

Clustering

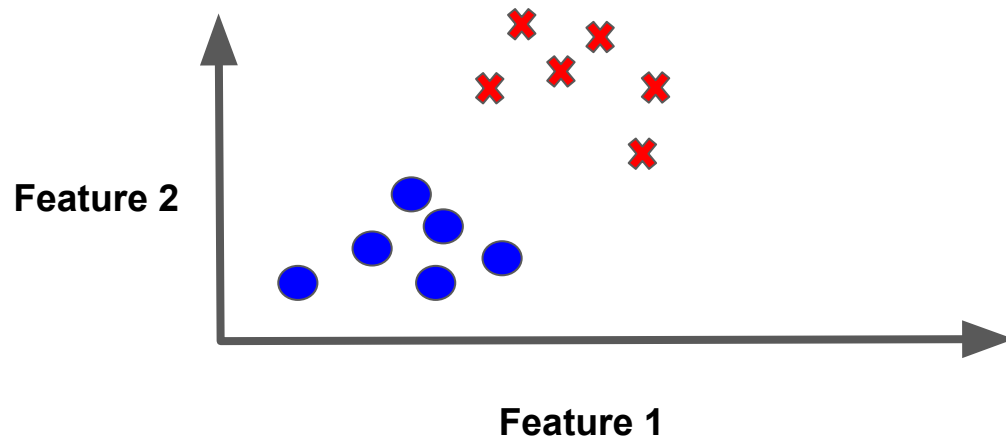
Wajahat Hussain



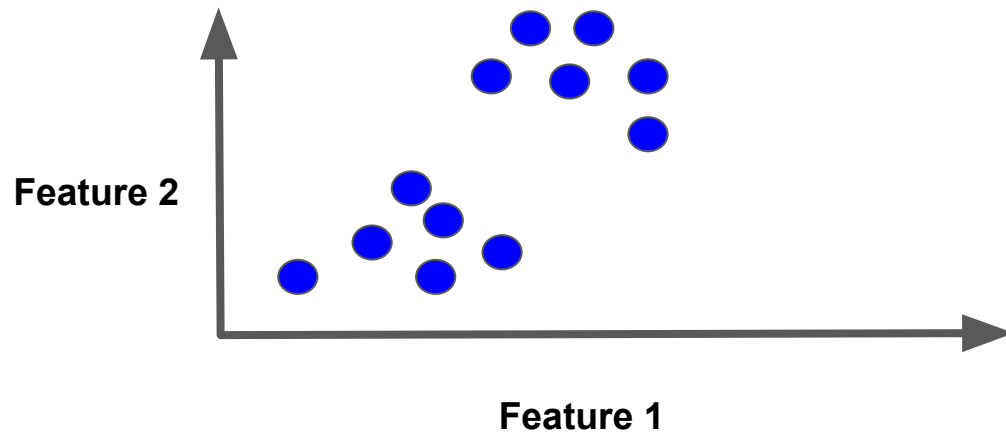
Divide this data into two clusters

X	Y
1	2
2	3
3	4
4	5
5	6
6	7
7	8
8	9
3	2
4	3
5	4
6	5
7	6
8	7
9	8
10	9

Supervised Learning

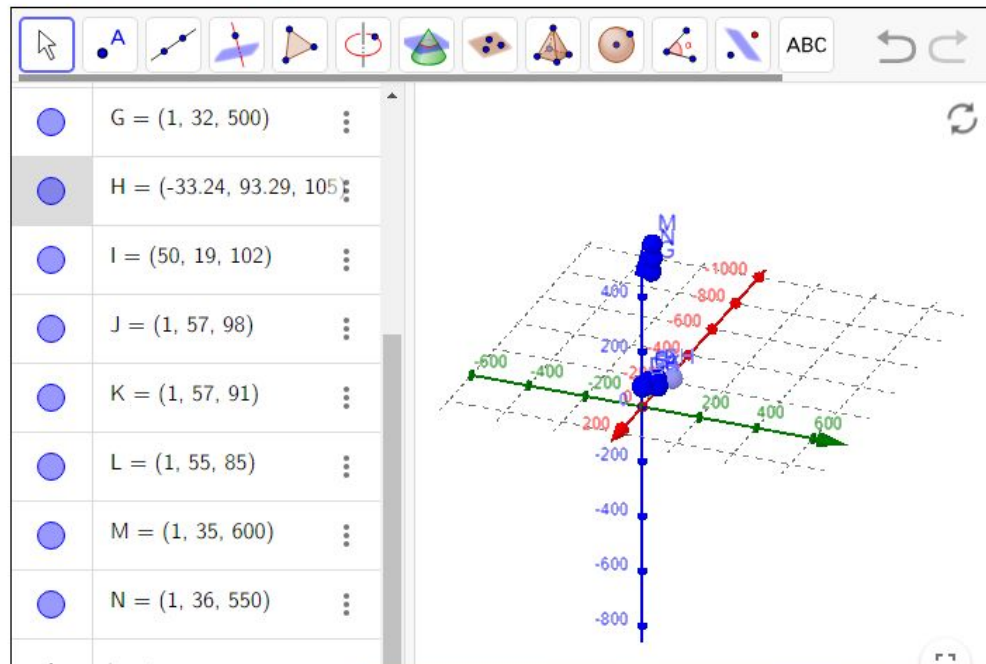


Unsupervised Learning



Geogebra - 3D Graphing

Author: Scott R. Franklin, Ph.D.



Easy Assigning

Assign to your Google classes, create a GeoGebra lesson and much more.



DISMISS

Divide this data into two clusters

X	Y
1	2
2	3
3	4
4	5
5	6
6	7
7	8
8	9
3	2
4	3
5	4
6	5
7	6
8	7
9	8
10	9

Unsupervised Learning

← → ↻ <https://news.google.com>

Google

News Pakistan edition Modern

Top Stories

- Hillary Clinton
- Mosul
- FC Barcelona
- Manchester United F.C.
- Arsenal F.C.
- Jacob Zuma
- Shah Rukh Khan
- Brexit
- Iowa
- MacBook Pro
- Islamabad, Islamabad...
- Suggested for you
- World
- Pakistan
- Business
- Technology
- Sports
- Entertainment
- Health
- Science
- More Top Stories

Top Stories

At least 20 killed, 65 injured in Karachi train collision

Geo News, Pakistan - 2 hours ago

KARACHI: Two trains collided in the city's Landhi area early Thursday morning leaving at least 20 people dead and resulting in injuries to 65 others.

PM Nawaz orders probe into Karachi train accident Pakistan Today

Why do Pakistan Railway trains crash all the time? Daily Pakistan

See realtime coverage

Related Karachi »

The Nation Daily Pakis... Daily Times DAWN.com Geo News, Yahoo News The Expre... DunyaNew... Reuters

In response to SC on Panamagate petitions, PM denies holding offshore companies

DAWN.com - 3 hours ago

A five-member bench of the Supreme Court resumed hearing of the Panamagate petitions on Thursday. Prime Minister Nawaz Sharif in a written statement denied holding any offshore company.

