Foundations of Machine Learning (ECE 5984)

- K-means/Gaussian Mixture Models -

Eunbyung Park

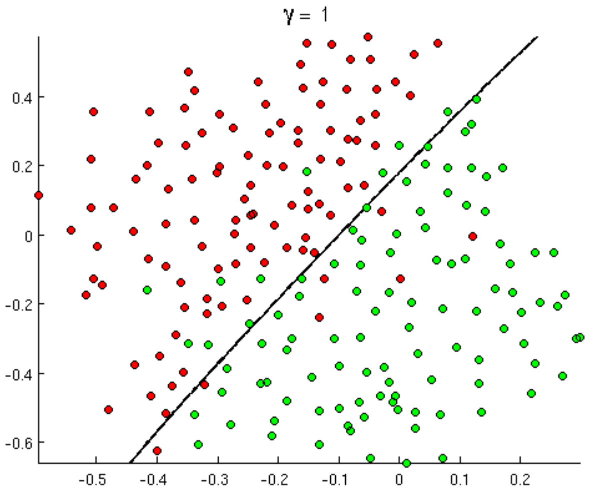
Assistant Professor

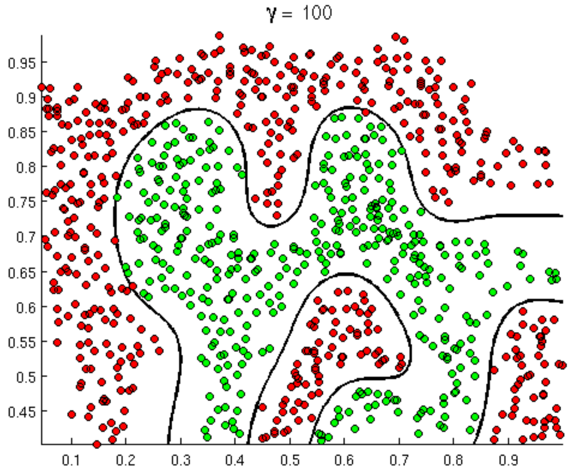
School of Electronic and Electrical Engineering

Eunbyung Park (silverbottlep.github.io)

Supervised Learning

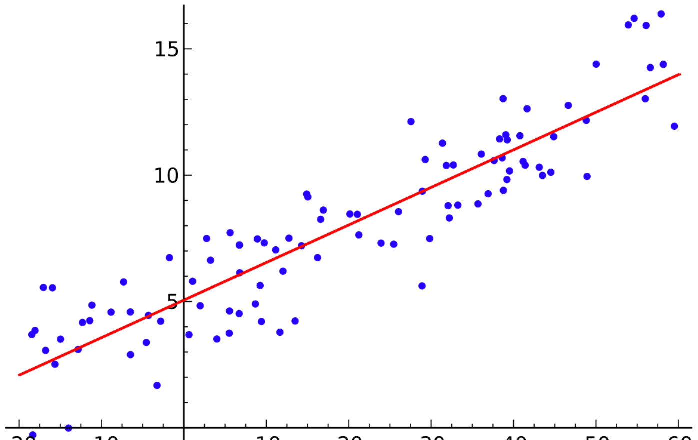
- Classification -> Learning boundaries
  - Logistic regression
  - Support Vector Machines (SVM)
  - K-nearest neighbors
  - Decision Trees, Neural networks

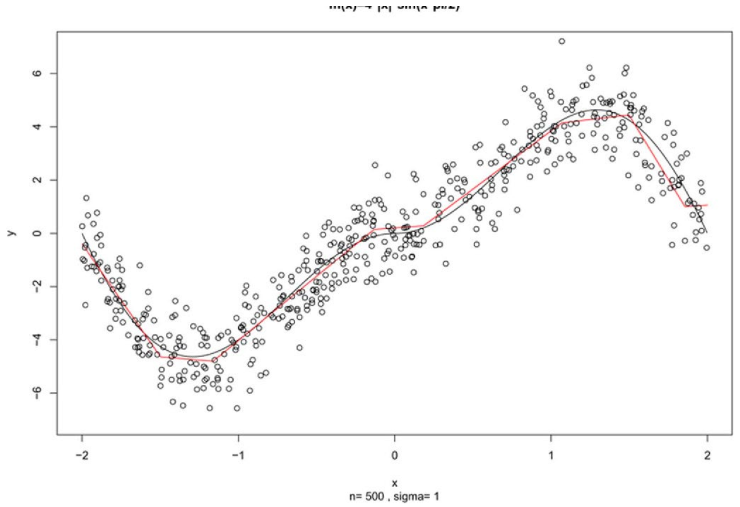




Supervised Learning

- Regression -> predicting real values
  - Linear regression
  - Polynomial regression
  - Neural networks
  - Gaussian process
  - Etc..

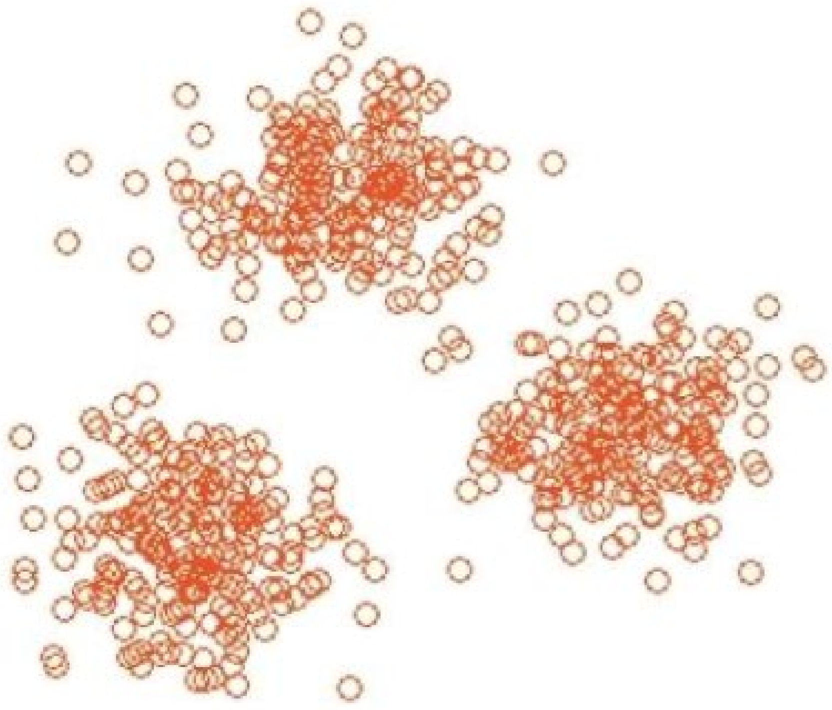




Slides adopted from prof Moon

Unsupervised Learning

- Clustering

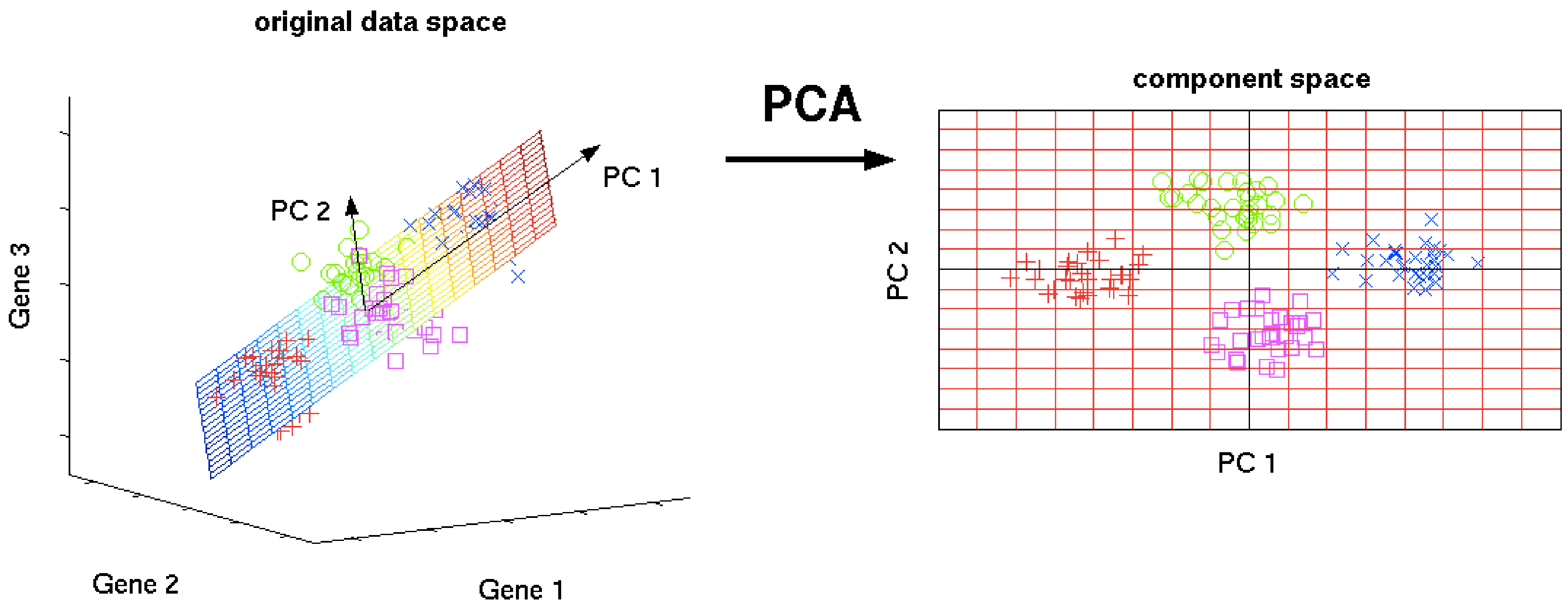


- Documents
- Users
- Webpages
- Diseases
- Pictures
- Vehicles

...

