

07-380 AI & ML II

Latent Factors:

Clustering and k-means

GMM (Gaussian Mixture Models)

EM (Expectation-Maximization)

Instructors: Pat Virtue & Mohammad Salameh

Slide credits: CMU AI with drawings from <http://ai.berkeley.edu>

Clustering & Unsupervised Learning

Learning Paradigms

Paradigm	Data
Supervised	$\mathcal{D} = \{\mathbf{x}^{(i)}, y^{(i)}\}_{i=1}^N \quad \mathbf{x} \sim p^*(\cdot) \text{ and } y = c^*(\cdot)$
$\hookrightarrow$ Regression	$y^{(i)} \in \mathbb{R}$
$\hookrightarrow$ Classification	$y^{(i)} \in \{1, \dots, K\}$
$\hookrightarrow$ Binary classification	$y^{(i)} \in \{+1, -1\}$
$\hookrightarrow$ Structured Prediction	$\mathbf{y}^{(i)}$ is a vector
Unsupervised	$\mathcal{D} = \{\mathbf{x}^{(i)}\}_{i=1}^N \quad \mathbf{x} \sim p^*(\cdot)$
Semi-supervised	$\mathcal{D} = \{\mathbf{x}^{(i)}, y^{(i)}\}_{i=1}^{N_1} \cup \{\mathbf{x}^{(j)}\}_{j=1}^{N_2}$
Online	$\mathcal{D} = \{(\mathbf{x}^{(1)}, y^{(1)}), (\mathbf{x}^{(2)}, y^{(2)}), (\mathbf{x}^{(3)}, y^{(3)}), \dots\}$
Active Learning	$\mathcal{D} = \{\mathbf{x}^{(i)}\}_{i=1}^N$ and can query $y^{(i)} = c^*(\cdot)$ at a cost
Imitation Learning	$\mathcal{D} = \{(s^{(1)}, a^{(1)}), (s^{(2)}, a^{(2)}), \dots\}$
Reinforcement Learning	$\mathcal{D} = \{(s^{(1)}, a^{(1)}, r^{(1)}), (s^{(2)}, a^{(2)}, r^{(2)}), \dots\}$

Clustering, Informal Goals

Goal: Automatically partition **unlabeled** data into groups of similar datapoints.

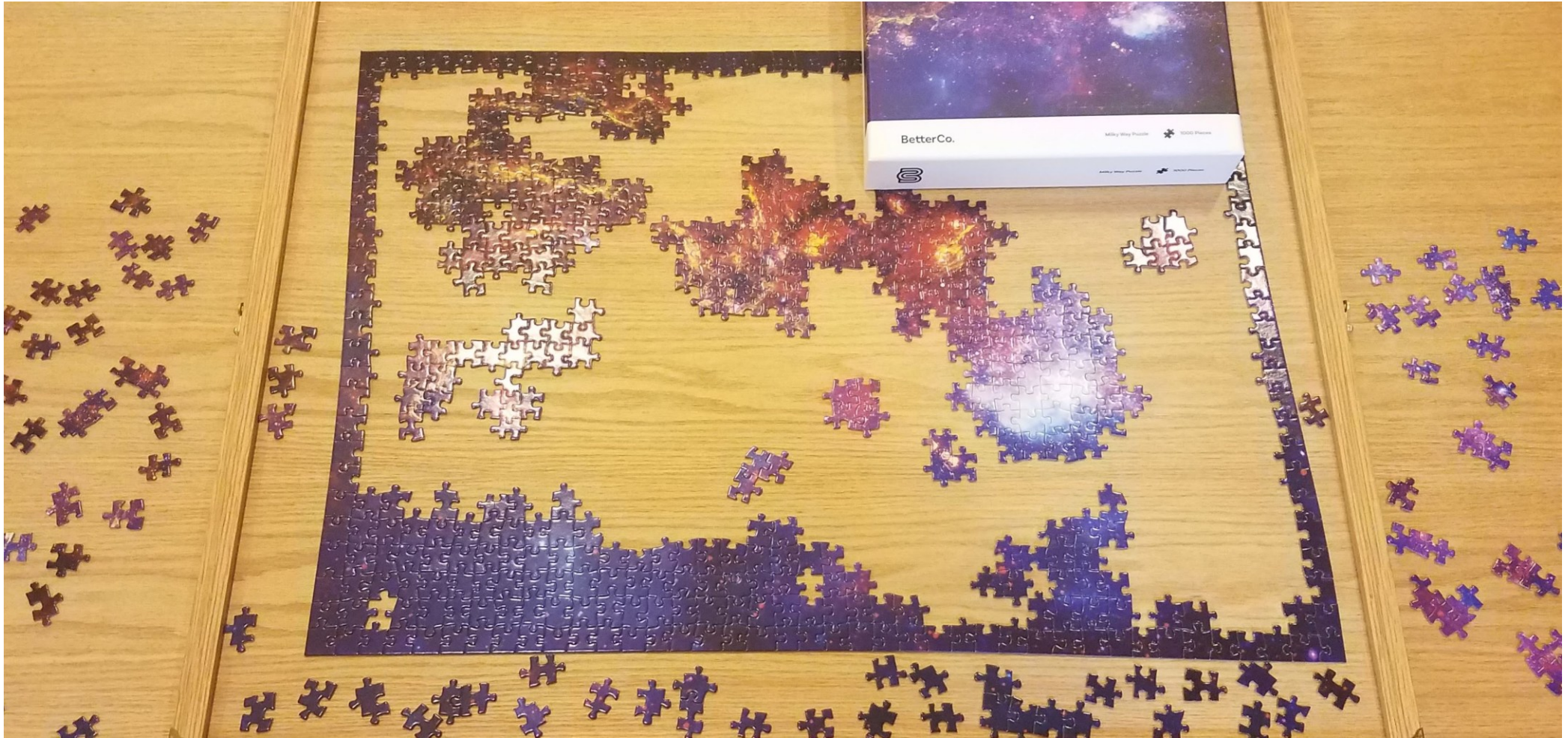
Question: When and why would we want to do this?