SC gives PMs children last chance to submit replies in Panama Leaks case Geo News, Pakistan

Panama hearing: PM Nawaz denies holding off shore companies, SC to form single-member commission Pakistan Today

Related Nawaz Sharif » Pakistan Tehreek-e-Insaf »

Opinion: Will Supreme Court save democracy? The Nation

In Depth: Supreme Court averts political showdown The Express Tribune

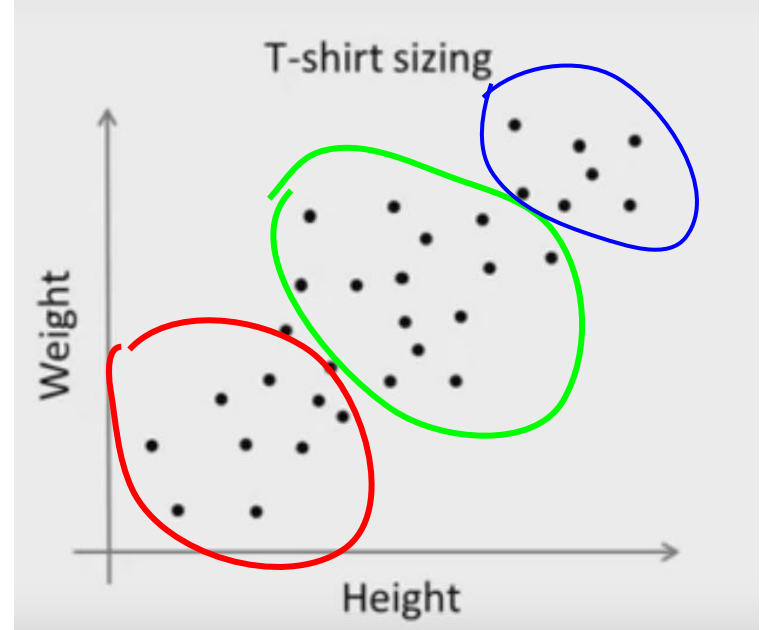
PANAMA PAPER SC COURT OF MINDS SUPREME COURT OF PAKISTAN

Daily Times Geo News, The Nation The Expre... DunyaNew... The News ... DunyaNew... Daily Pakis... The Expre...

