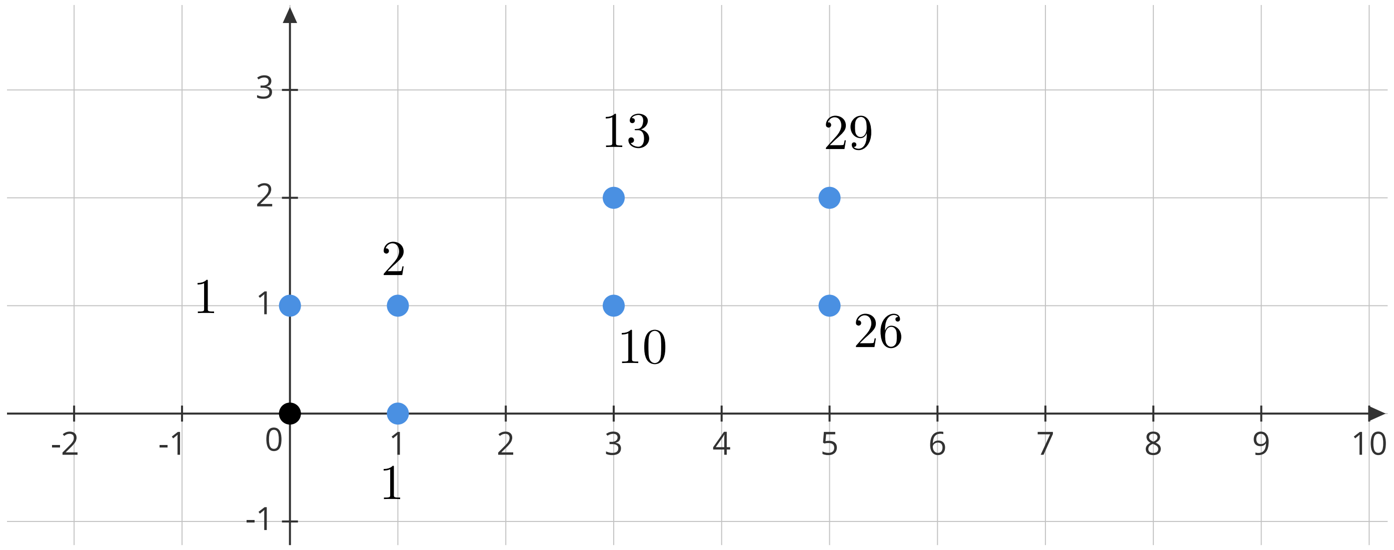


$$P(x_1) = \frac{1}{8}$$

$$\mu_1 = (0, 0)$$

Step-2: Choose the second mean

Step-2.1: Compute the scores (squared distances) for the remaining 7 data-points from μ_1



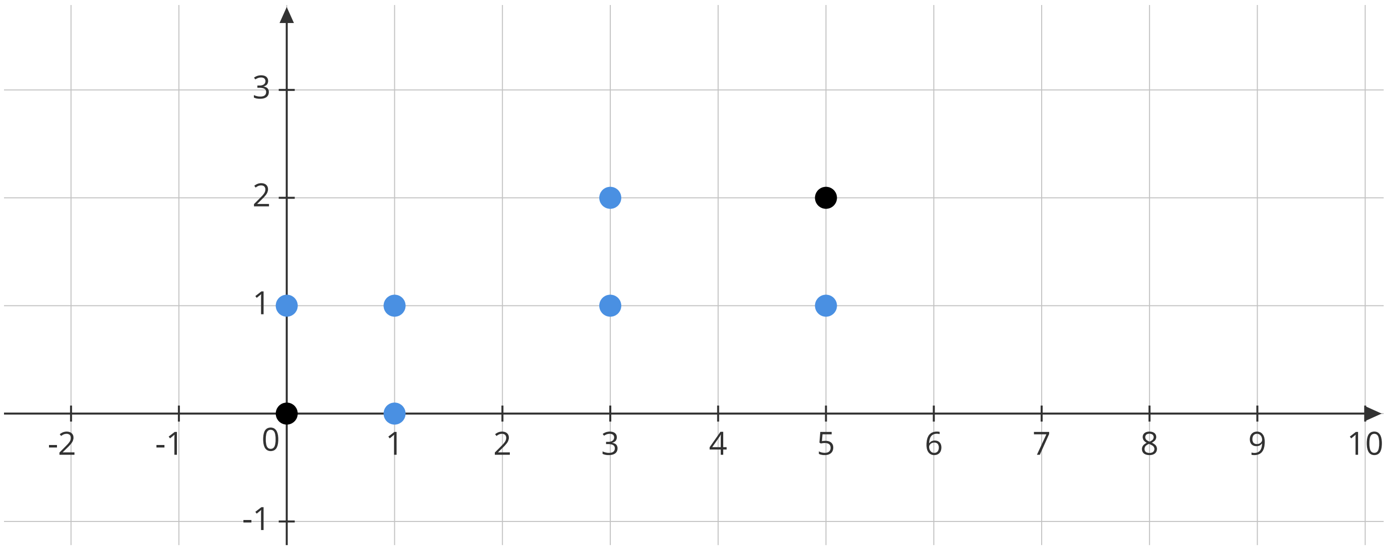
Step-2.2: Form the probability distribution over the 7 data-points using these scores