Useful for:

- Automatically organizing data.
- Understanding hidden structure in data.
- Preprocessing for further analysis.
 - Representing high-dimensional data in a low-dimensional space (e.g., for visualization purposes).

Clustering Applications

Jigsaw puzzles!



Semi-supervised Learning

Supervised: $\mathcal{D} = \{\mathbf{x}^{(i)}, y^{(i)}\}_{i=1}^N$

Unsupervised: $\mathcal{D} = \{\mathbf{x}^{(i)}\}_{i=1}^N$

Google News

ON

Nor'easter dumps heavy snow and cuts off power to hundreds of thousands across Northeast as many roads remain impassable

4 hours ago · Gene Norman, Artemis Moshtagian & ...



FOX WEATHER

The Daily Weather Update from FOX Weather: Nor'easter continues dumping snow on New England

56 minutes ago



NASA Blogs

Engineers Pinpoint Cause of Voyager 1 Issue, Are Working on Solution – Voyager

17 hours ago

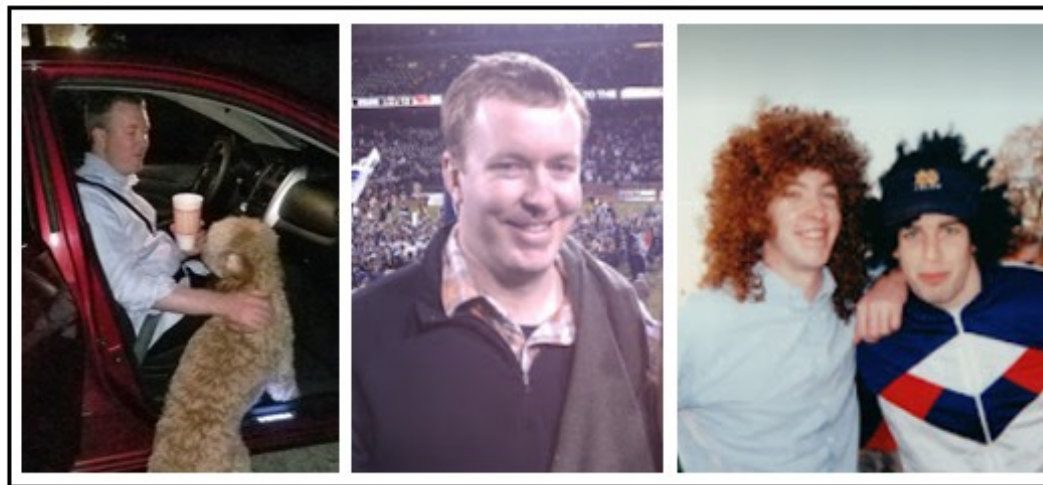
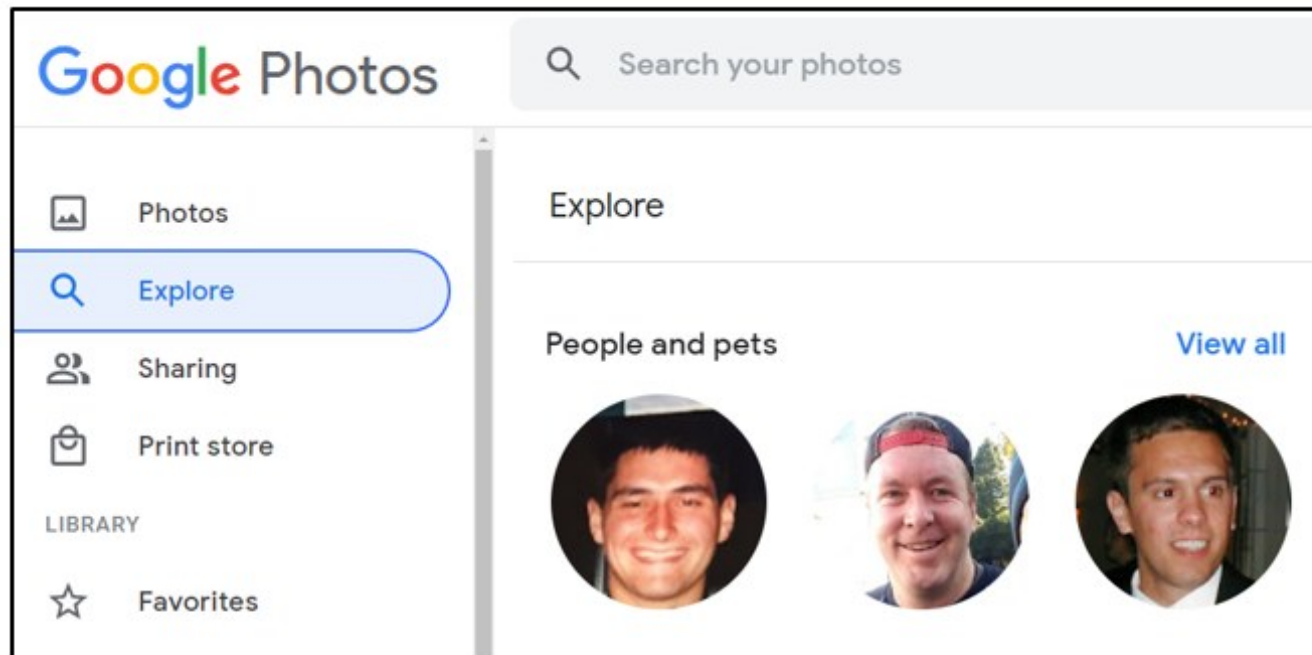
SPACEDAILY