Unsupervised Learning

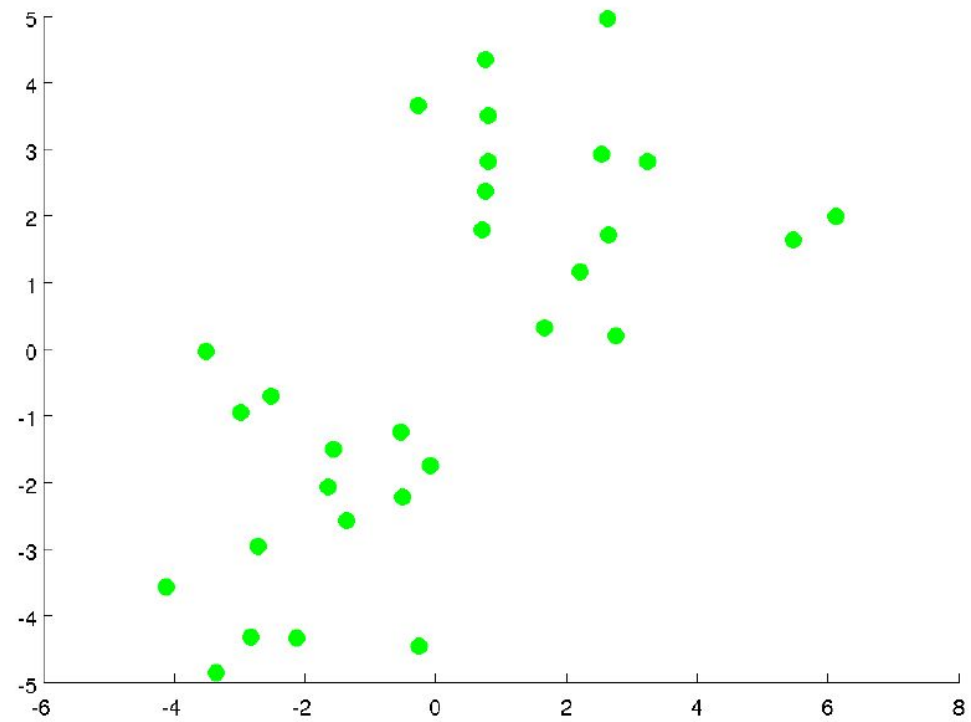


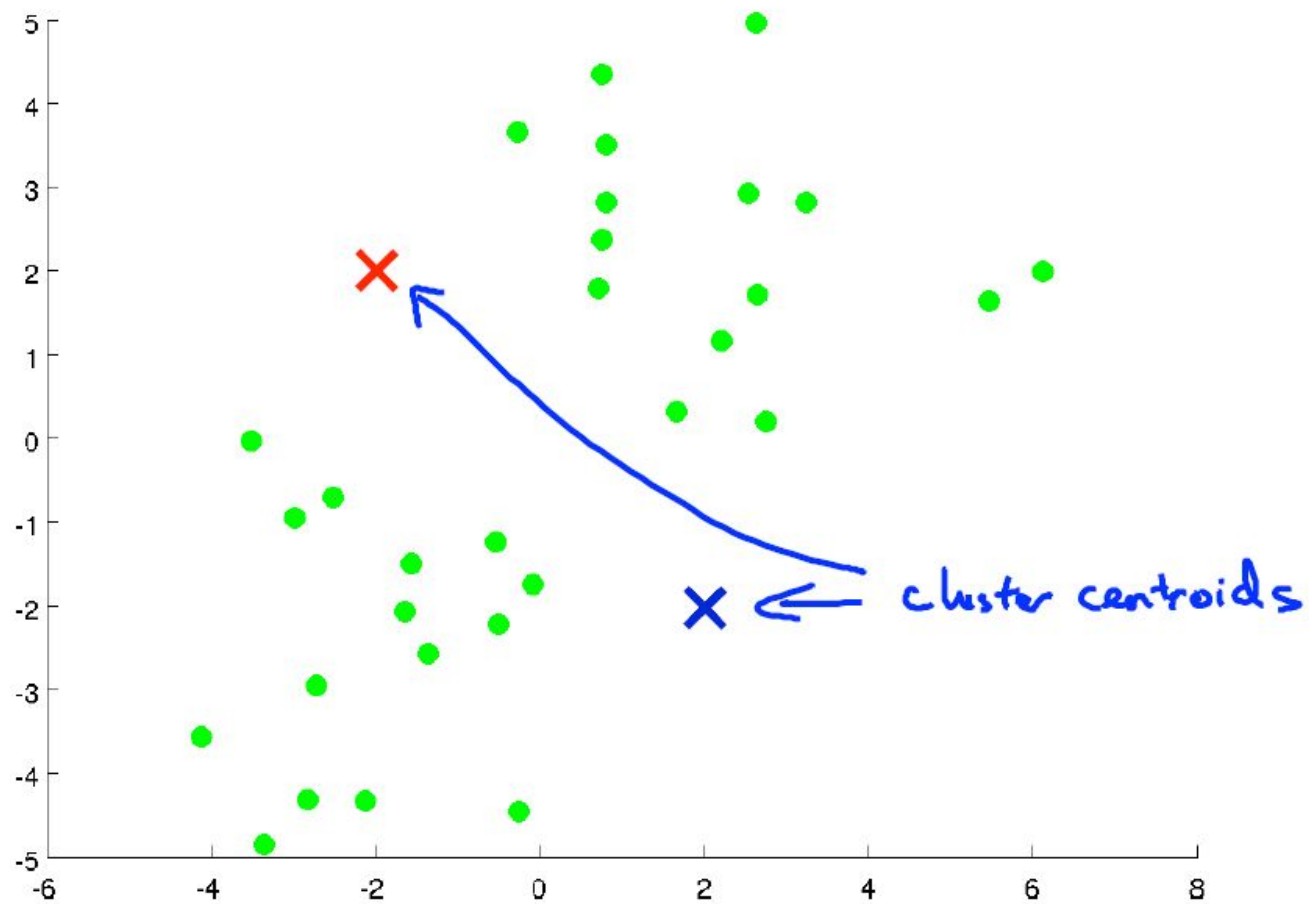
Market segmentation

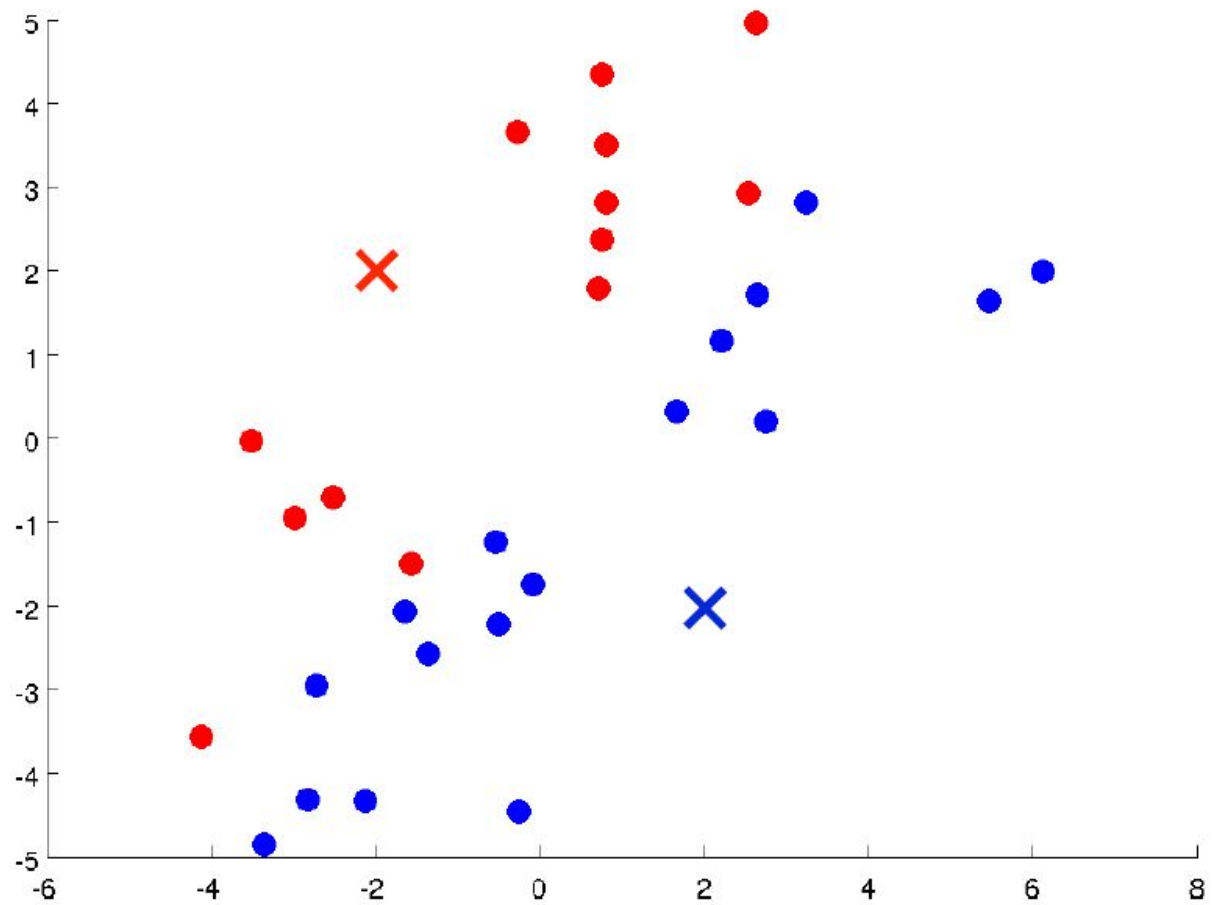


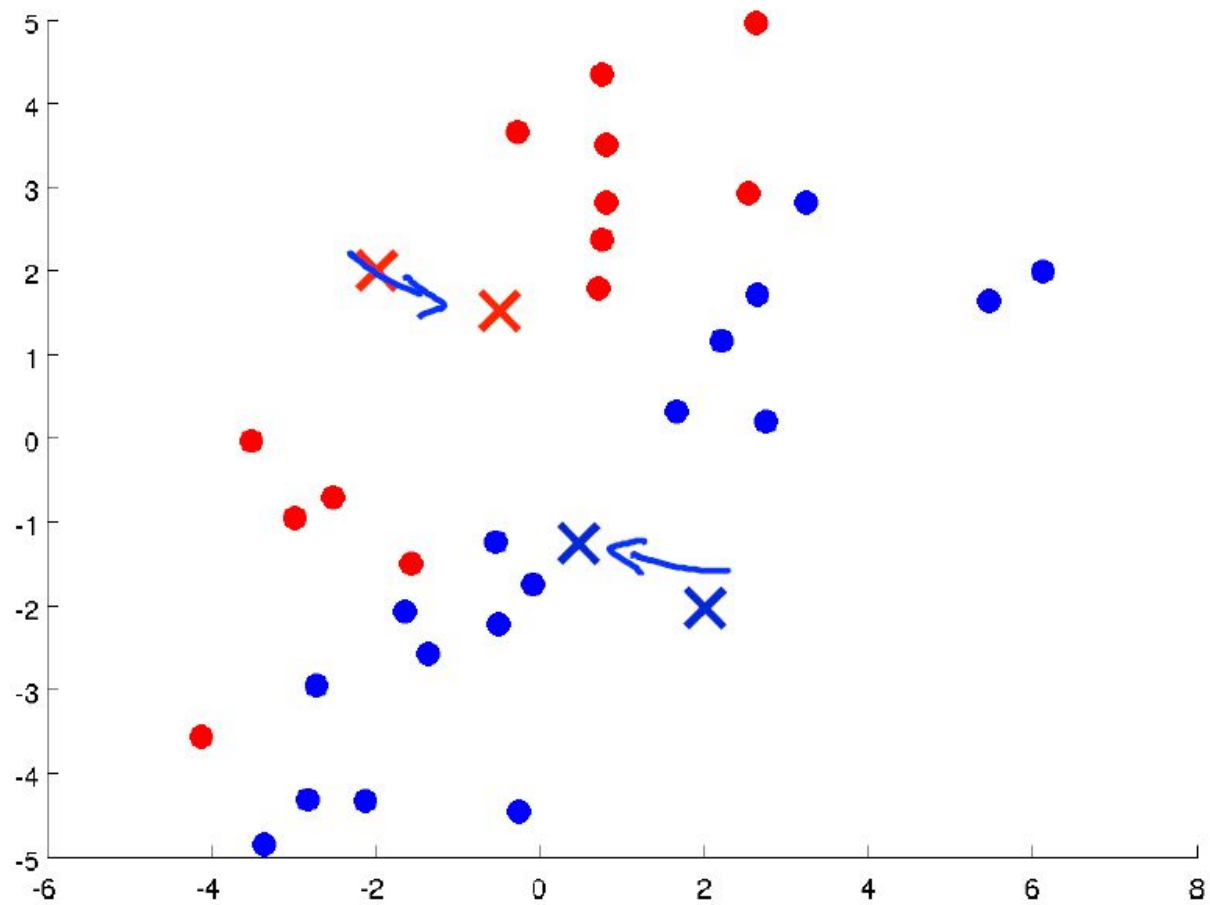
You want to design a shirt. How many sizes should there be?