x_i	$P(\mu_2 = x_i \mid \mu_1 = (0,0))$
$(0,1)$	$\frac{1}{82}$
$(1,0)$	$\frac{1}{82}$
$(1,1)$	$\frac{2}{82}$
...	...

Step-2.3: Sample a point from this distribution

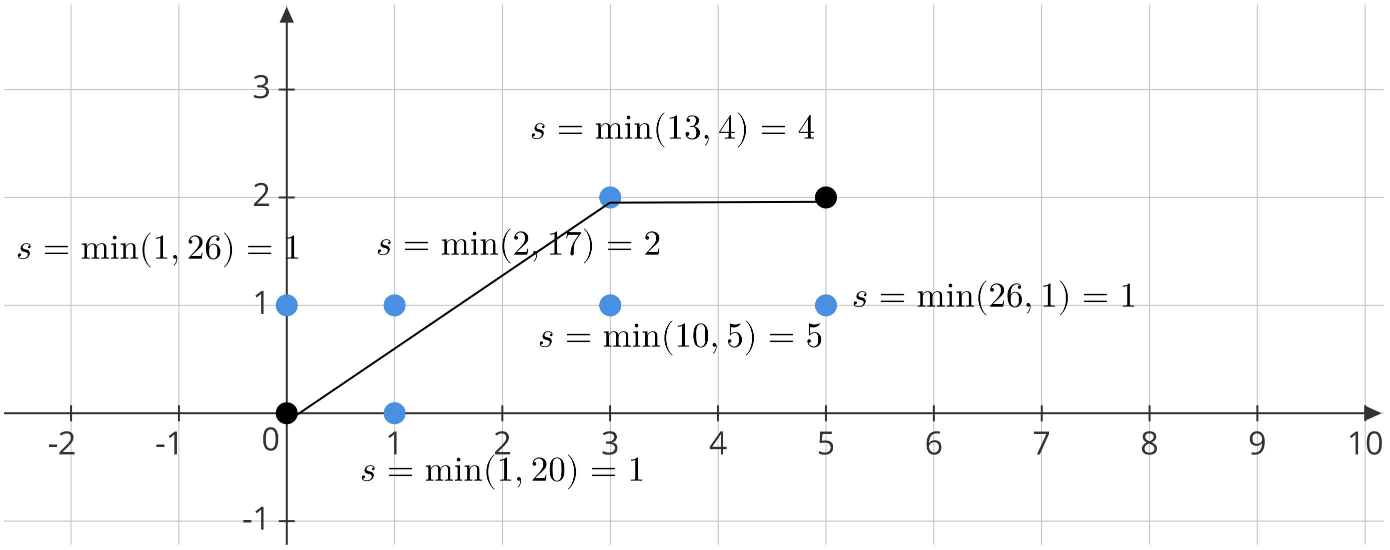
For this run, let us assume that $\mu_2 = (5,2) = x_8$. The probability associated with this:

$$P[\mu_2 = (5,2) \mid \mu_1 = (0,0)] = \frac{29}{82}$$



Step-3: Choose the third mean

Step-3.1: Compute the distances of each of the six points to the two means



Score is distance * distance

Form the probability distribution using the scores

$$\sum s_i = 14$$

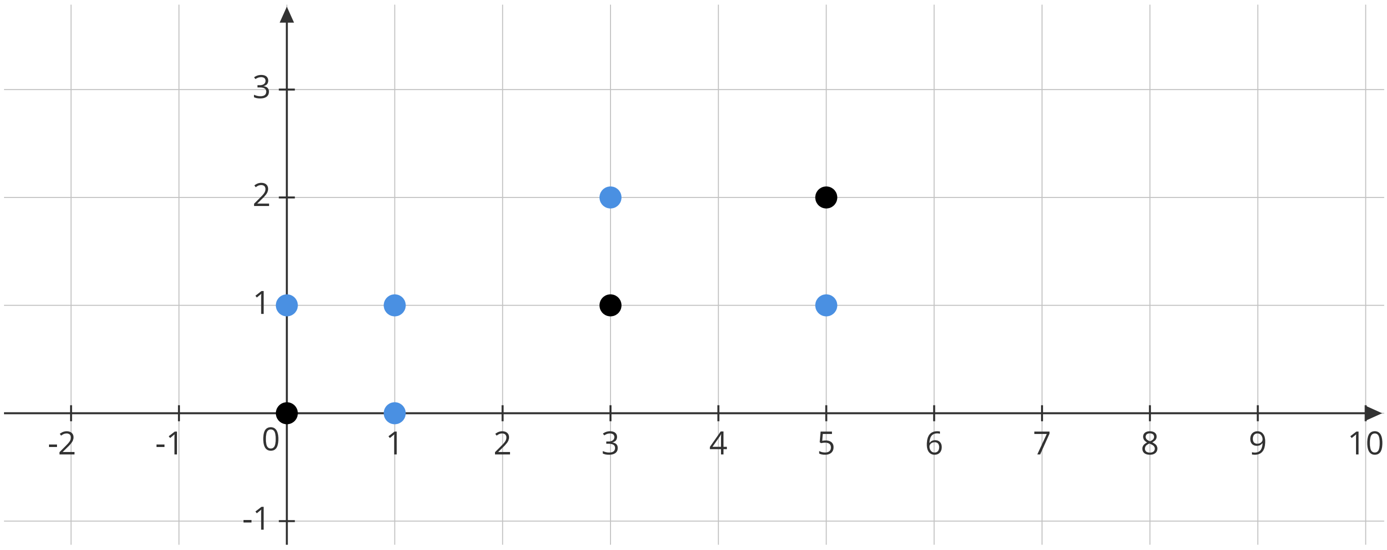
The probability of choosing $(3, 1)$ as the third mean condition on the first two means is $\frac{5}{14}$.

$A \rightarrow$ choose the first mean
$B \rightarrow$ choose the second mean
$C \rightarrow$ choose the third mean

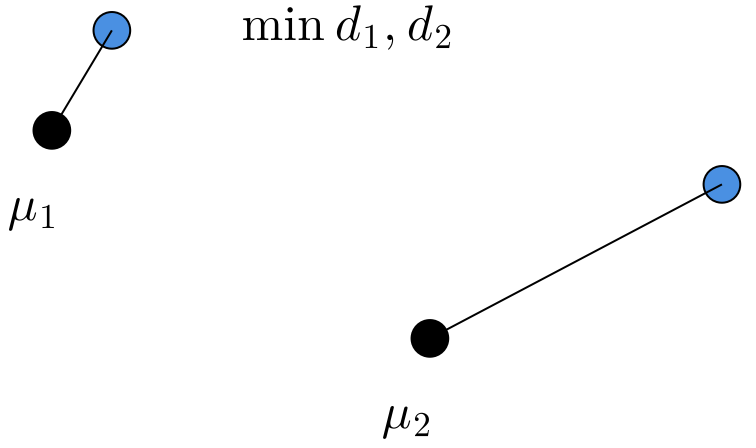
$$P(A) = \frac{1}{8}$$
$$P(B|A) = \frac{29}{82}$$

$$P(C|A, B) = \frac{5}{14}$$

$$P(A, B, C) = \frac{1}{8} \times \frac{29}{82} \times \frac{5}{14}$$



The three means should be as far away from each other as possible.