Voyager 1's Data Transmission Issue Traced to Memory Corruption, Fix in Progress

1 hour ago



Semi-supervised Learning

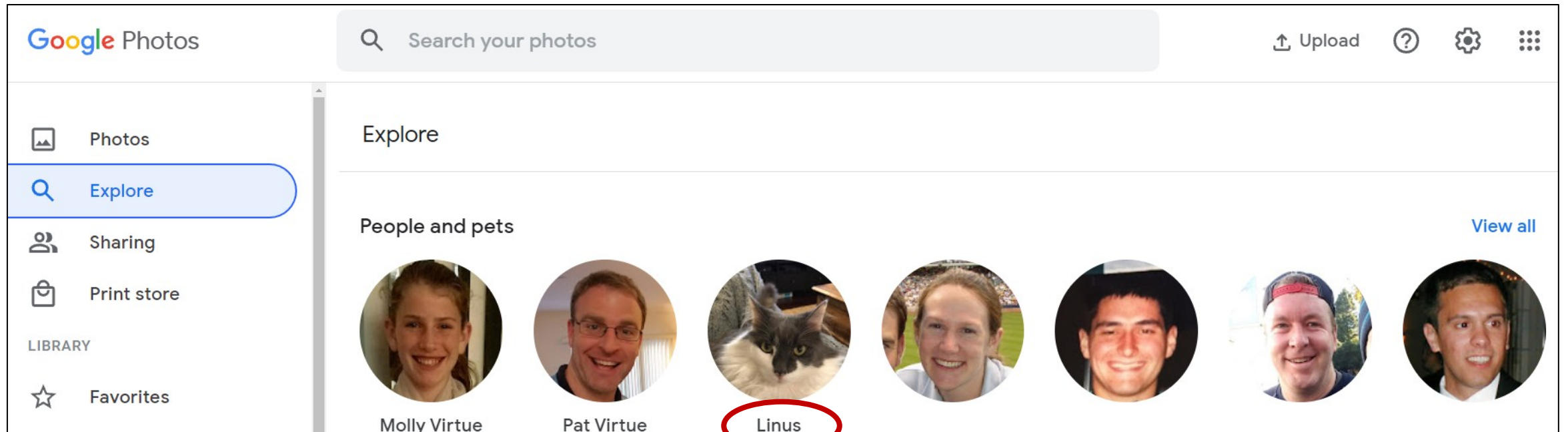


Semi-supervised Learning

Supervised: $\mathcal{D} = \{\mathbf{x}^{(i)}, y^{(i)}\}_{i=1}^N$

Unsupervised: $\mathcal{D} = \{\mathbf{x}^{(i)}\}_{i=1}^N$

Semi-supervised: $\mathcal{D} = \{\mathbf{x}^{(i)}\}_{i=1}^{N_1} \cup \{\mathbf{x}^{(j)}, y^{(j)}\}_{j=1}^{N_2}$