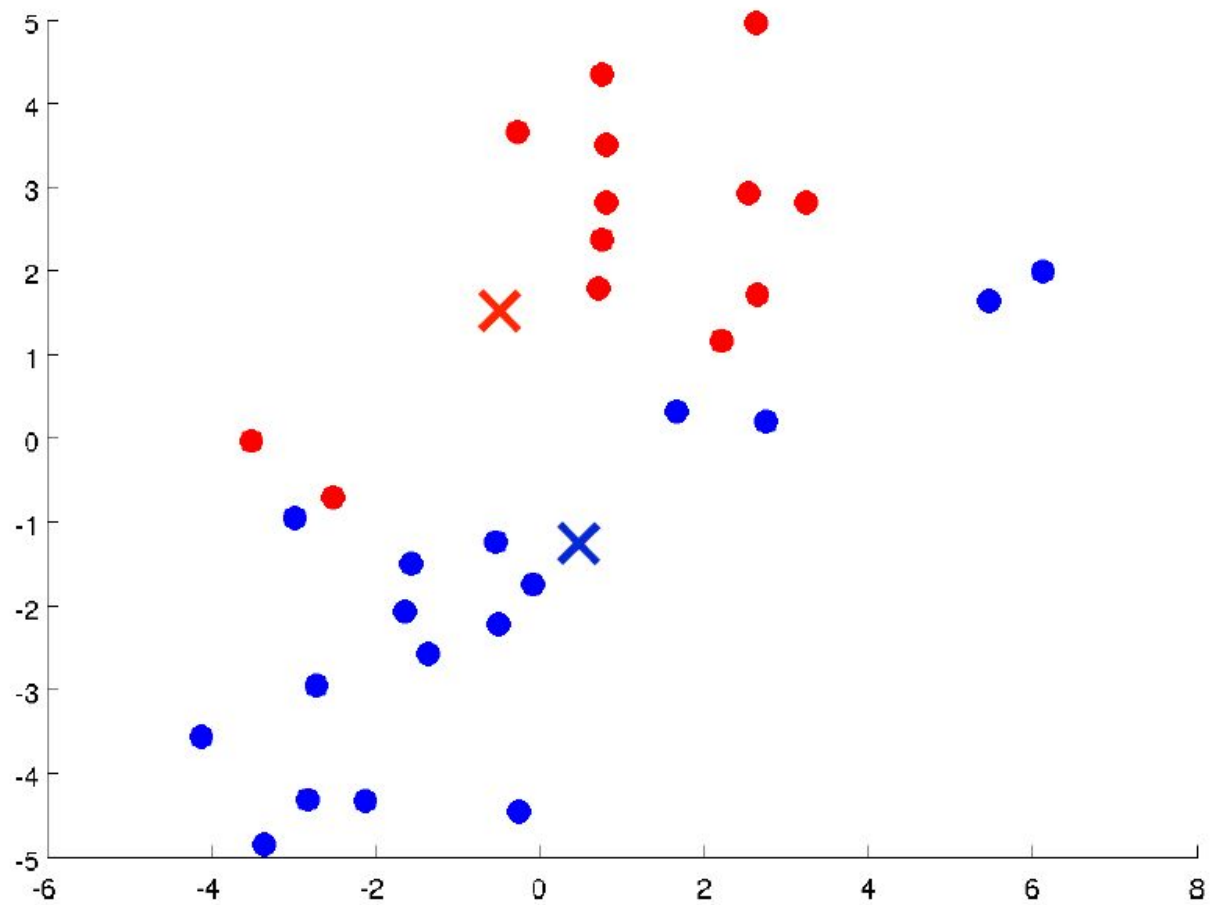
K-means Algorithm

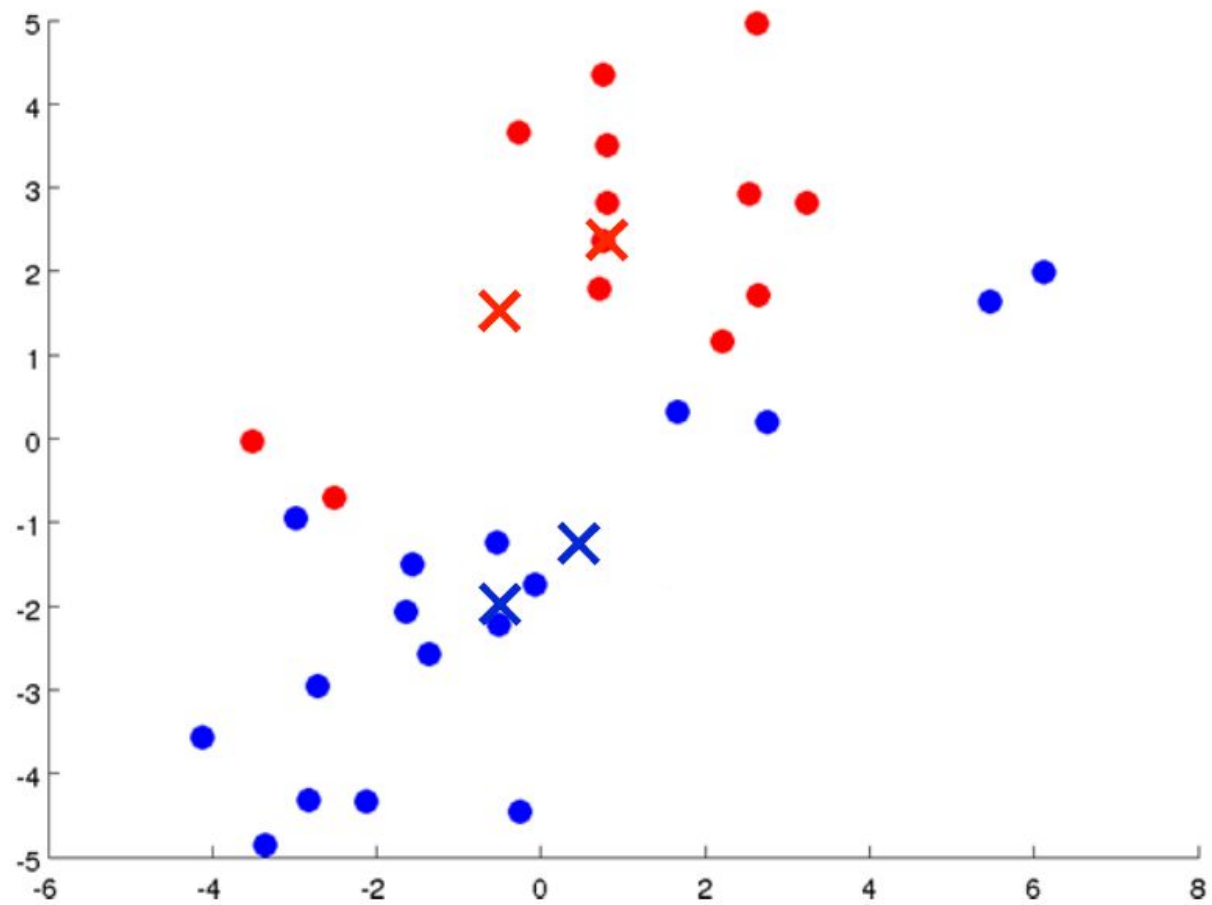


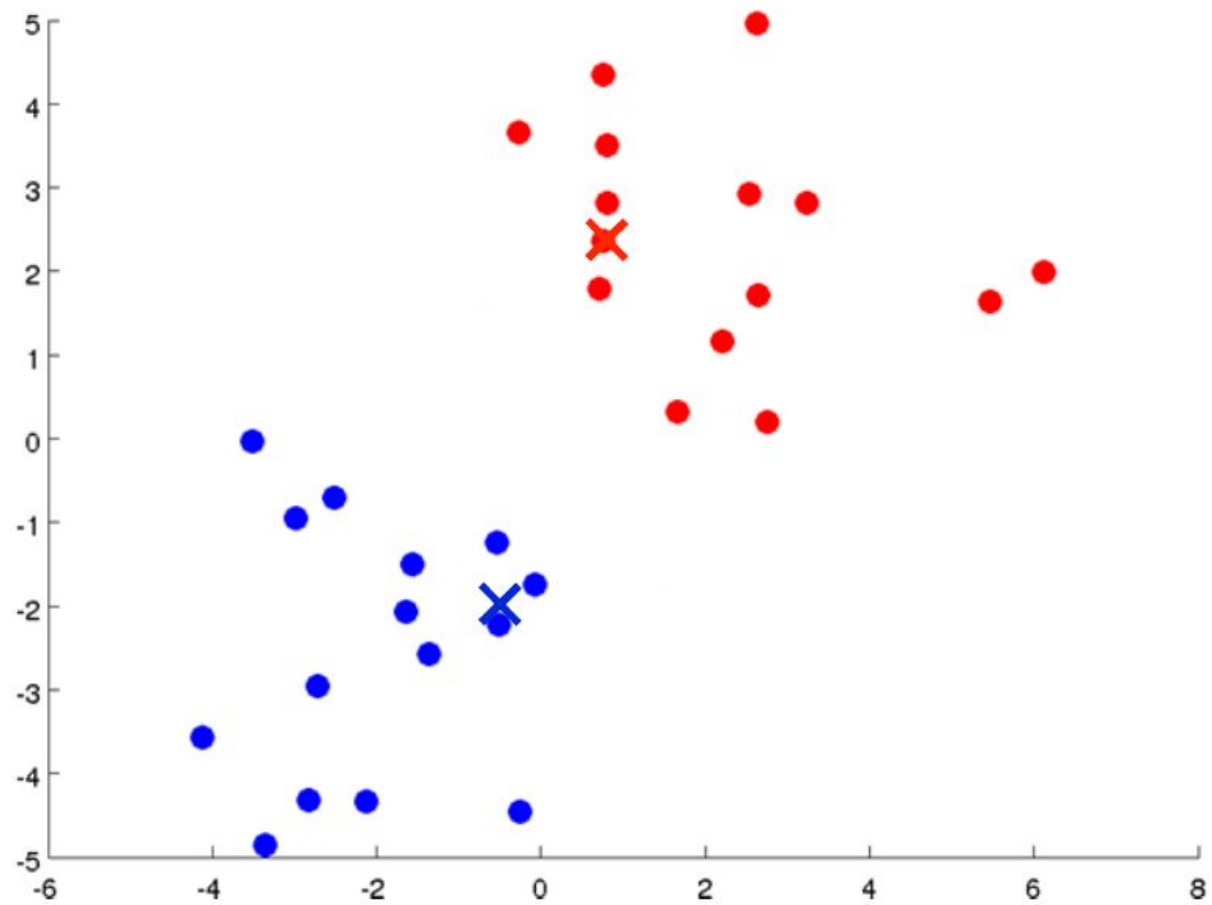


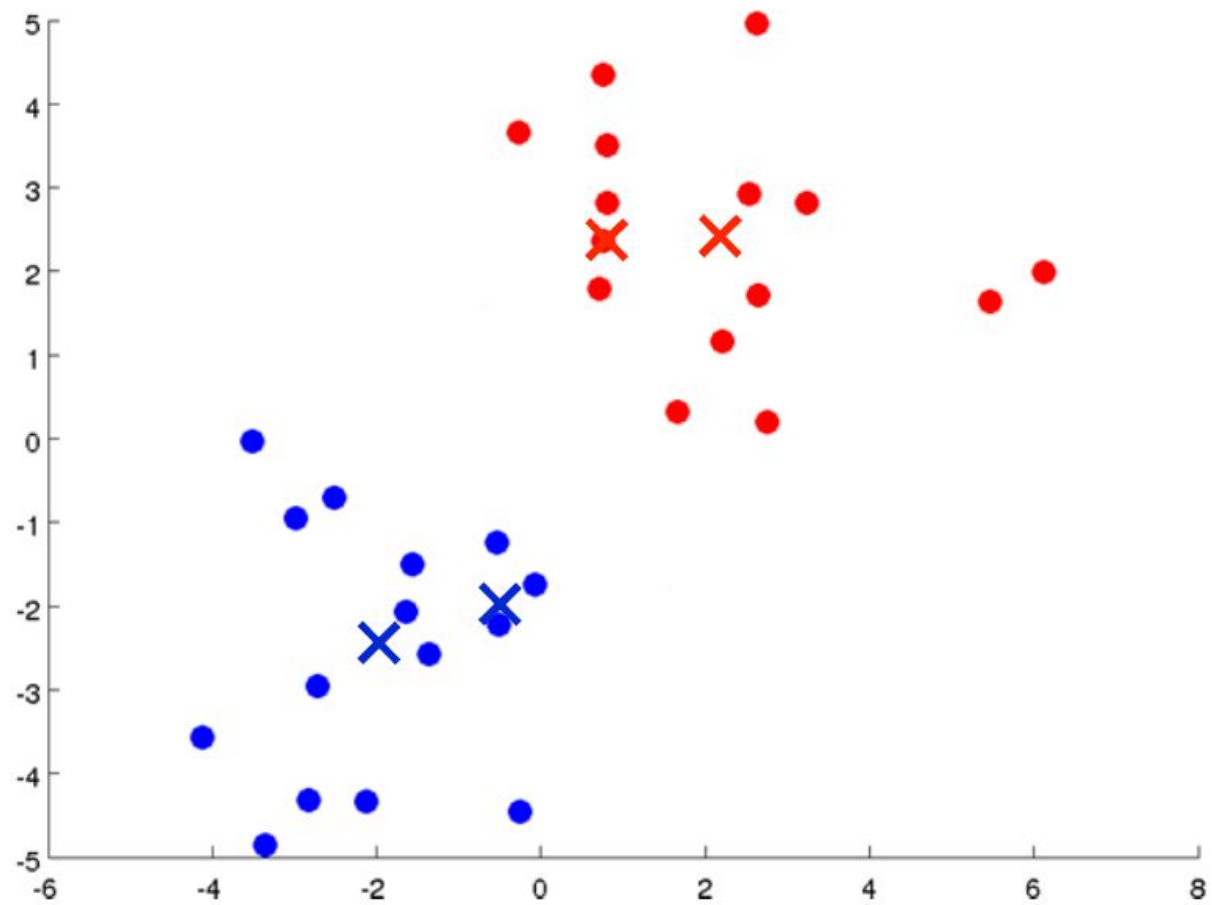


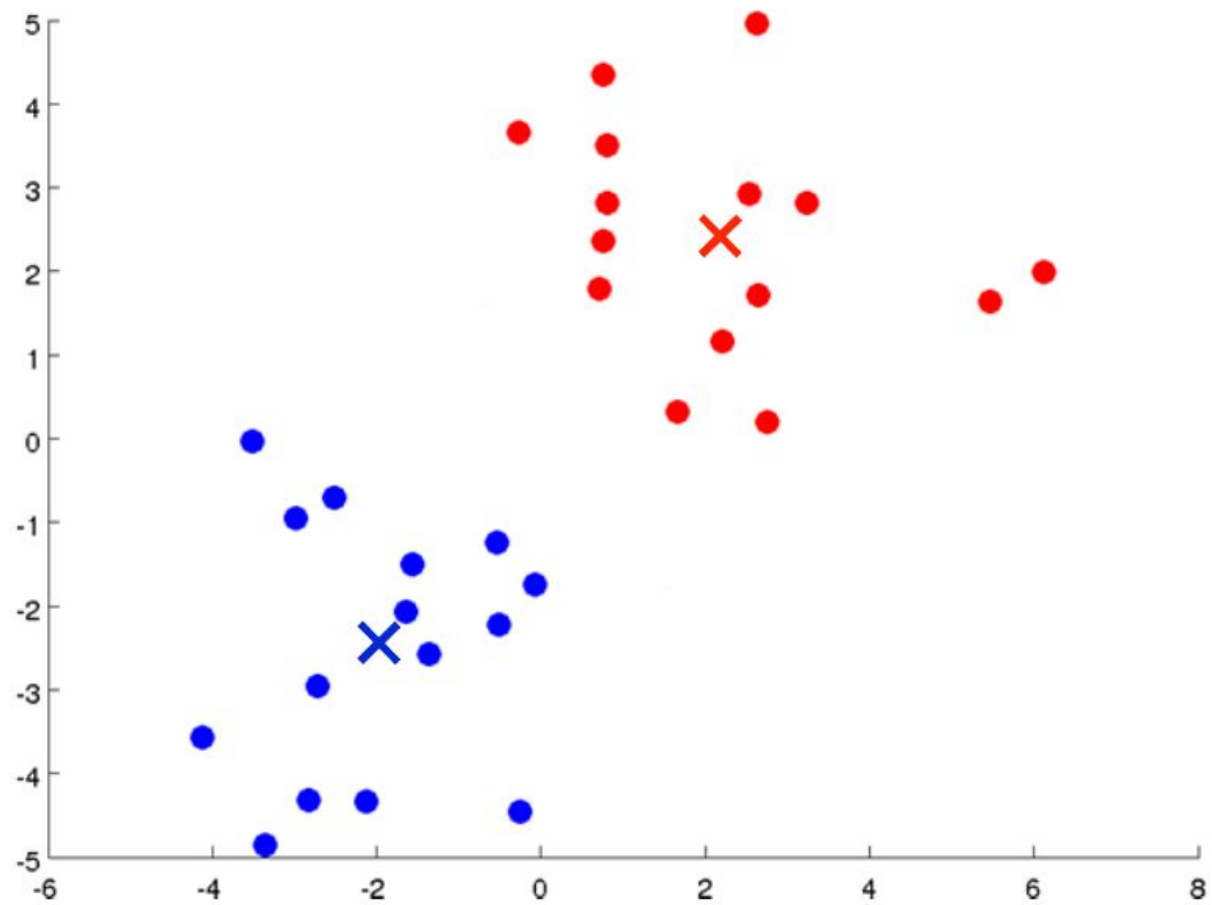


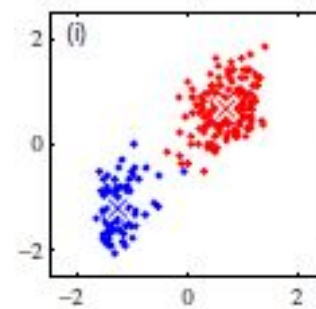
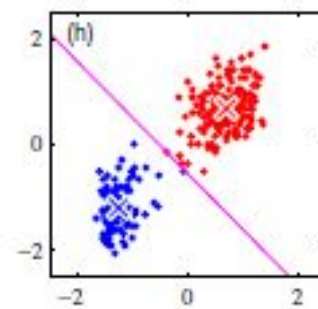
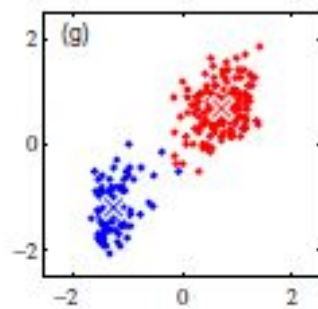
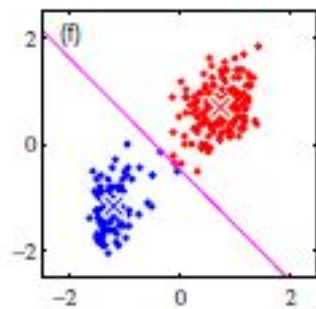
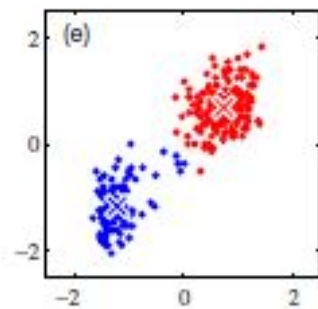
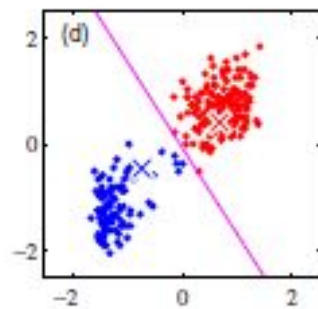
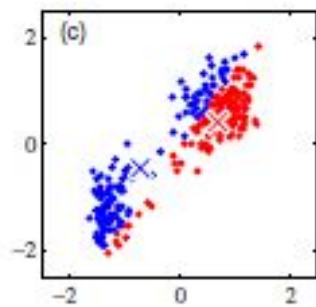
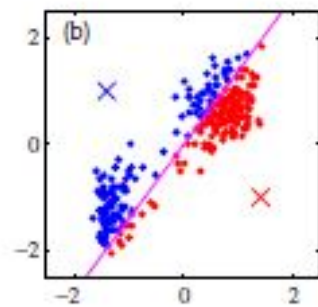
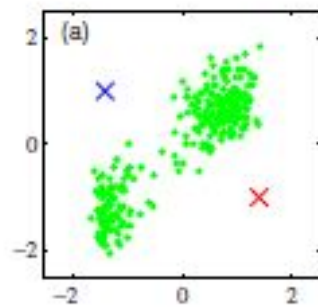






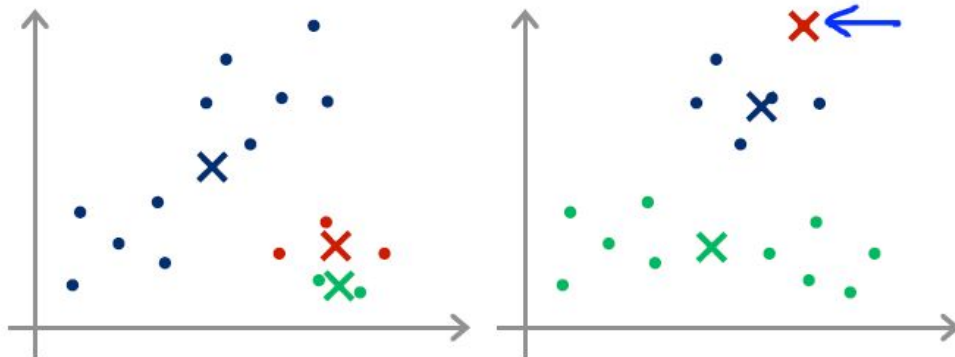