2. Demo

PRML: Pattern Recognition and Machine Learning by Chis Bishop (Microsoft)

Old Faithful Geyser

[Credits: Wikipedia]

Figure 1: Eruption of Old Faithful in 1948

Old Faithful Geyser Data

Description: (From R manual):

> Waiting time between eruptions and the duration of the eruption for the Old Faithful geyser in Yellowstone National Park, Wyoming, USA.
>
> A data frame with 272 observations on 2 variables.

eruptions	numeric	Eruption time in mins
waiting	numeric	Waiting time to next eruption

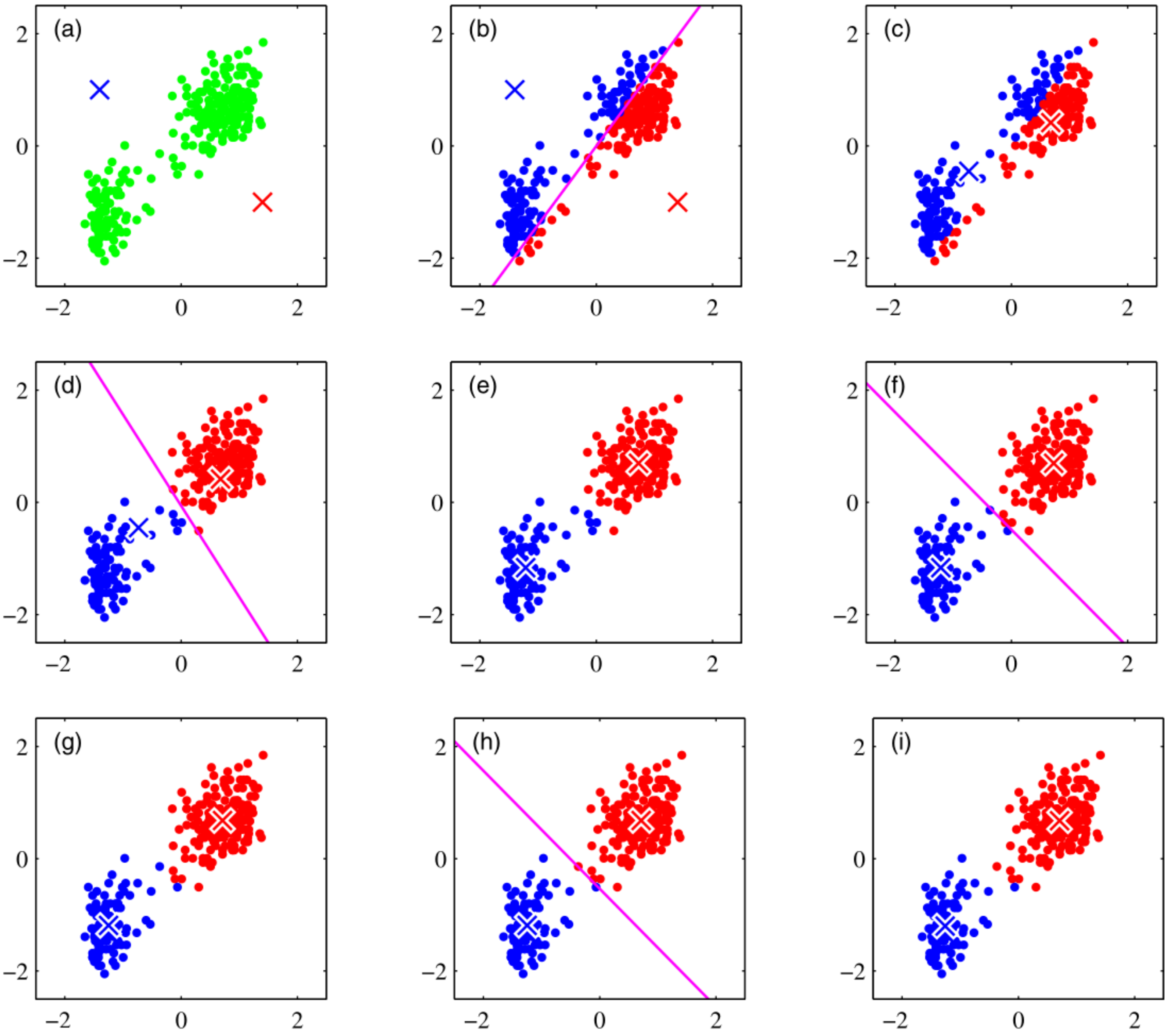
	eruptions	waiting
1	3.600	79
2	1.800	54
3	3.333	74
4	2.283	62
5	4.533	85