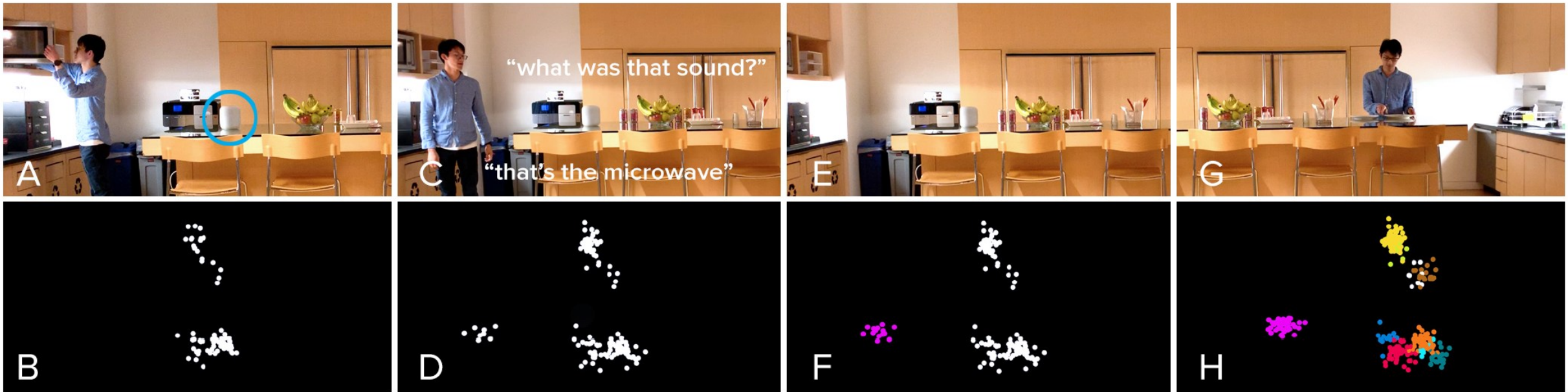


Semi-supervised Learning

Supervised: $\mathcal{D} = \{\mathbf{x}^{(i)}, y^{(i)}\}_{i=1}^N$

Unsupervised: $\mathcal{D} = \{\mathbf{x}^{(i)}\}_{i=1}^N$

Semi-supervised: $\mathcal{D} = \{\mathbf{x}^{(i)}\}_{i=1}^{N_1} \cup \{\mathbf{x}^{(j)}, y^{(j)}\}_{j=1}^{N_2}$



Clustering

k-means algorithm

K-Means

Algorithm

Input: Data, $x^{(i)}$, Desired number of clusters, K

Initialize: the K cluster centers (randomly if necessary)

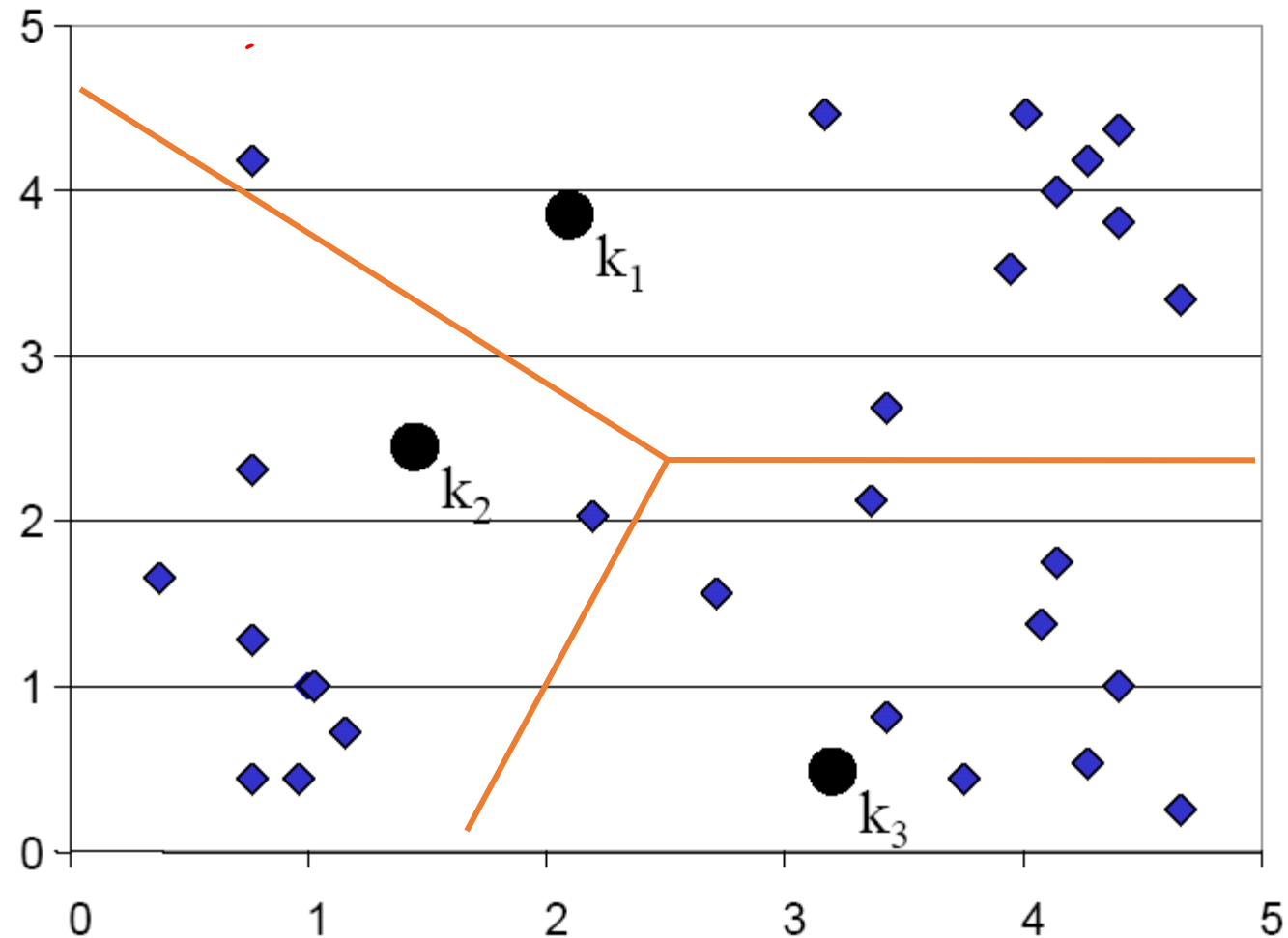
Iterate:

1. Assign points to the nearest cluster centers
2. Re-estimate the K cluster centers (aka the **centroid** or **mean**), by assuming the memberships found above are correct.