Unsupervised Learning

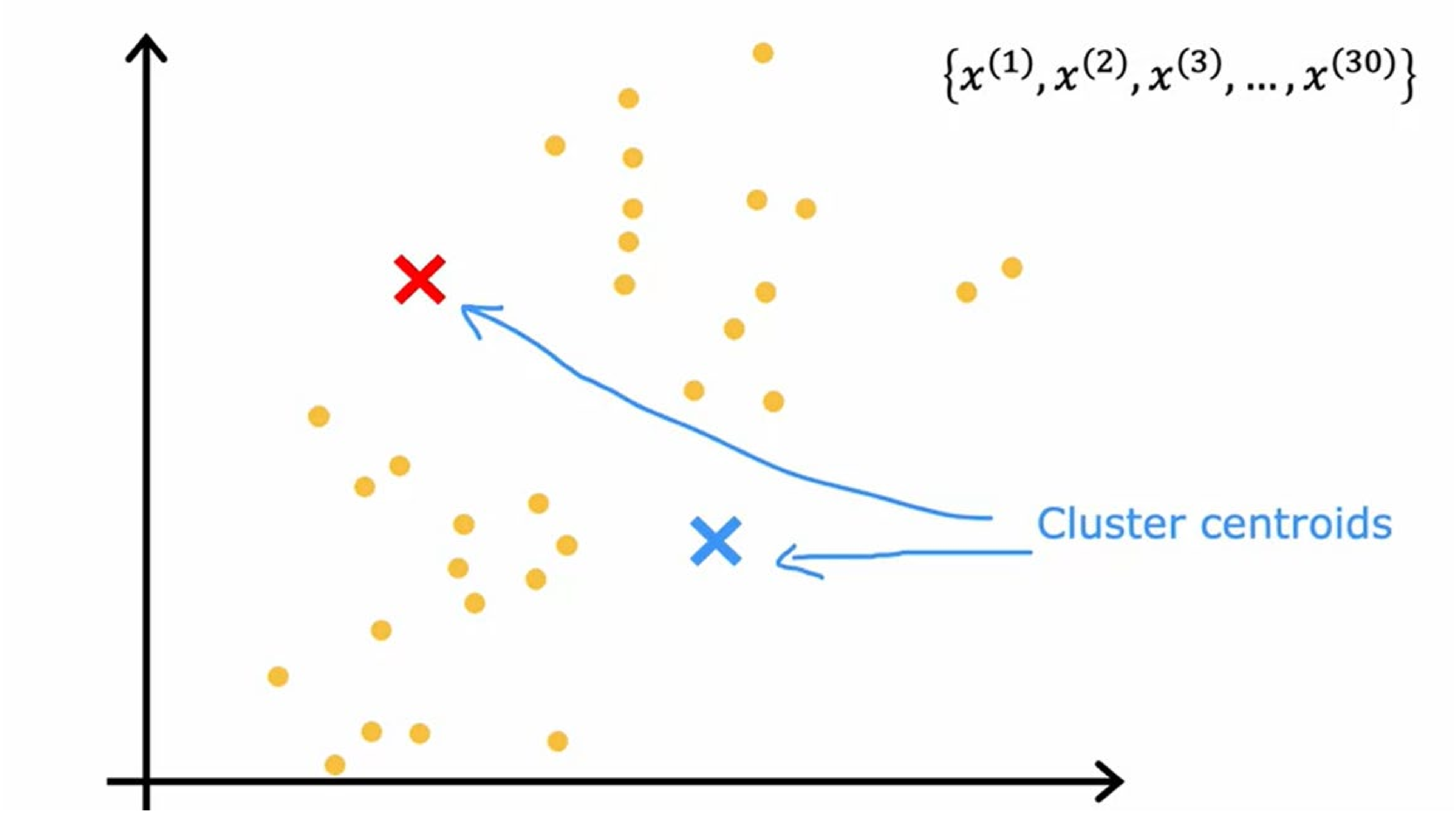
- Principal Component Analyses (Dimensionality reduction)



Slides adopted from prof Moon

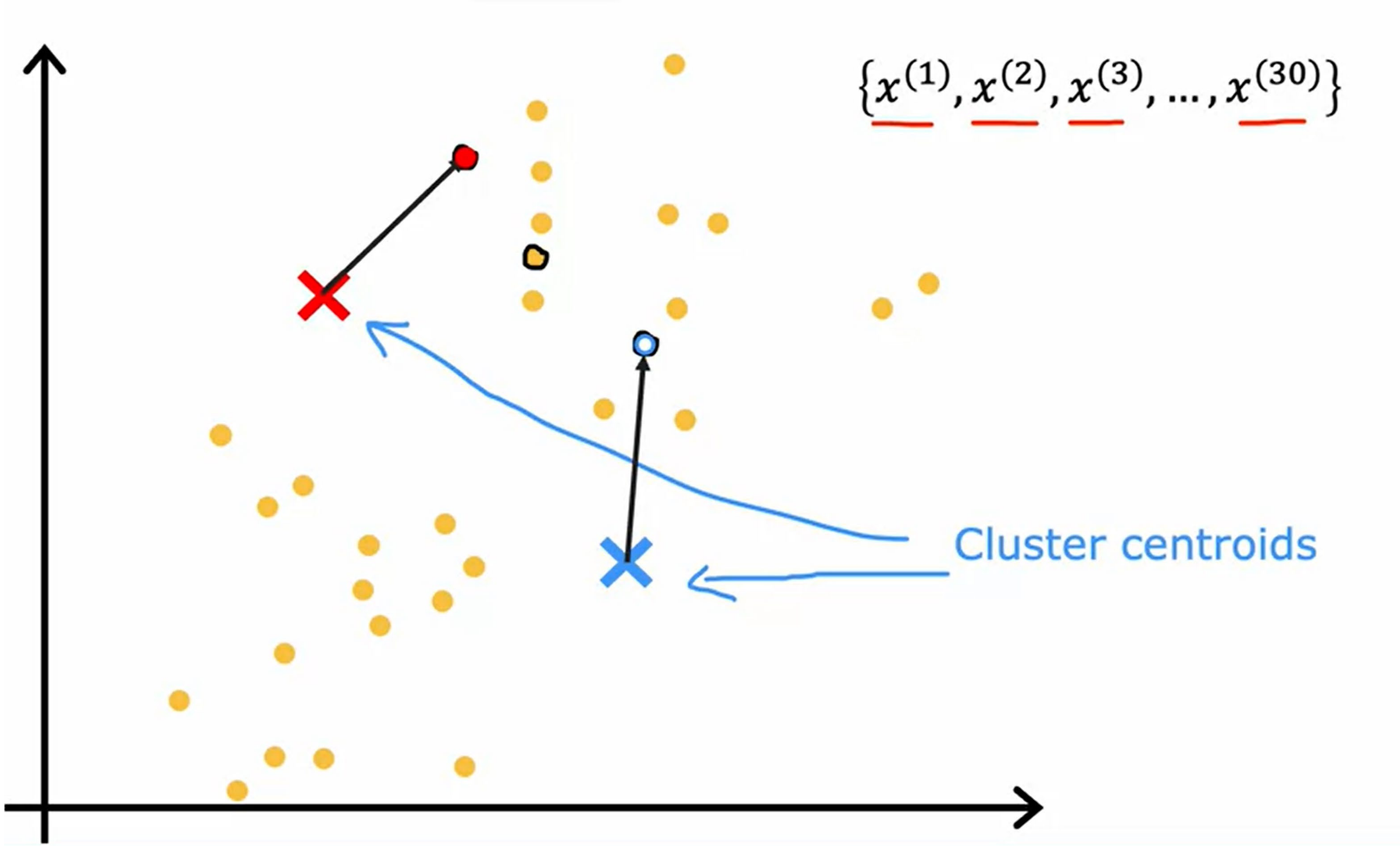
K-means Clustering

K-means



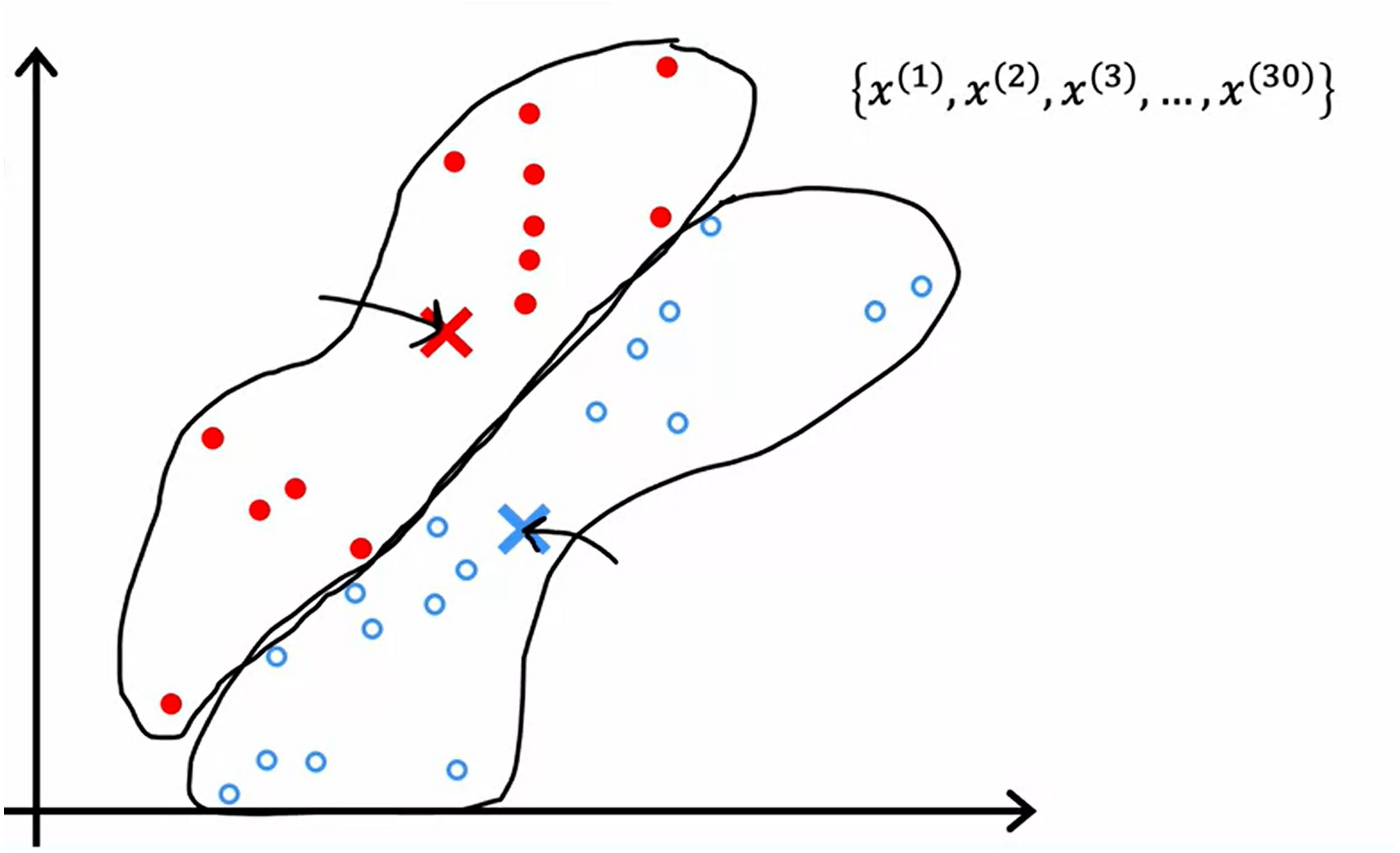
[Machine Learning Specialization [3 courses] (Stanford) | Coursera]