Notice that the scales of the two features are different. One good idea is to normalize them

K-means on normalized dataset:

Mean-variance normalization:

$$(x_i - \mu) \, / \, \sigma$$



Credits: Page 426, Bishop, PRML [Jordan, Microsoft]

Image Segmentation

Try to identify regions in an image that are homogeneous. For example, this could mean an object, a face. This is problem in computer vision $\rightarrow$ helping computers make sense of the visual world.

An image is a rectangular grid of pixels. This image could have 400×200 pixels. $80,000$ pixels. Each pixel in this image is a data-point. Each pixel is a vector in $\mathbb{R}^3$. This is a color image:

- Red channel: $(0, 255)$, 8 bits
- Blue channel: $(0, 255)$
- Green channel: $(0, 255)$

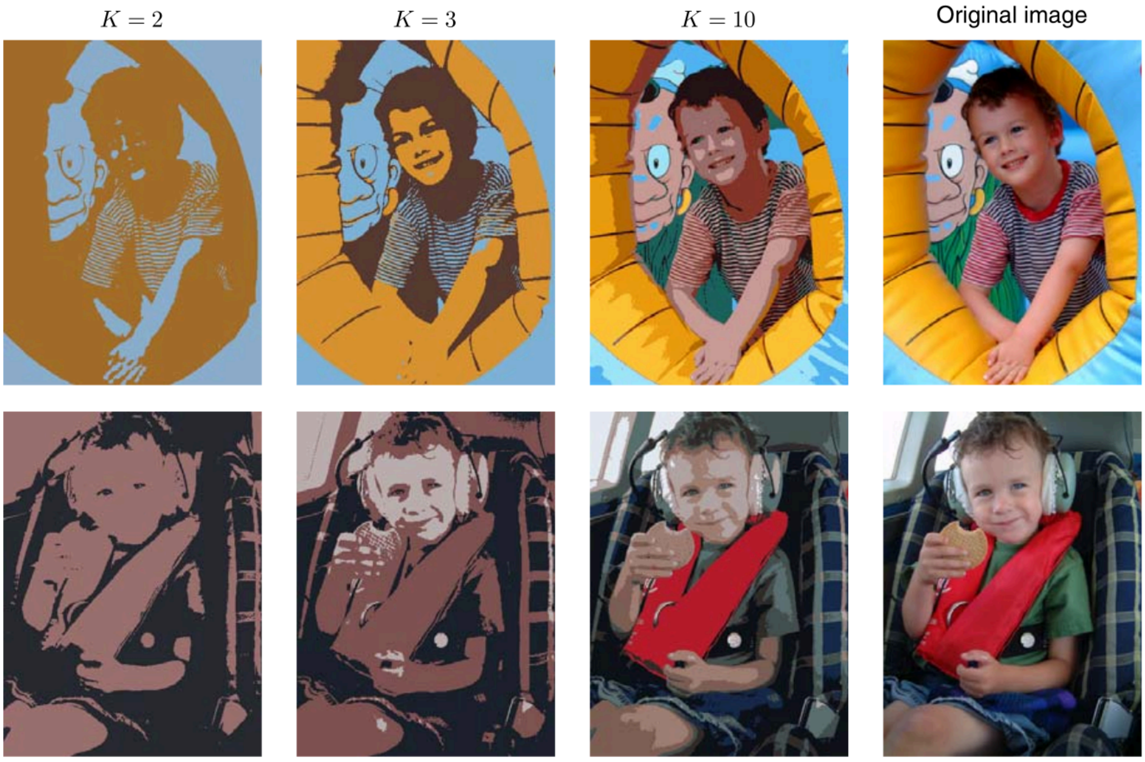
$$\begin{bmatrix} 100 \\ 125 \\ 200 \end{bmatrix}$$

The shape of the data-matrix:

$$3 \times 80,000$$

Run k-means on this with different values of k:

$$\mu_1 = \begin{bmatrix} 100 \\ 125 \\ 200 \end{bmatrix}$$



Credits: Page 429, Bishop, PRML [Jordan, Microsoft]

Image compression