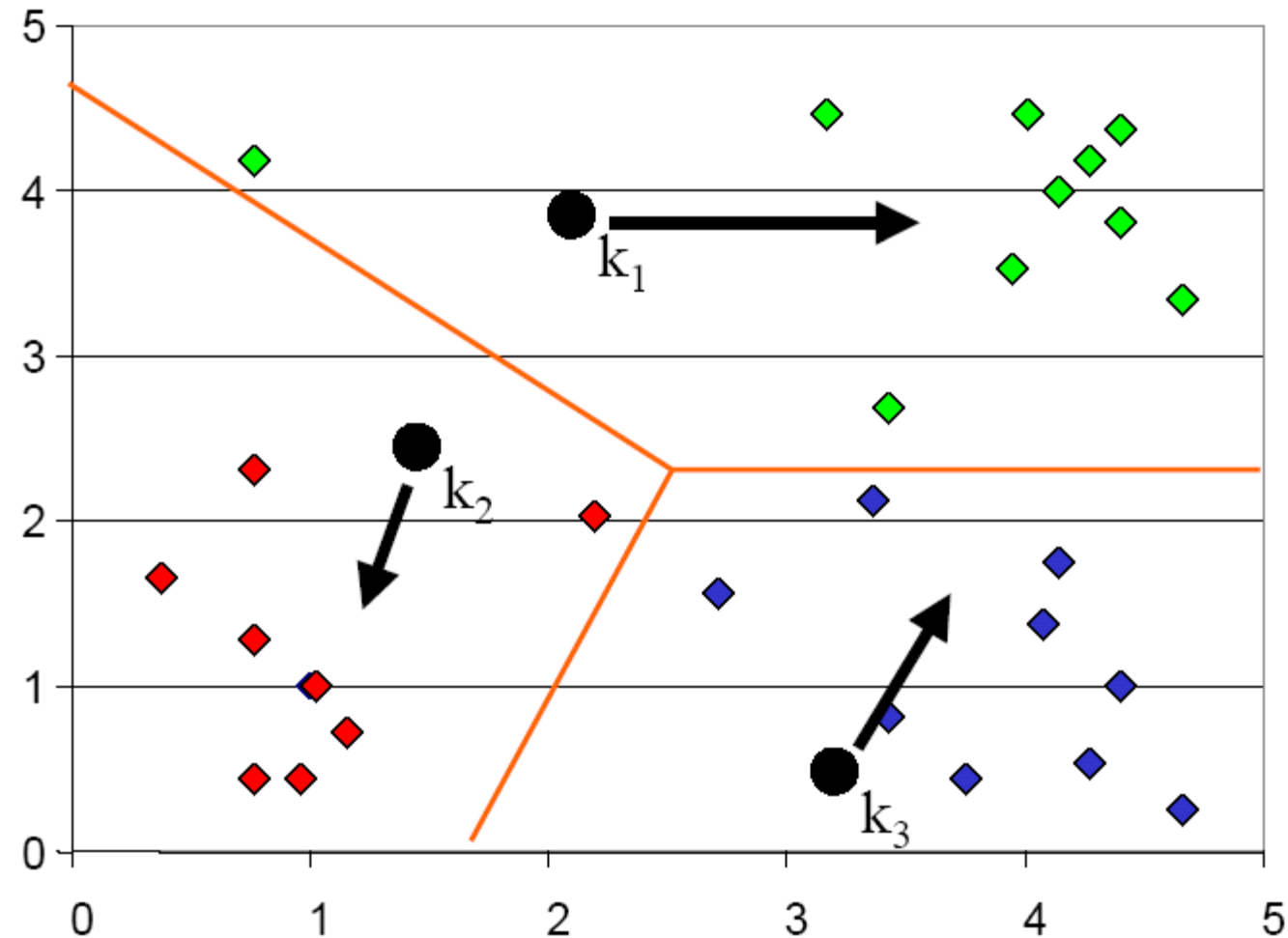
Termination:

If none of the objects changed membership in the last iteration, exit. Otherwise go to 1.