K-means

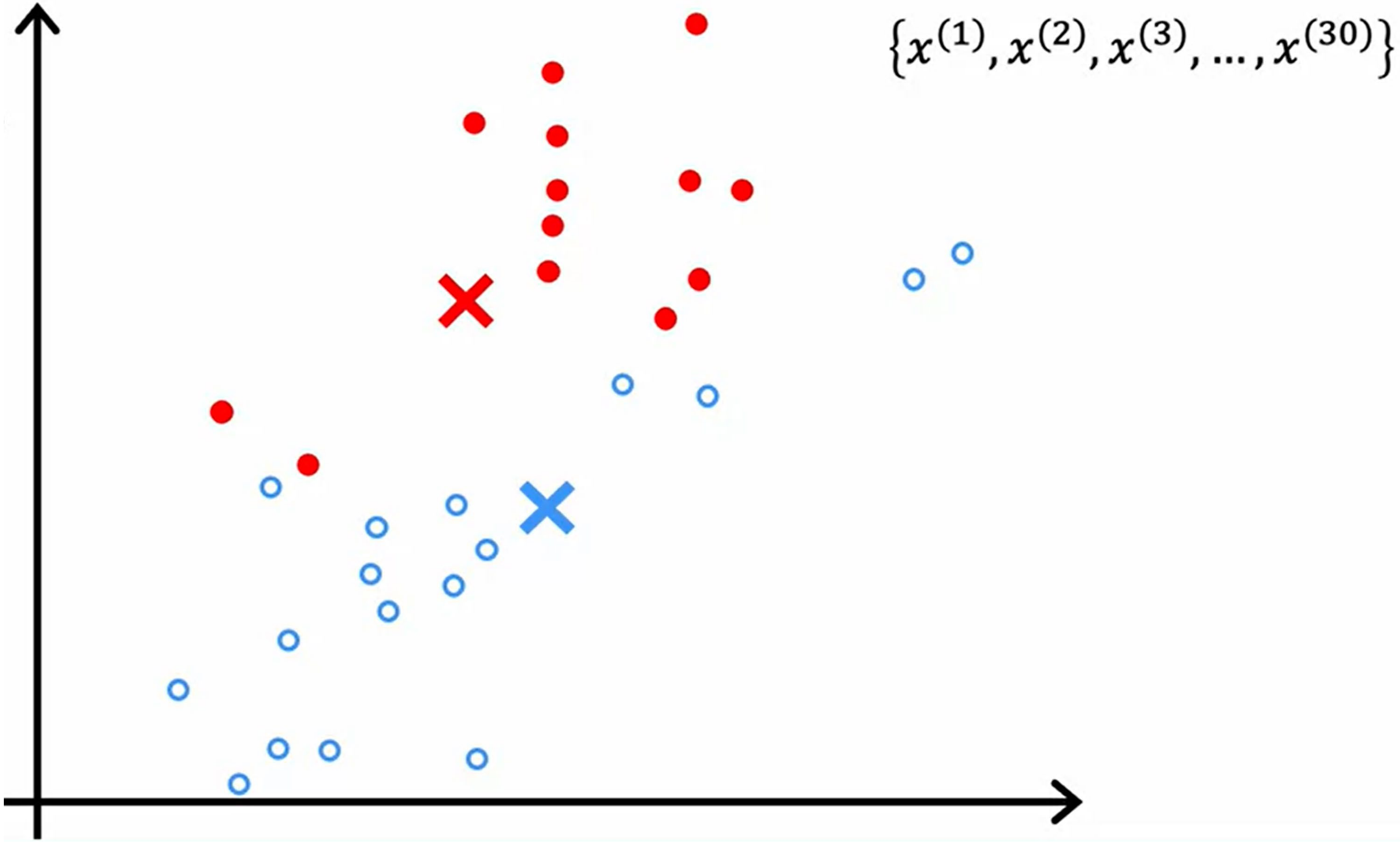


[Machine Learning Specialization [3 courses] (Stanford) | Coursera]