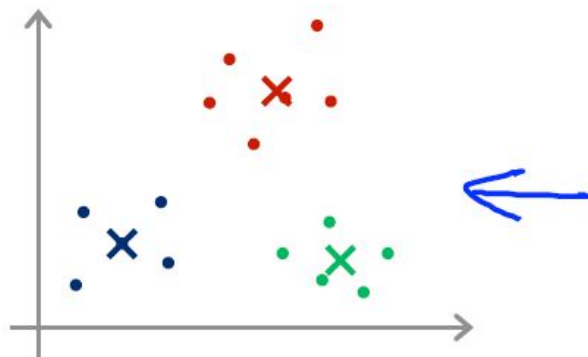






Is there any failure case?

Local optima



K-means algorithm

Input:

- K (number of clusters)
- Training set $\{x^{(1)}, x^{(2)}, \dots, x^{(m)}\}$

$$x^{(i)} \in \mathbb{R}^n$$

K-means algorithm

Randomly initialize K cluster centroids $\mu_1, \mu_2, \dots, \mu_K \in \mathbb{R}^n$

Repeat {

 for $i = 1$ to m

$c^{(i)} :=$ index (from 1 to K) of cluster centroid
 closest to $x^{(i)}$

 for $k = 1$ to K

$\mu_k :=$ average (mean) of points assigned to cluster k

}

K-means optimization objective

$c^{(i)}$ = index of cluster $(1, 2, \dots, K)$ to which example $x^{(i)}$ is currently assigned

μ_k = cluster centroid \underline{k} ($\mu_k \in \mathbb{R}^n$)

$\mu_{c^{(i)}}$ = cluster centroid of cluster to which example $x^{(i)}$ has been assigned

K
 $k \in \{1, 2, \dots, K\}$
 $x^{(i)} \rightarrow 5$
 $\underline{c^{(i)}} = 5$
 $\underline{\mu_{c^{(i)}}} = \mu_5$

Optimization objective:

$$J(c^{(1)}, \dots, c^{(m)}, \mu_1, \dots, \mu_K) = \frac{1}{m} \sum_{i=1}^m \|x^{(i)} - \mu_{\underline{c^{(i)}}}\|^2$$

$$\min_{\substack{c^{(1)}, \dots, c^{(m)}, \\ \mu_1, \dots, \mu_K}} J(c^{(1)}, \dots, c^{(m)}, \mu_1, \dots, \mu_K)$$

One dimensional example. Why average the points?