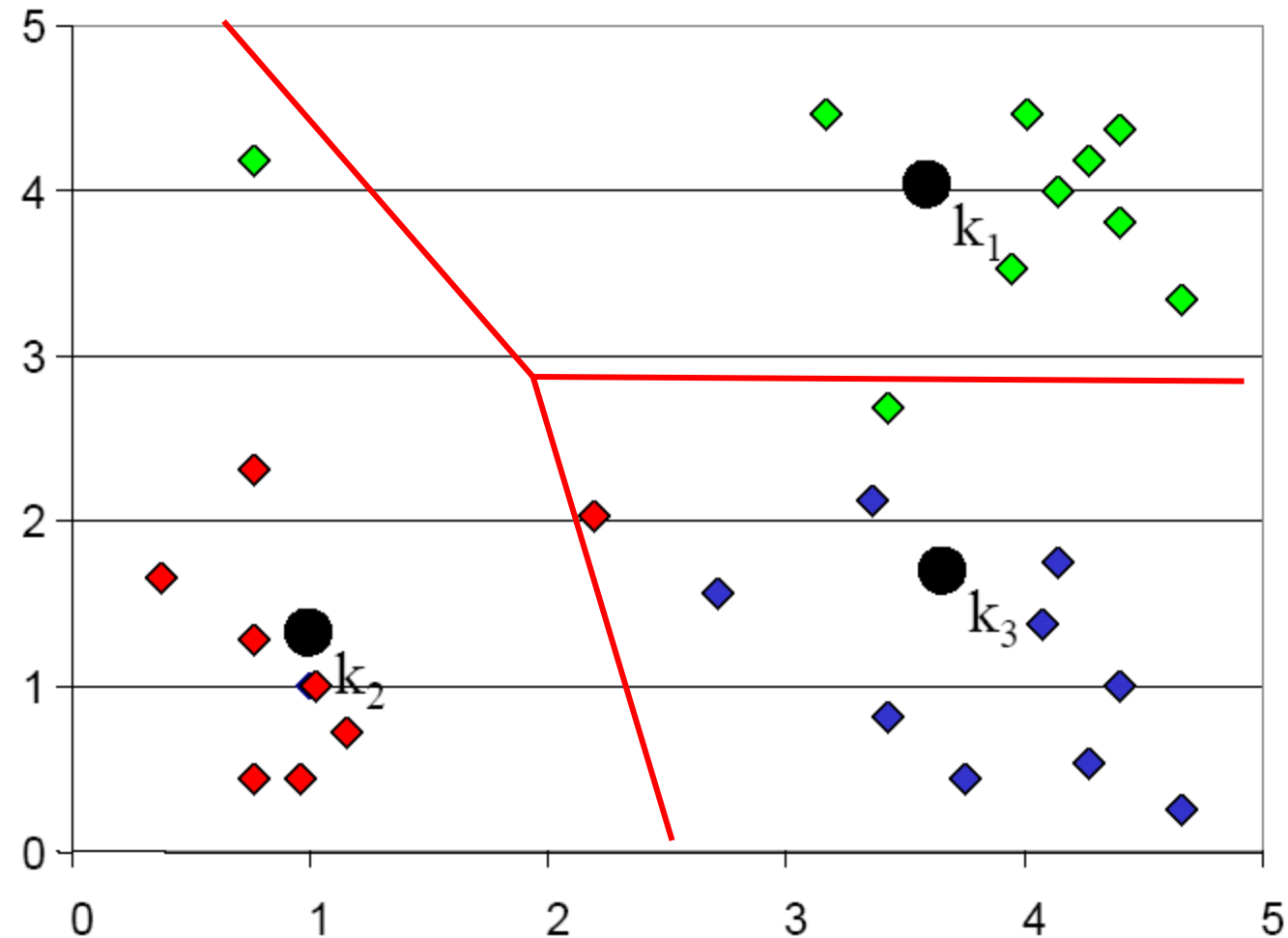
K-means Clustering: Assign points



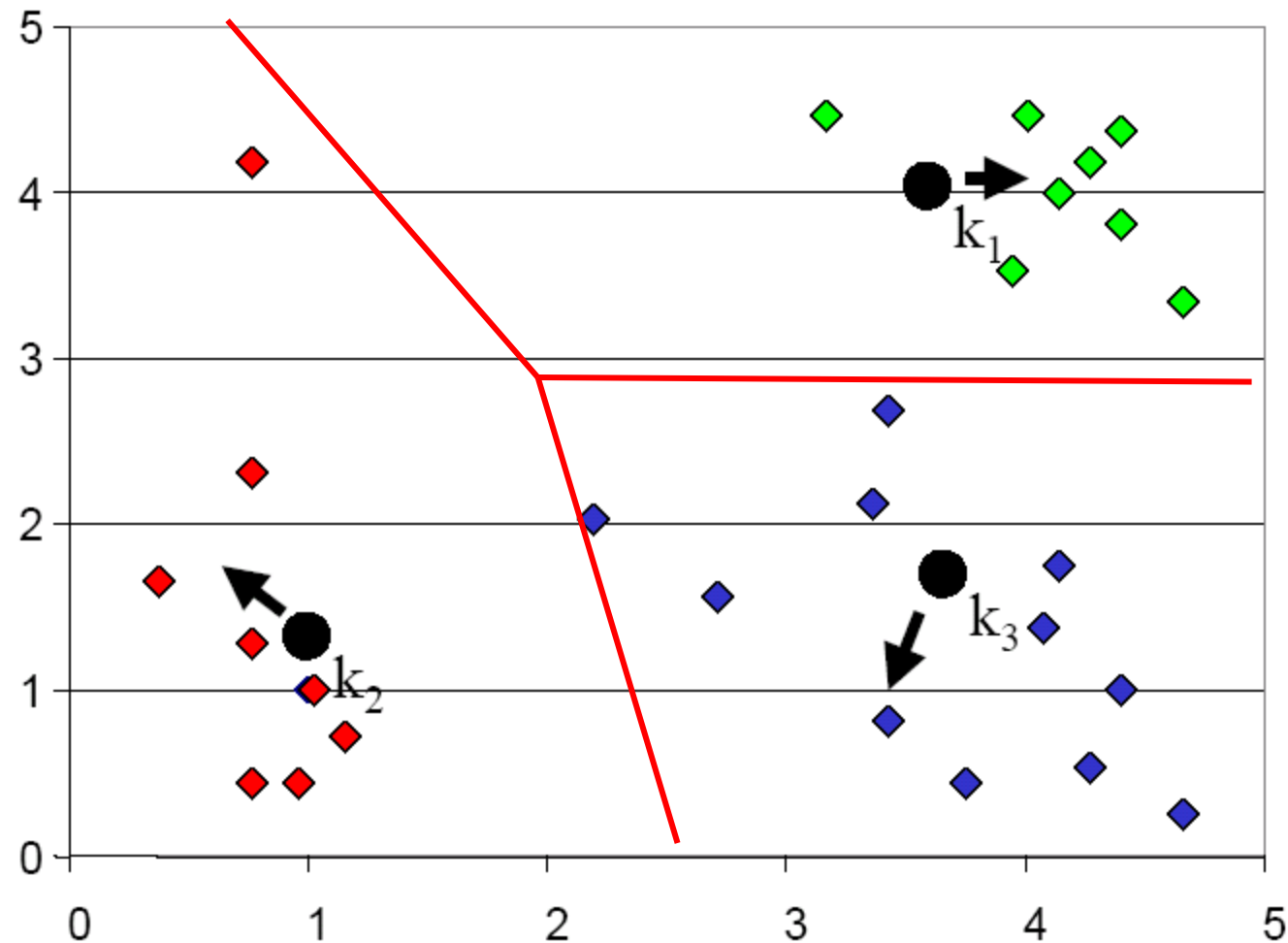
K-means Clustering: Update centers



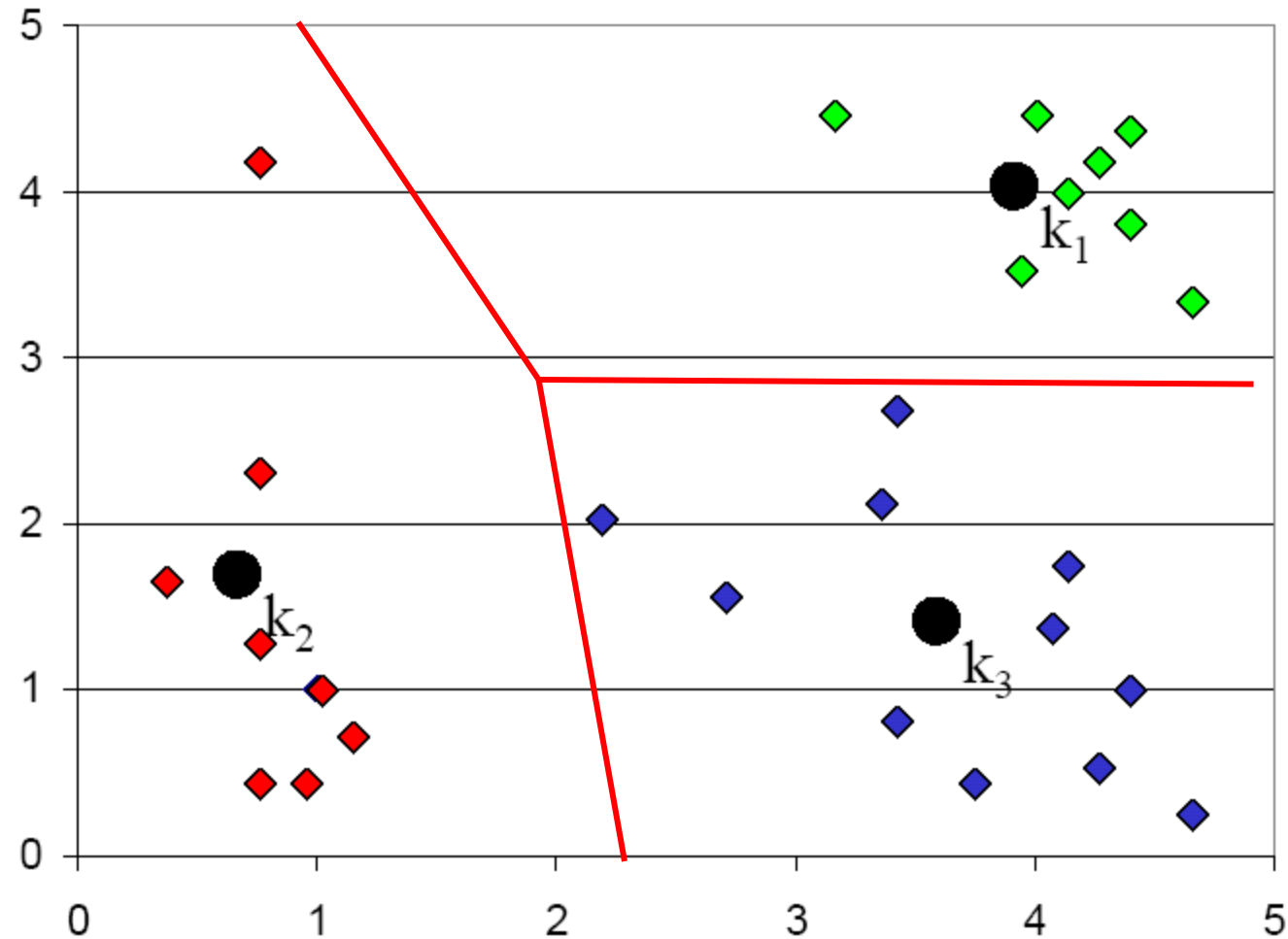
K-means Clustering: Assign points



K-means Clustering: Update centers