K-means

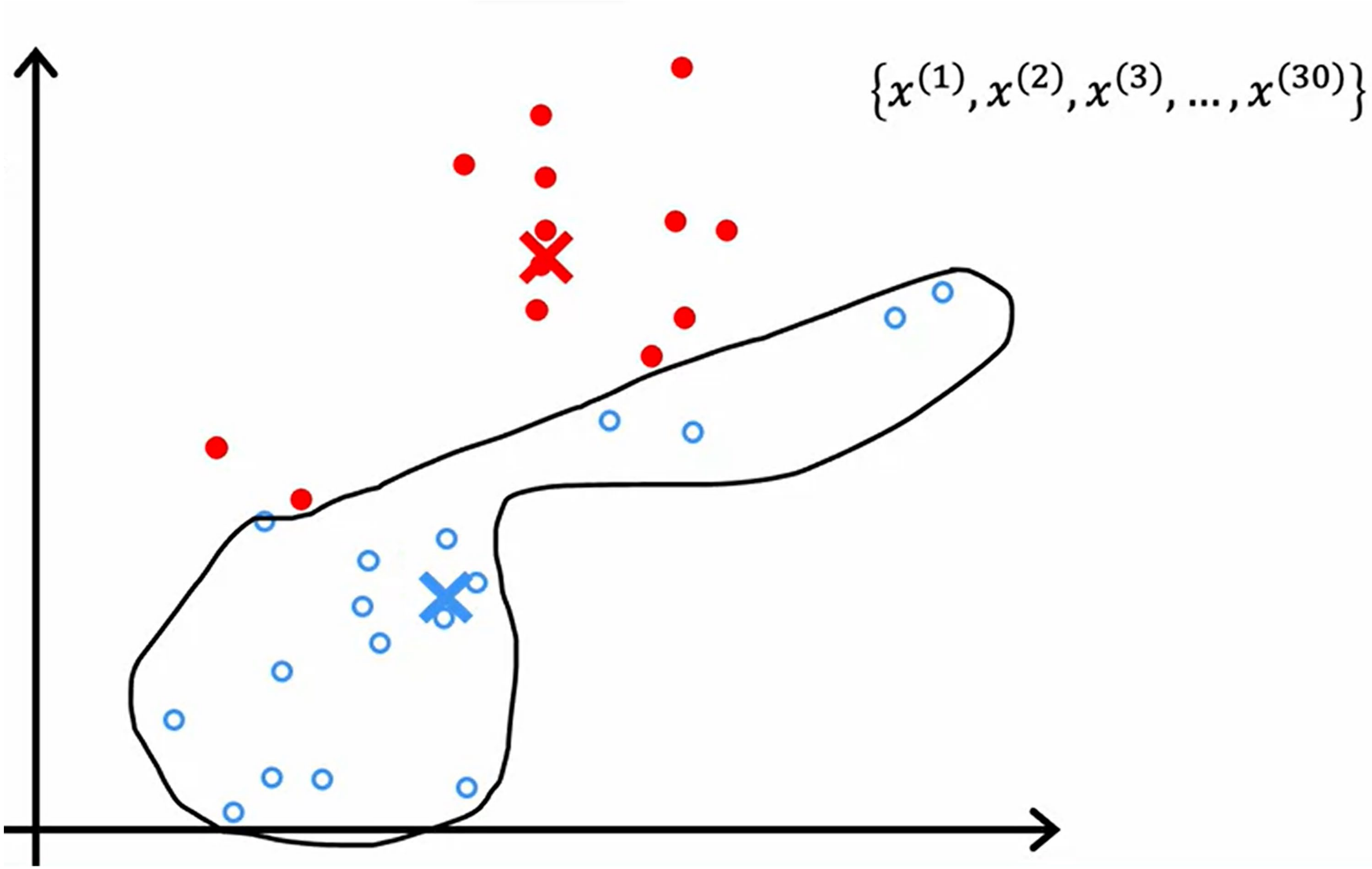


K-means



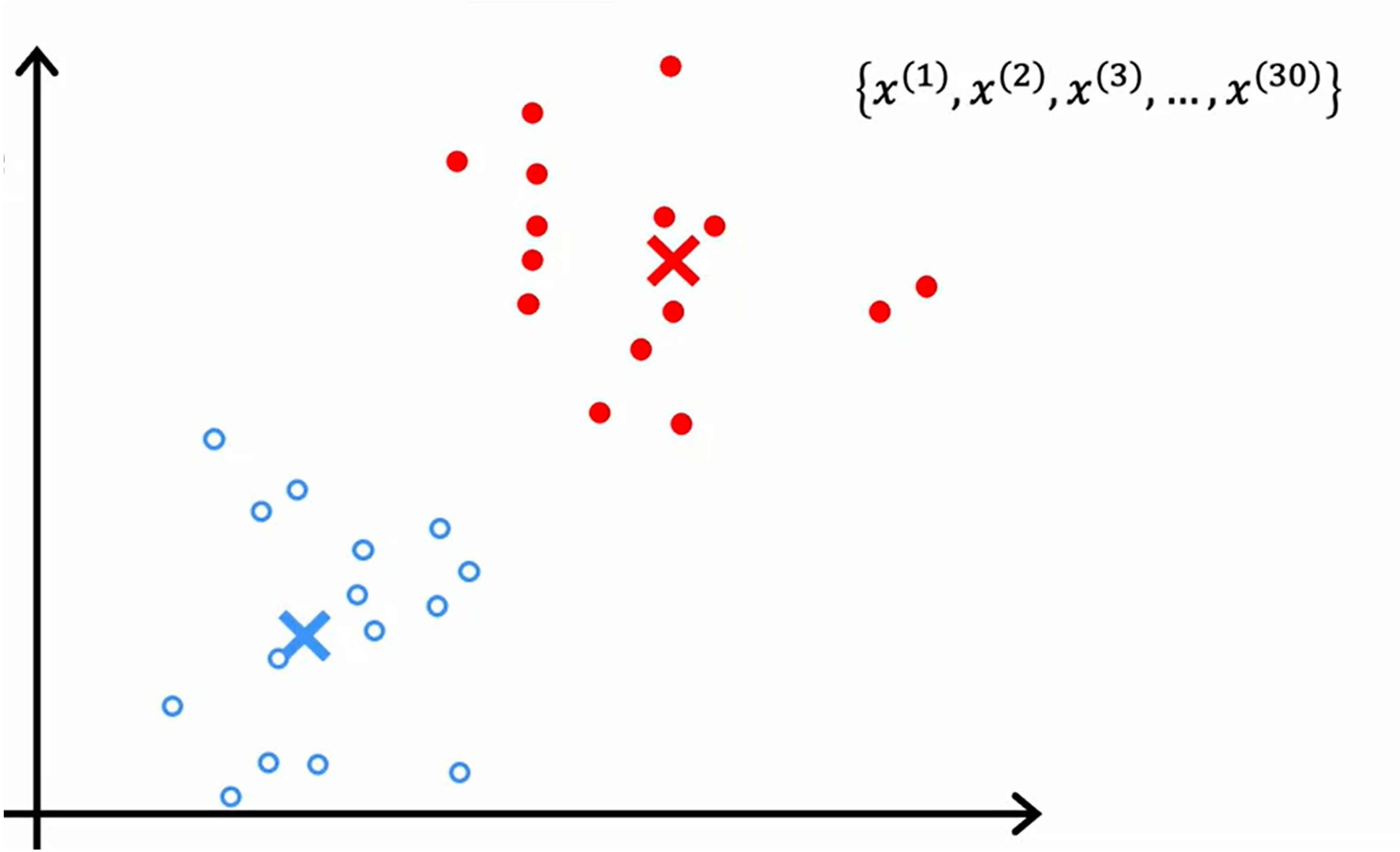
Machine Learning Specialization [3 courses] (Stanford) | Coursera

K-means



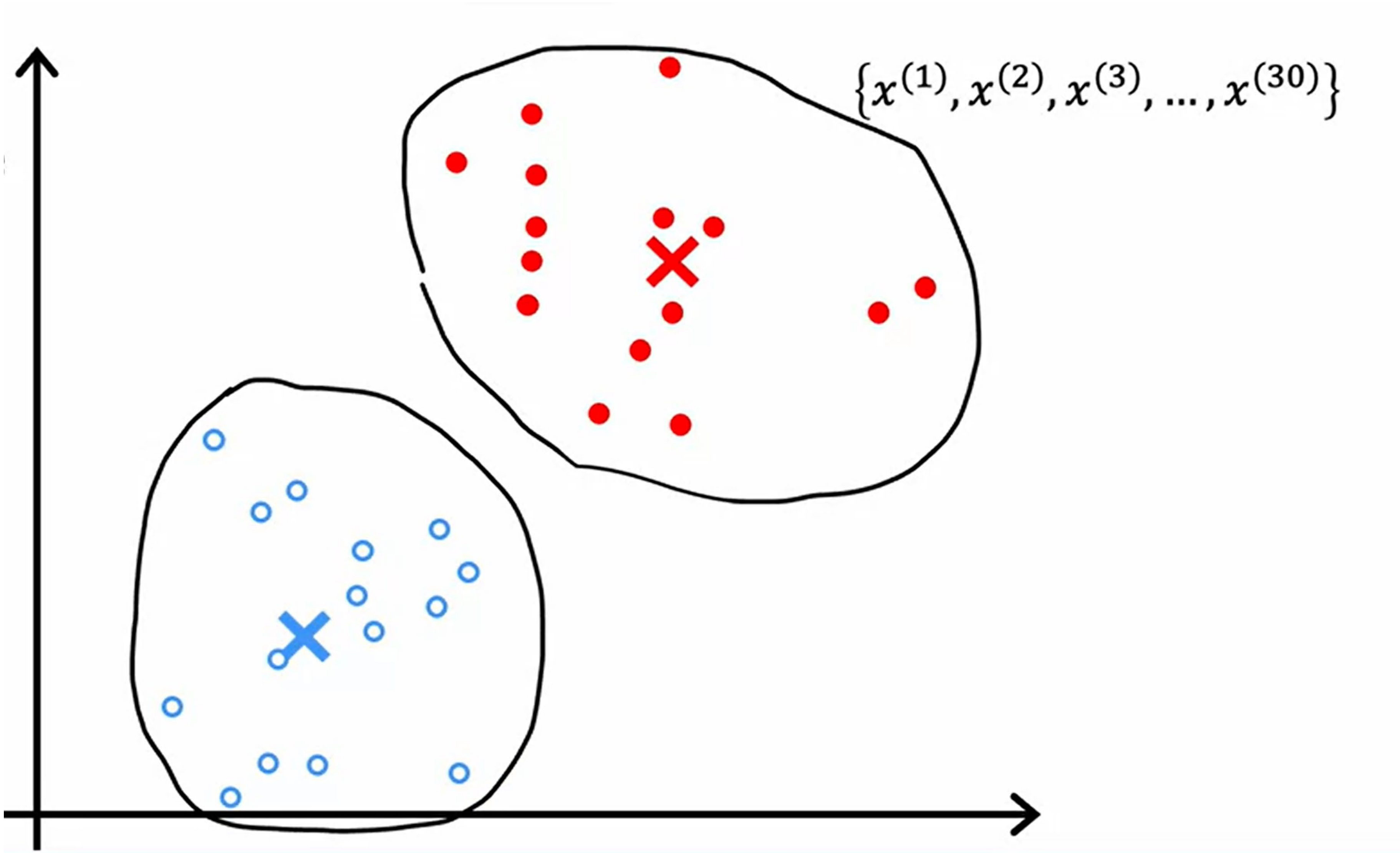
[Machine Learning Specialization [3 courses] (Stanford) | Coursera](https://www.coursera.org/specializations/machine-learning-introduction)

K-means



[Machine Learning Specialization [3 courses] (Stanford) | Coursera]

K-means



[Machine Learning Specialization [3 courses] (Stanford) | Coursera]