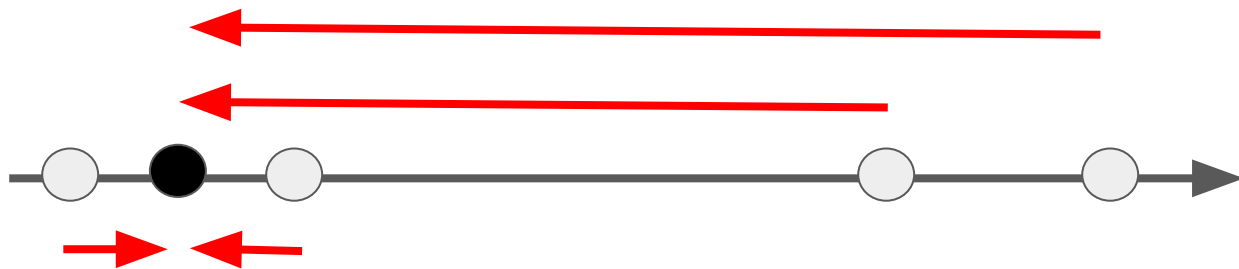
- We assume there is only a single cluster. The cluster center is blue/black circle. Which is a better cluster center? Black or blue?
- Red arrows shows the total error.
- Average of data points represents better cluster representation.



$$J(c^{(1)}, \dots, c^{(m)}, \mu_1, \dots, \mu_K) = \frac{1}{m} \sum_{i=1}^m \|x^{(i)} - \mu_{c(i)}\|^2$$

One dimensional example

- We assume there is only a single cluster. The cluster center is black circle
- Red arrows shows the total error.



$$J(c^{(1)}, \dots, c^{(m)}, \mu_1, \dots, \mu_K) = \frac{1}{m} \sum_{i=1}^m \|x^{(i)} - \mu_{c(i)}\|^2$$

One dimensional example

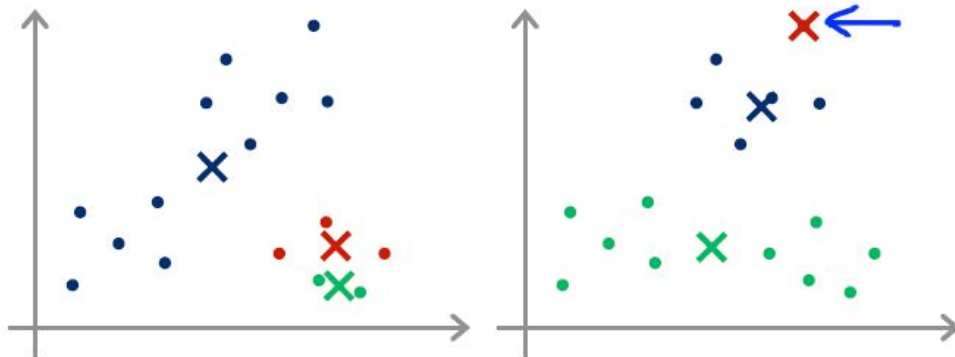
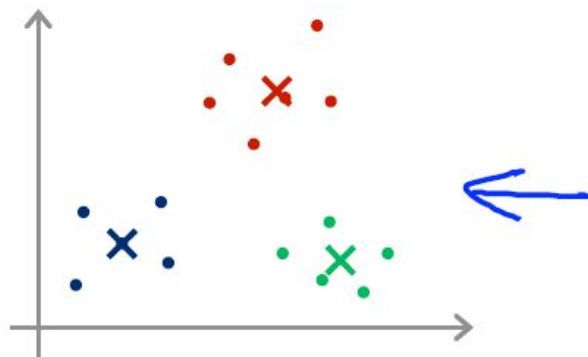
- Now we assume there are two clusters. The cluster centers are black circles.
- Red arrows shows the total error.
- Increasing clusters reduces the error.



$$J(c^{(1)}, \dots, c^{(m)}, \mu_1, \dots, \mu_K) = \frac{1}{m} \sum_{i=1}^m \|x^{(i)} - \mu_{c(i)}\|^2$$

Is there any failure case?

Local optima



How to avoid the local optima?

Random initialization

For $i = 1$ to 100 {

Randomly initialize K-means.

Run K-means. Get $c^{(1)}, \dots, c^{(m)}, \mu_1, \dots, \mu_K$.

Compute cost function (distortion)

$$J(c^{(1)}, \dots, c^{(m)}, \mu_1, \dots, \mu_K)$$

}

Pick clustering that gave lowest cost $J(c^{(1)}, \dots, c^{(m)}, \mu_1, \dots, \mu_K)$

One dimensional example



How to randomly initialize the clusters centers?

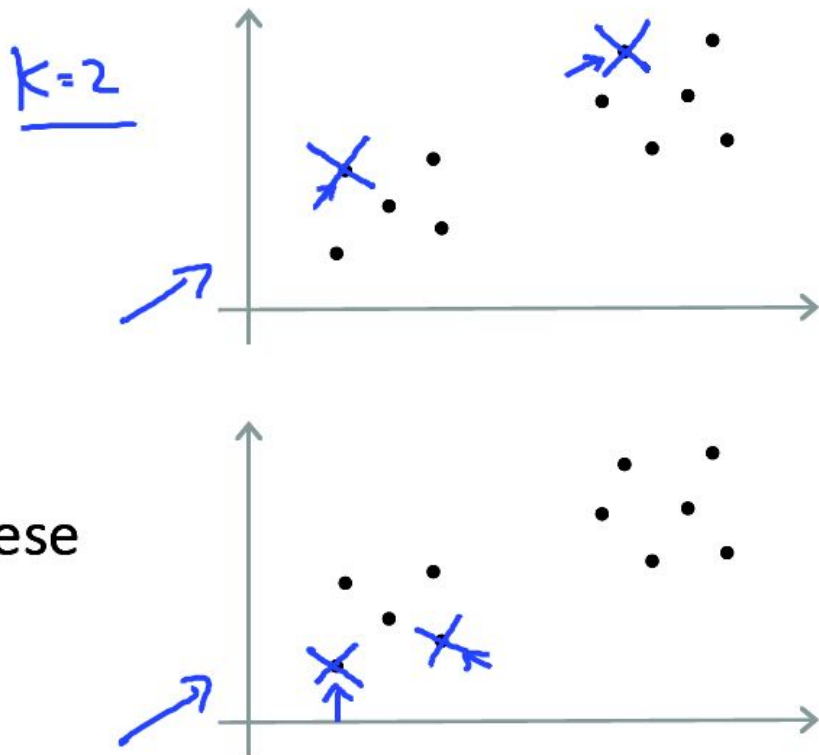
Random initialization

Should have $K < m$

Randomly pick K training examples.

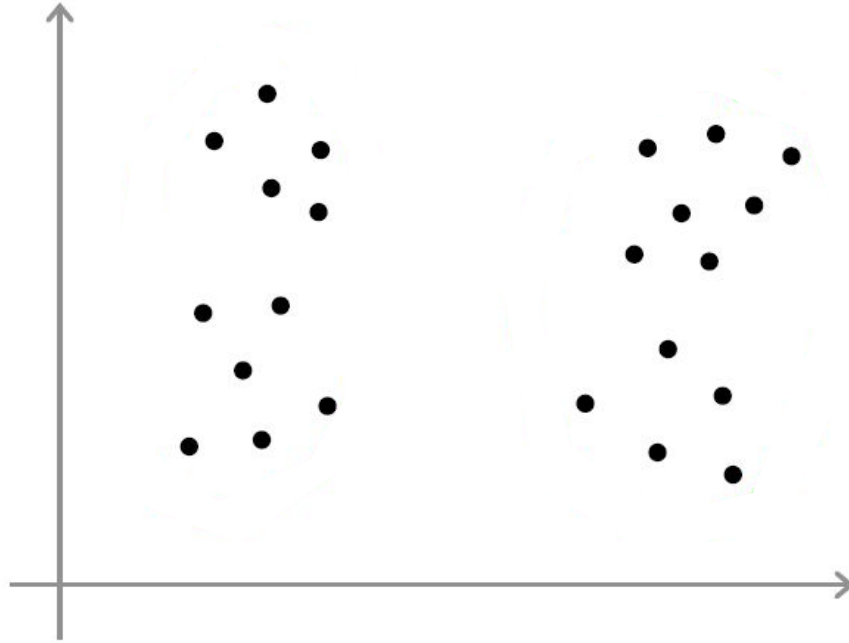
Set μ_1, \dots, μ_K equal to these K examples.