K-means Algorithm

Randomly initialize K cluster centroids $\mu_1, \mu_2, \mu_3, \ldots, \mu_K$

Repeat {

}

K-means Algorithm

Randomly initialize K cluster centroids $\mu_1, \mu_2, \mu_3, \dots, \mu_K$

Repeat {

For all $\forall i$,

$$c^{(i)} := \underset{j}{\operatorname{argmin}} \|x^{(i)} - \mu_j\|^2$$

Assign each training example $x^{(i)}$ to the closest cluster centroids μ_j

}

K-means Algorithm

Randomly initialize K cluster centroids $\mu_1, \mu_2, \mu_3, \dots, \mu_K$

Repeat {

For all $\forall i$,

$$c^{(i)} := \underset{j}{\operatorname{argmin}} \|x^{(i)} - \mu_j\|^2$$

Assign each training example $x^{(i)}$ to the closest cluster centroids μ_j

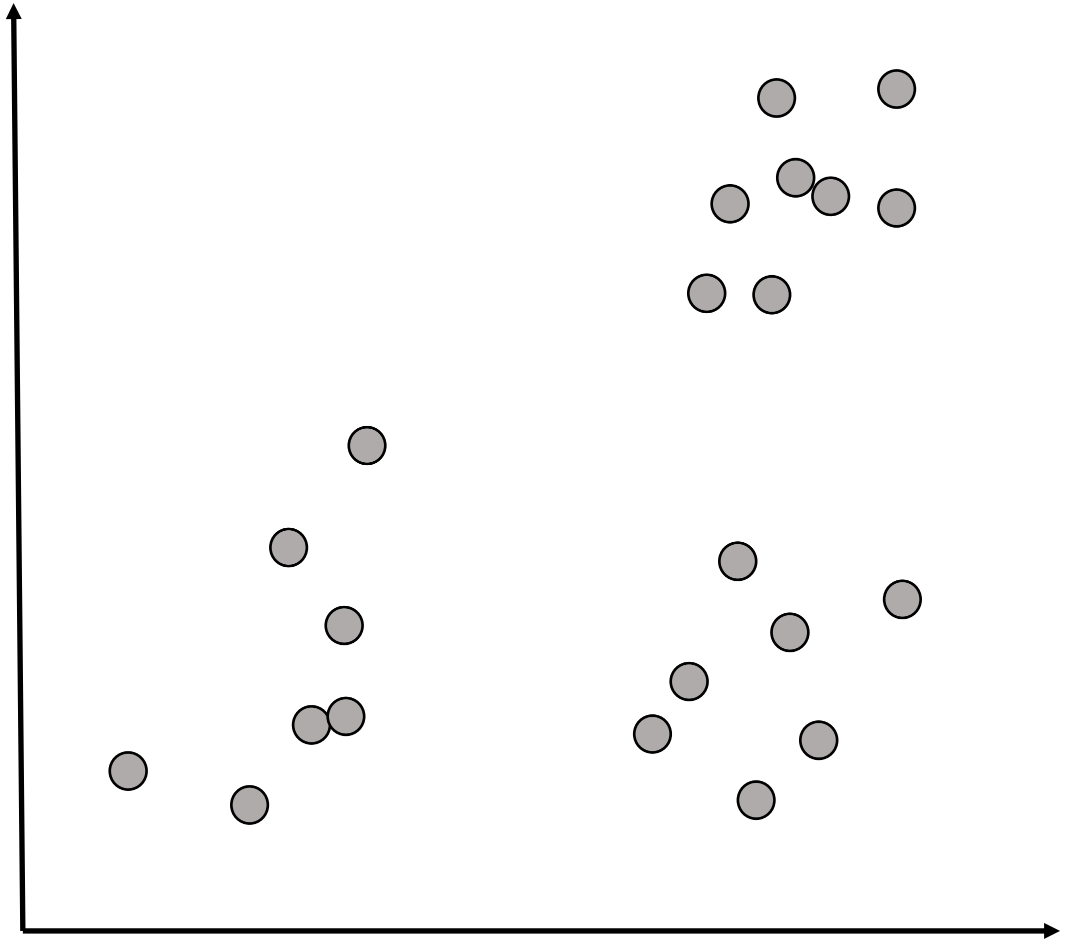
For all $\forall j$,

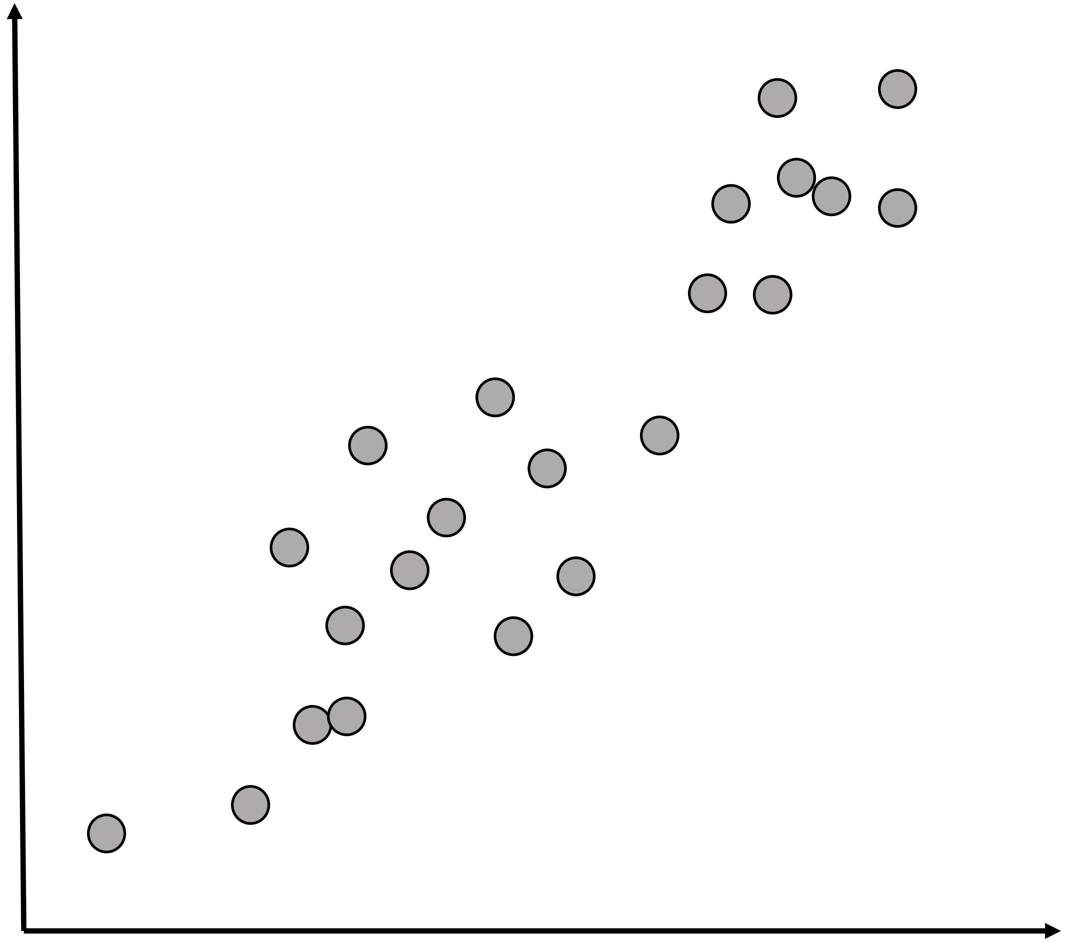
$$\mu_j := \frac{\sum_{i=1}^{m} 1\{c^{(i)} == j\} x^{(i)}}{\sum_{i=1}^{m} 1\{c^{(i)} == j\}}$$

Move μ_j to the mean of the points assigned to it

}

K-means for Ambiguous Data





K-means Optimization Objective

- Measure the Sum of Squared Errors (Distortion)
- K-means algorithms actually does ‘coordinate-descent on L’

$$L(c,\mu) = \underbrace{\sum_{i=1}^{m} \|x^{(i)} - \mu_{c^{(i)}}\|^2}_{\text{Distortion}} = \sum_{i=1}^{m}\sum_{k=1}^{K} 1\{c^{(i)} == k\}\|x^{(i)} - \mu_k\|^2$$

K-means Optimization Objective

- Measure the Sum of Squared Errors (Distortion)
- K-means algorithms actually does ‘coordinate-descent on L’

$$L(c,\mu) = \sum_{i=1}^{m} \|x^{(i)} - \mu_{c^{(i)}}\|^2 = \sum_{i=1}^{m}\sum_{k=1}^{K} 1\{c^{(i)} == k\}\|x^{(i)} - \mu_k\|^2$$

Distortion

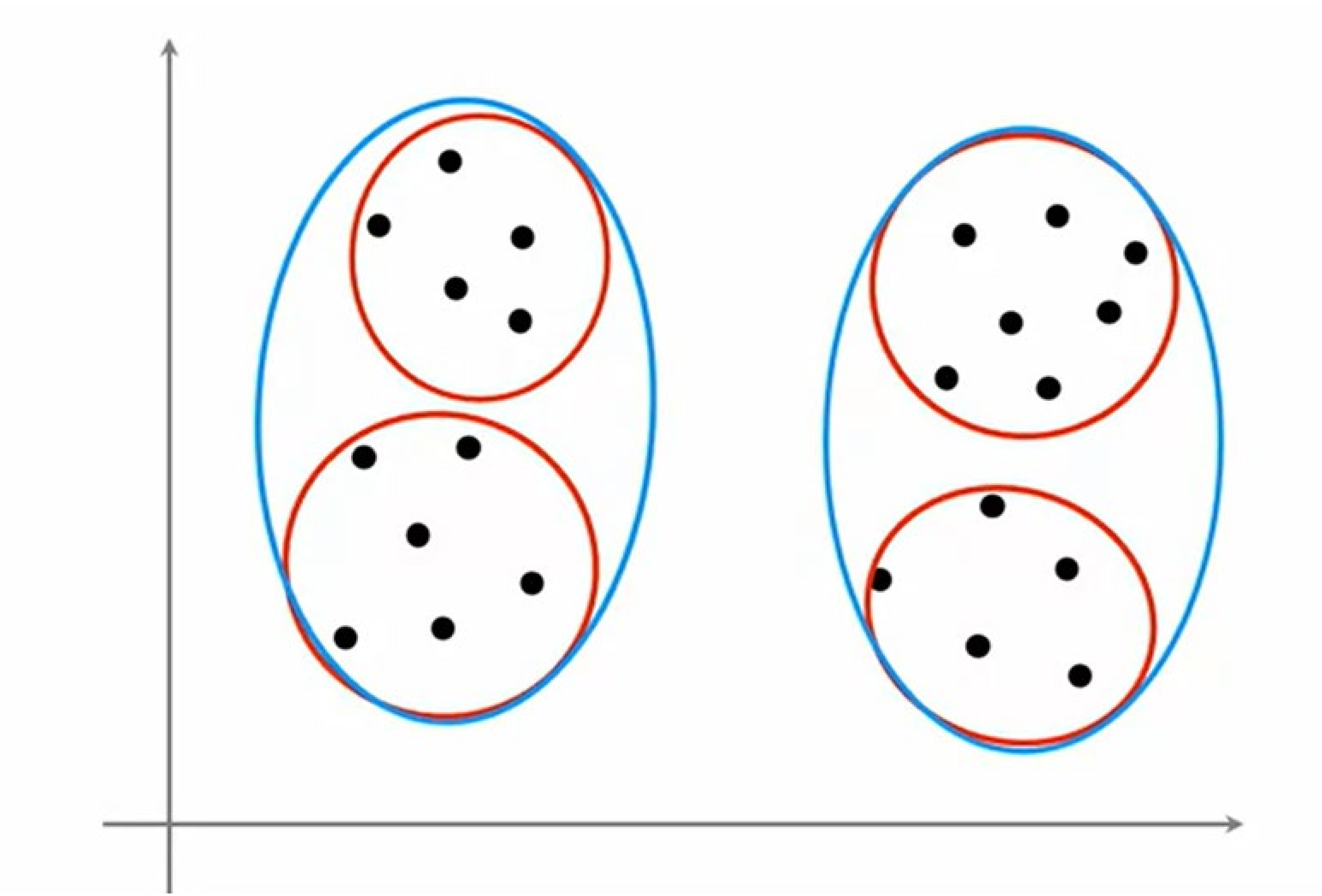
Convex?
Convergence?

K-means Initialization

- K < m
- Randomly pick K training examples
- Set μ_k equal to these K examples

Run multiple K-means with different initializations,
Then pick one that gave lowest loss $L(c, \mu)$

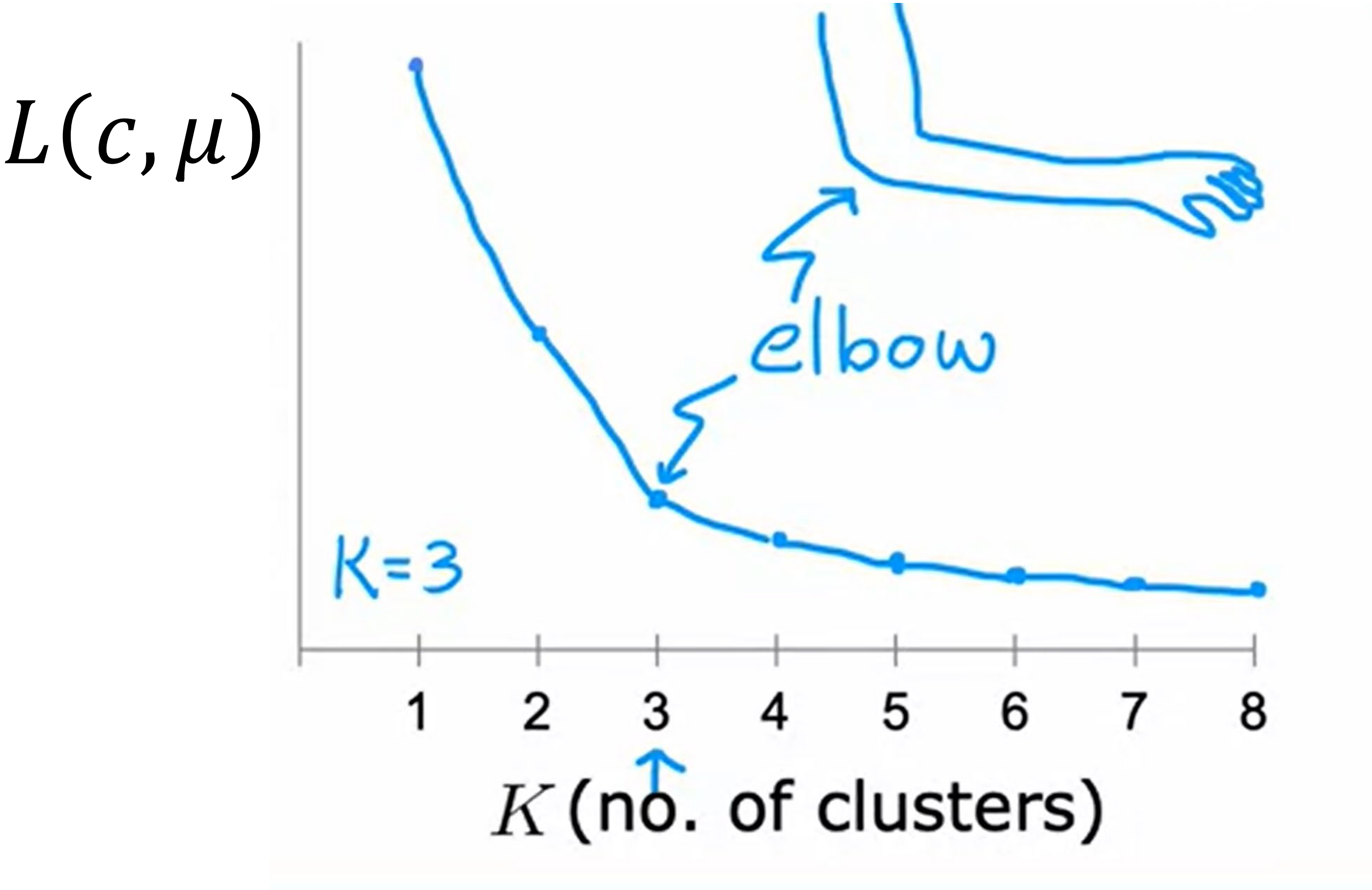
What is the Ideal K?



[Machine Learning Specialization [3 courses] (Stanford) | Coursera]

What is the Ideal K?

- Elbow method



[Machine Learning Specialization [3 courses] (Stanford) | Coursera]

What is the Ideal K?

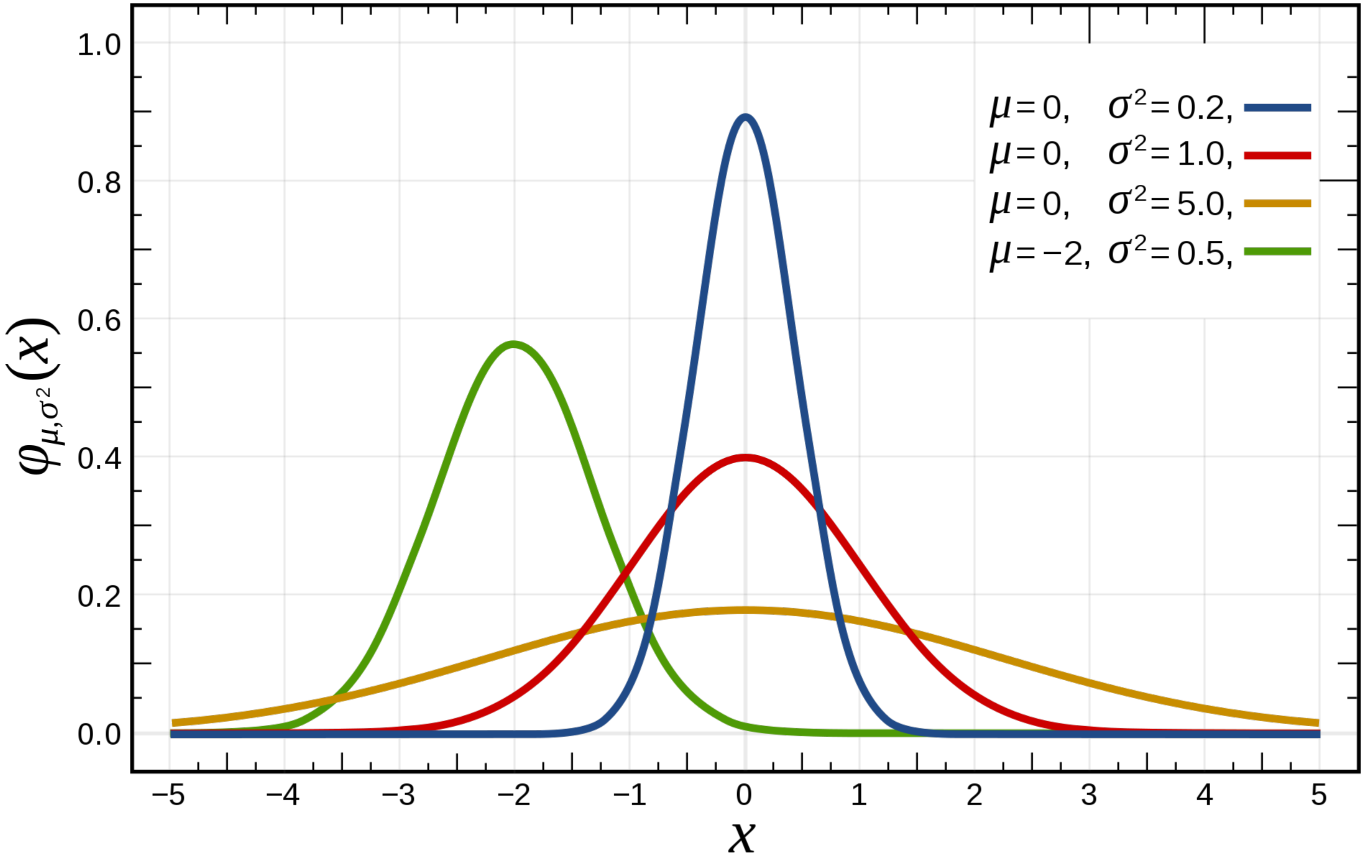
- Often, you want to get clusters for some later downstream tasks
- Evaluate K-means based on how well it performs on that tasks

Gaussian Mixture Models

Gaussian Distribution

- Normal distribution
- Widely used model for the distribution of continuous variable

$$p(x;\mu,\sigma^2)=\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

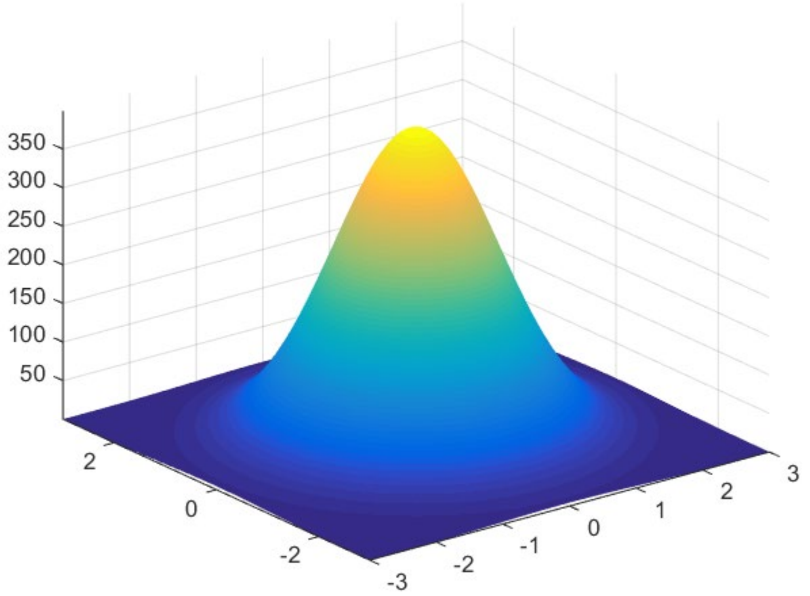


Multivariate Gaussian Distribution

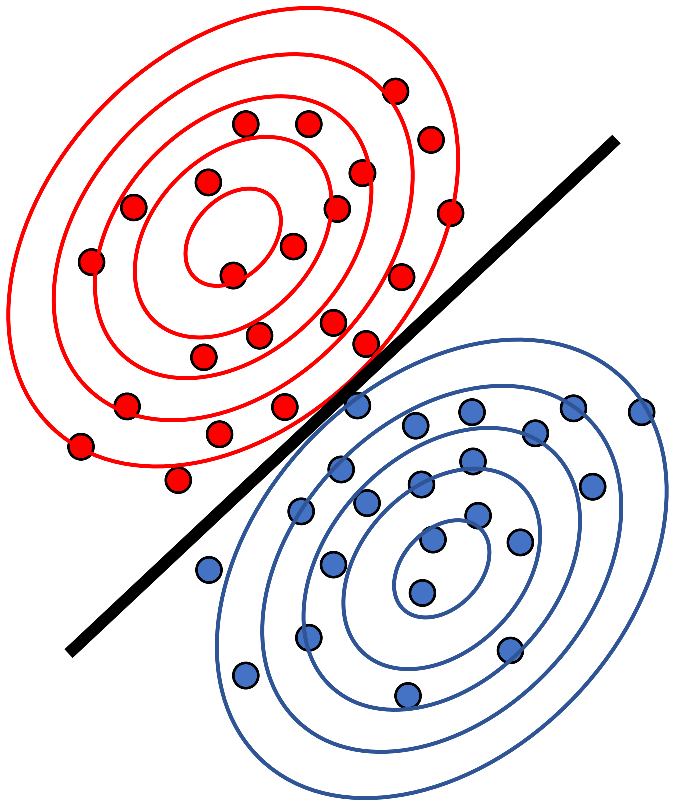
$x, \mu \in \mathbb{R}^d$

$\Sigma \in \mathbb{R}^{d \times d}$

$$p(x; \mu, \Sigma) = \frac{1}{(2\pi)^{d/2}} \frac{1}{|\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right)$$