$$\begin{aligned}\mu_1 &= x^{(i)} \\ \mu_2 &= x^{(j)} \\ &\vdots\end{aligned}$$



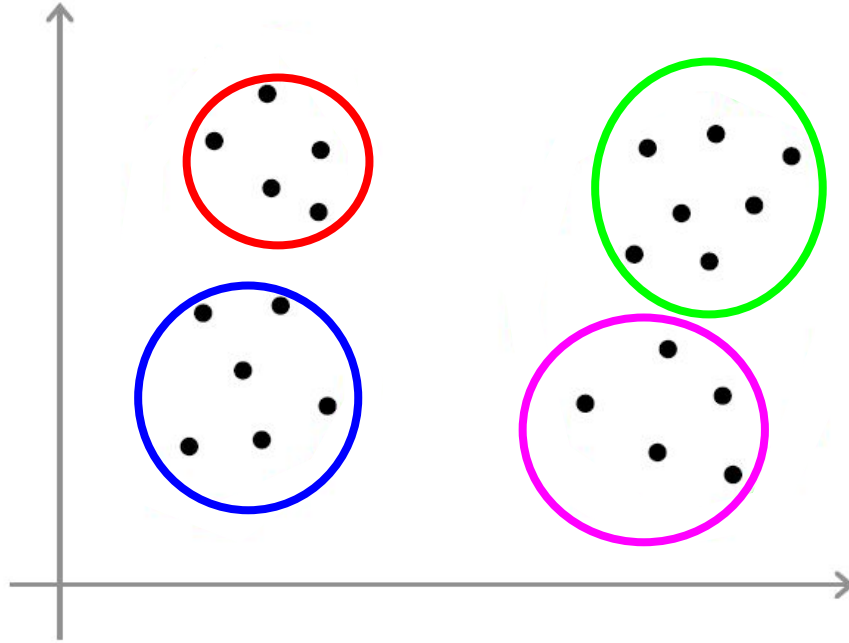
What is the right value of K?

What is the right value of K?



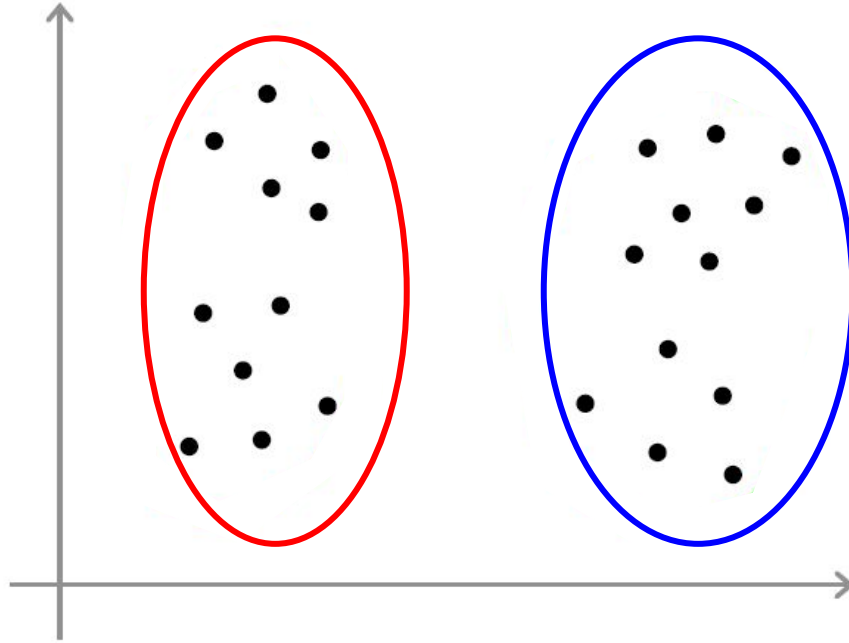
What is the right value of K? 4?

What is the right value of K?



What is the right value of K? 2?

What is the right value of K?



Choosing the value of K

Elbow method:

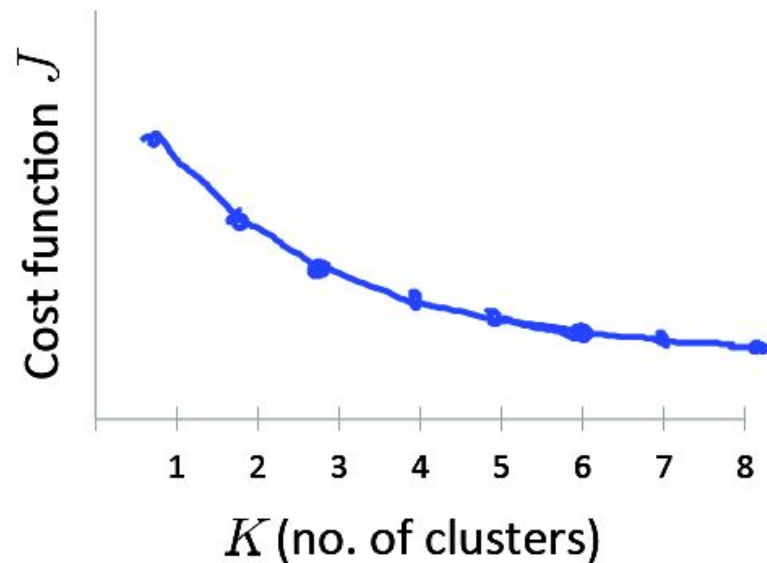
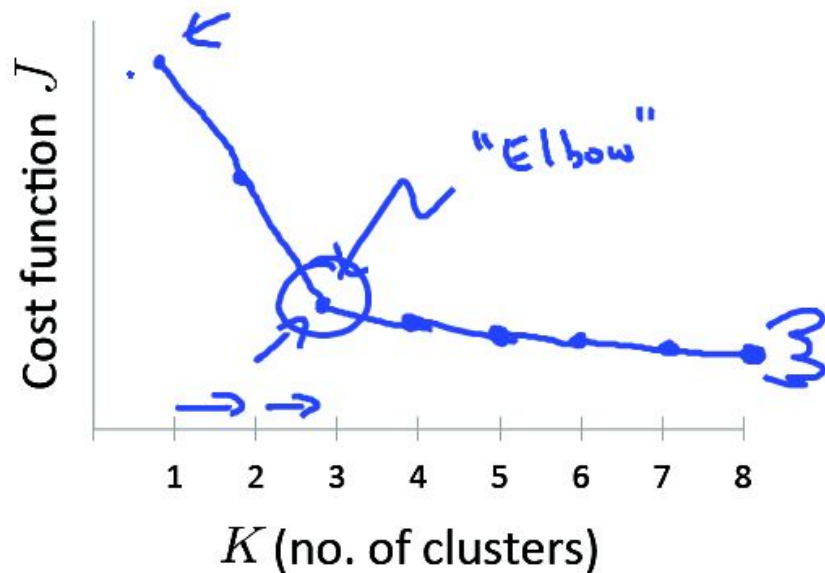


Image Compression Using K-means



Image Compression Using K-means



N pixels

R, G, B Channel

8 bits per value

Total bits = **$24 N$** bits

K clusters

Total bits = $24 K + N \log_2 K$ bits

<http://ieeexplore.ieee.org> > document



Love it or leave it? A new look at Signal Fidelity Measures

by Z Wang · 2009 · Cited by 2787 — In this article, we have reviewed the reasons why we (collectively) want to **love** or **leave** the venerable (but perhaps hoary) **MSE**.

DOI: 10.1109/MSP.2008.930649