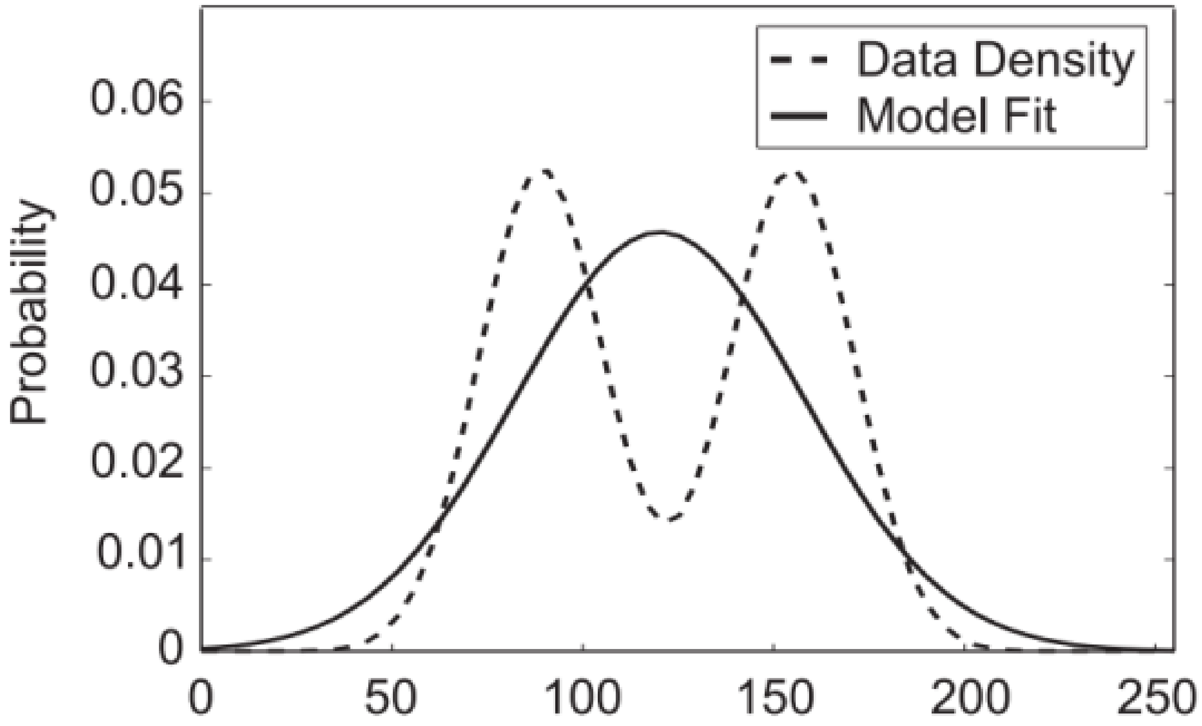


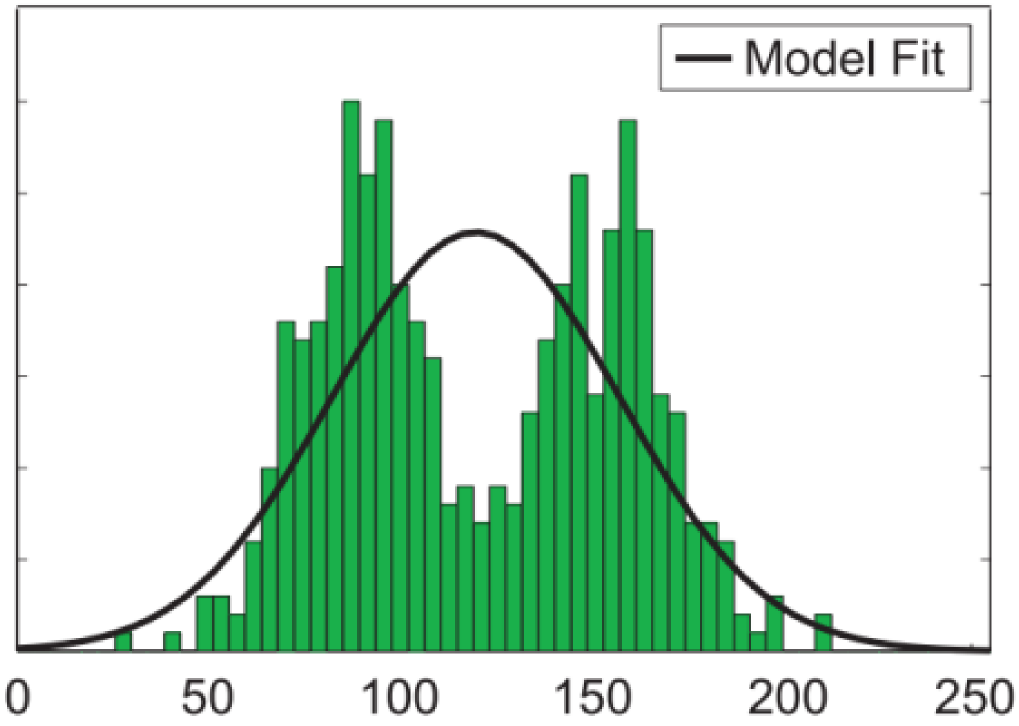
GDA



GMM Overview

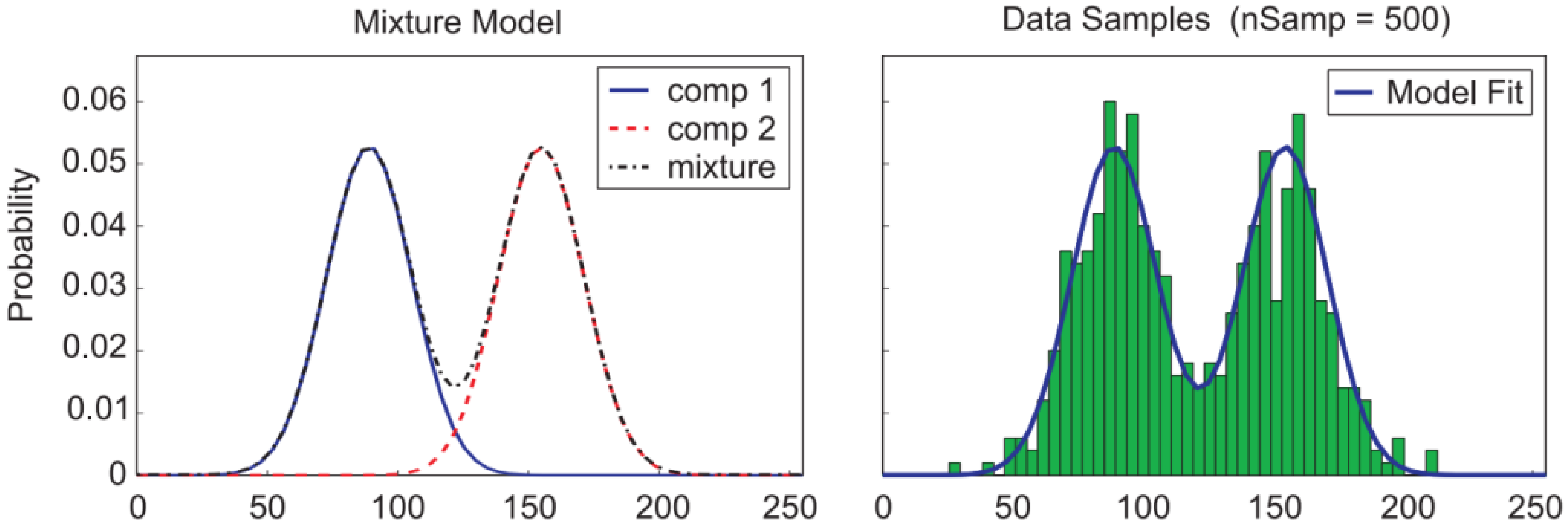
- A Gaussian





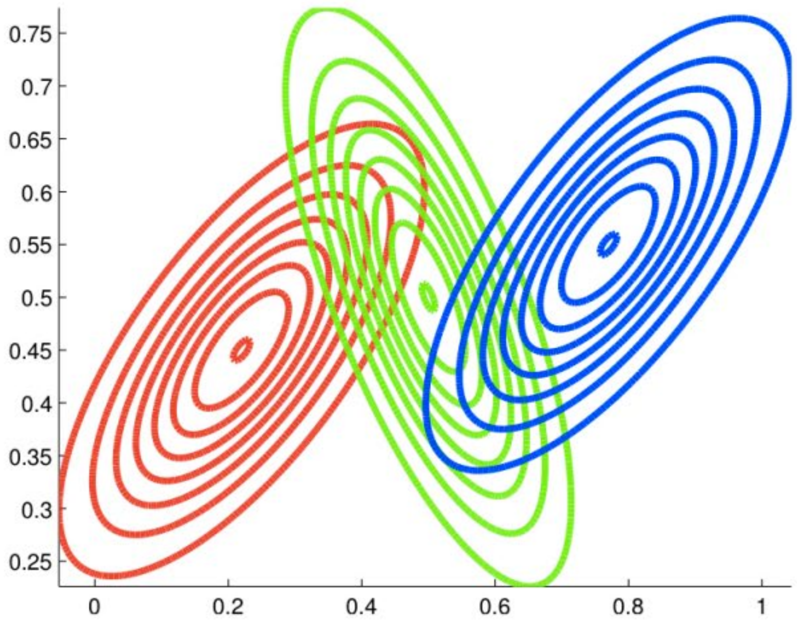
GMM Overview

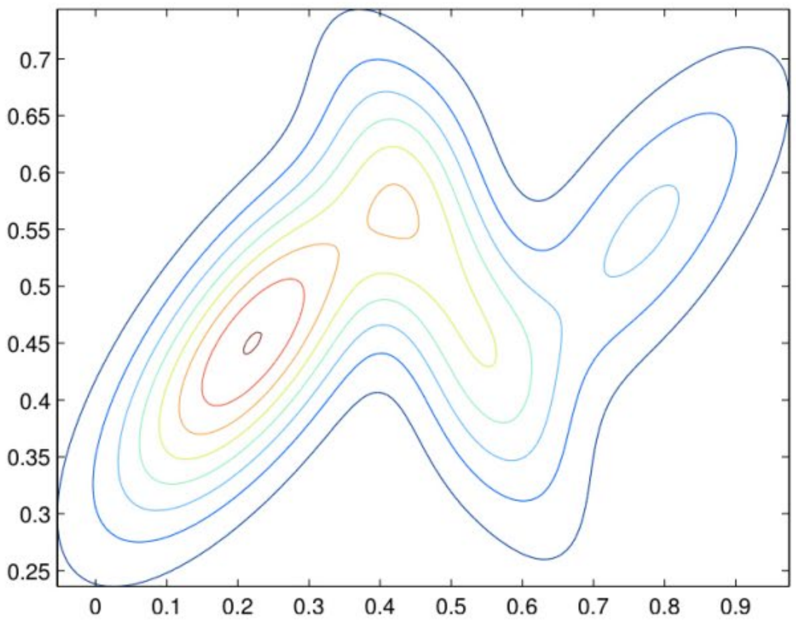
- A Mixture of Gaussians

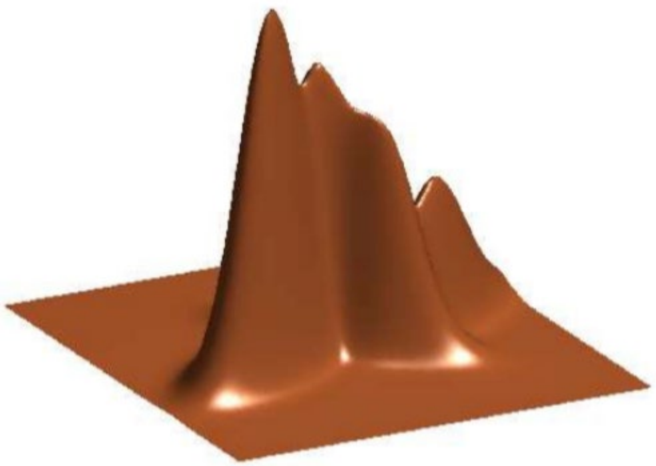


GMM Overview

- Visualizing a Mixture of Gaussians







Generative Models w/ Latent Variables

- We do not observe z and we only observe x, so we need to marginalize latent variables

$$p(x, z) = p(x|z)p(z)$$

(joint distribution, chain rule)

$$p(x) = \int p(x|z)p(z)dz$$

(z is latent variable, marginalization)

Maximum Likelihood

- Latent Variable - Categorical distribution

$$p(z) \coloneqq \mathrm{Cat}(\phi), \quad \text{where } \phi_j \geq 0, \quad \sum_{j=1}^{K} \phi_j = 1, \;\; p(z=j) = \phi_j$$

- Observed Variable - Gaussian distribution

$$p(x|z=j) \coloneqq N\left(x; \mu_j, \Sigma_j\right), \quad \mu_j \in \mathbb{R}^d, \Sigma_j \in \mathbb{R}^{d \times d}$$

Maximum Likelihood

- Log-likelihood

$$l(\phi,\mu,\Sigma) = \log \prod_{i=1}^{N} p\left(x^{(i)};\phi,\mu,\Sigma\right) = \sum_{i=1}^{N} \log p\left(x^{(i)};\phi,\mu,\Sigma\right)$$

$$= \sum_{i=1}^{N} \log \sum_{j=1}^{K} p\left(z^{(i)} = j;\phi\right) p\left(x^{(i)} | z^{(i)} = j;\mu_j,\Sigma_j\right)$$

$$= \sum_{i=1}^{N} \log \sum_{j=1}^{K} \phi_j \, N\left(x;\mu_j,\Sigma_j\right)$$

Can we do $\frac{dl}{d\mu_j} = 0$?

Maximum Likelihood

- What if we observed ‘z’?

$$l(\phi, \mu, \Sigma) = \log \prod_{i=1}^{N} p\left(x^{(i)}; \phi, \mu, \Sigma\right) = \sum_{i=1}^{N} \log p\left(x^{(i)}; \phi, \mu, \Sigma\right)$$

$$= \sum_{i=1}^{N} \log \sum_{j=1}^{K} p\left(z^{(i)} = j; \phi\right) p\left(x^{(i)} | z^{(i)} = j; \mu_j, \Sigma_j\right)$$

$$= \sum_{i=1}^{N} \log p\left(z^{(i)} = j; \phi\right) p\left(x^{(i)} | z^{(i)} = j; \mu_j, \Sigma_j\right)$$

Same as GDA!

EM Algorithm for GMM

(E-step): for each i, j:

$$w_j^{(i)} := p\left(z^{(i)} = j | x^{(i)}; \phi_j, \mu_j, \Sigma_j\right)$$

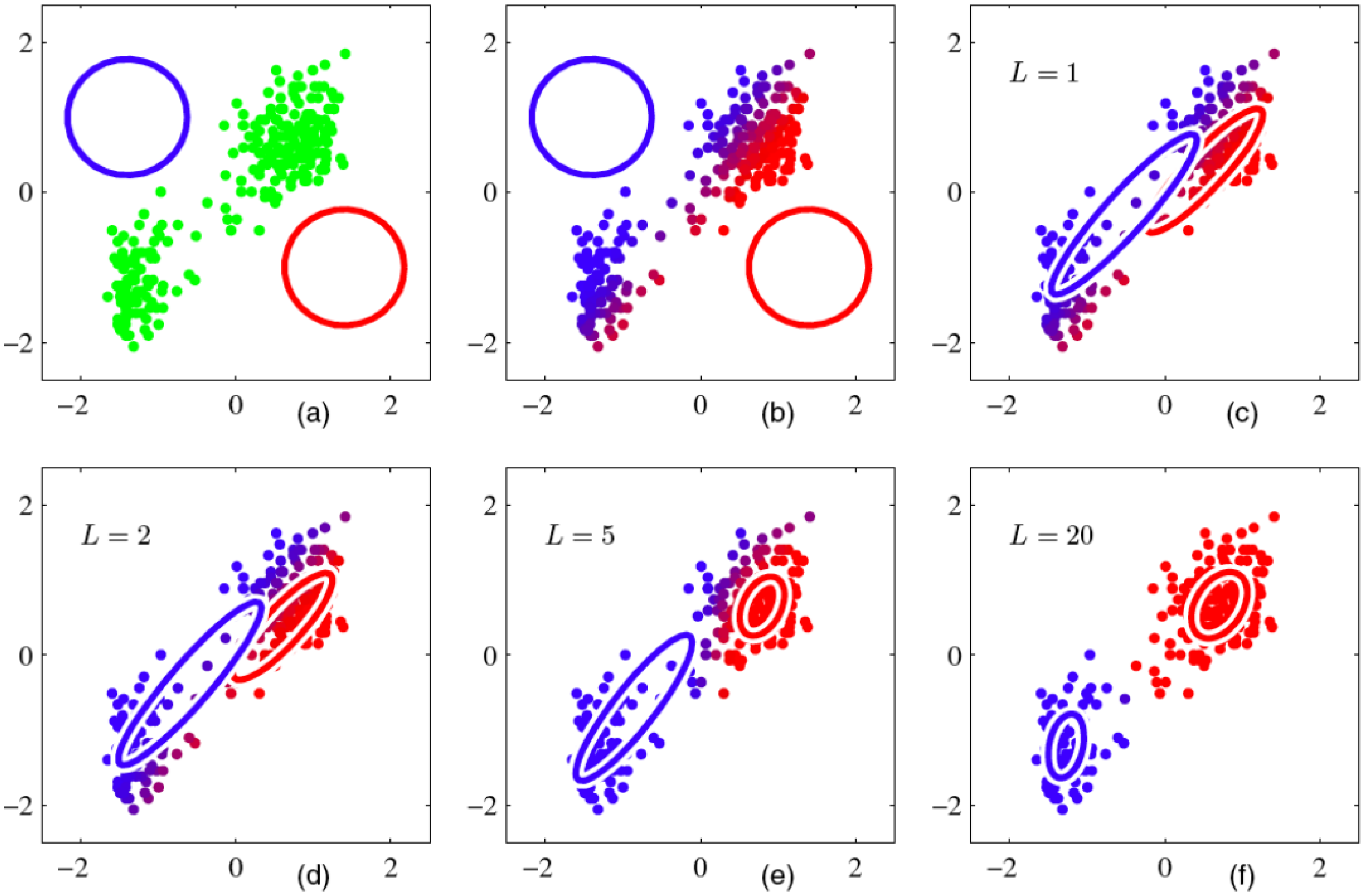
$$= \frac{p\left(x^{(i)} | z^{(i)} = j;\ \mu_j, \Sigma_j\right) p\left(z^{(i)} = j; \phi_j\right)}{\sum_{k=1}^{K} p(x^{(i)} | z^{(i)} = k;\ \mu_k, \Sigma_k) p(z^{(i)} = k; \phi_k)}$$

Posterior distribution
‘soft guess’
‘responsibility’

(M-step)

$$\phi_j := \frac{1}{N} \sum_{i=1}^{N} w_j^{(i)} \qquad \mu_j := \sum_{i=1}^{N} \frac{w_j^{(i)} x^{(i)}}{w_j^{(i)}} \qquad \Sigma_j := \sum_{i=1}^{N} \frac{w_j^{(i)} \left(x^{(i)} - \mu_j\right)\left(x^{(i)} - \mu_j\right)^{\mathsf{T}}}{w_j^{(i)}}$$

Examples



2
0
−2
−2
0
(a)
2
2
0
−2
−2
0
(b)
2
$L = 1$
2
0
−2
−2
0
(c)
2
$L = 2$
2
0
−2
−2
0
(d)
2
$L = 5$
2
0
−2
−2
0
(e)
2
$L = 20$
2
0
−2
−2
0
(f)
2