K-means Clustering: Assign points



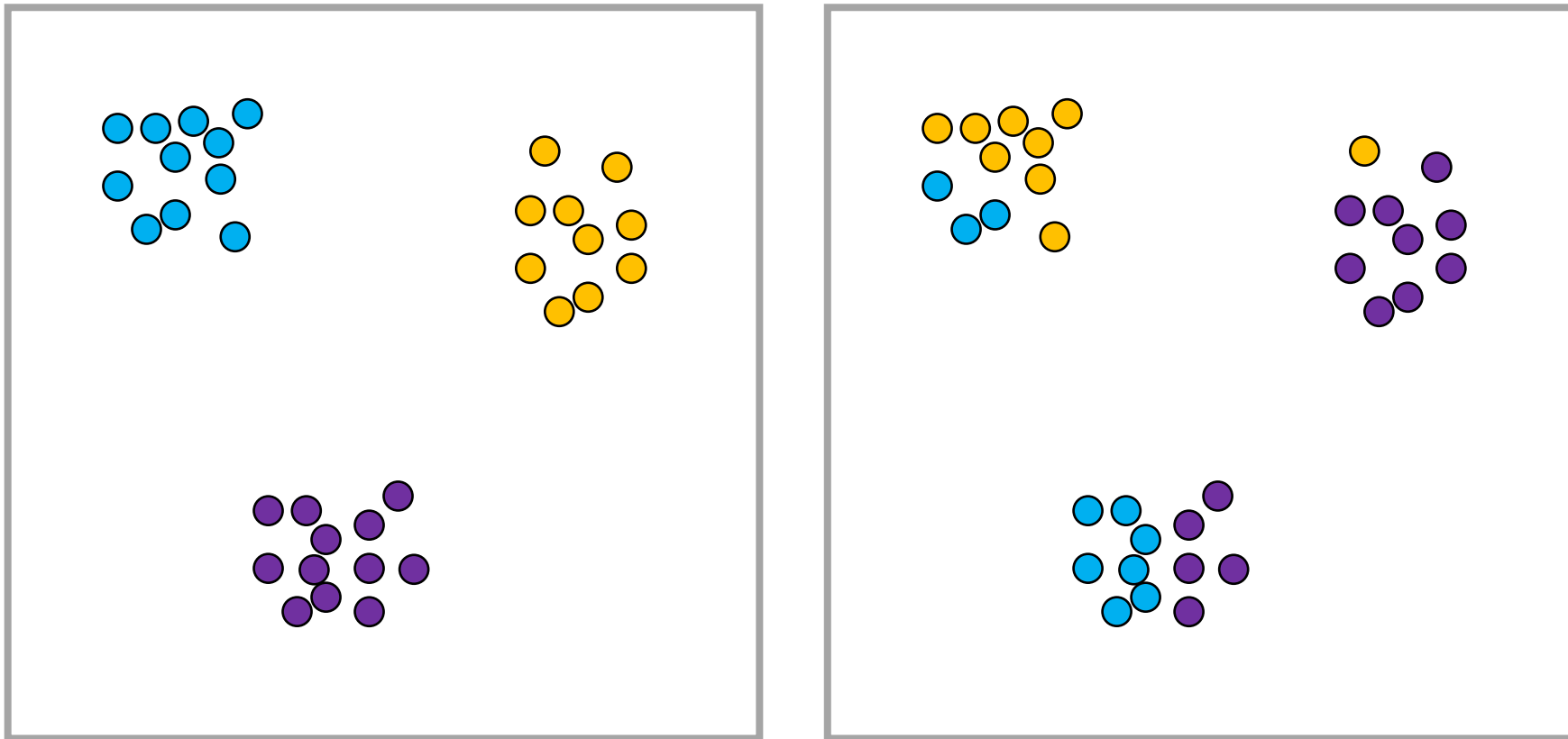
K-means Optimization

Optimization recipe

1. Formulate objective
2. Minimize objective

K-means Optimization

Question: Which of these clusterings is “better”?



K-means Optimization

Input: $K, \mathbf{x}^{(1)}, \dots, \mathbf{x}^{(N)}, \mathbf{x}^{(i)} \in \mathbb{R}^M$

Num clusters, unlabeled data

Output: $z^{(1)}, \dots, z^{(N)}, z^{(i)} \in \{1 \dots K\}$

Cluster assignments per point

Output: $\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_K, \boldsymbol{\mu}_k \in \mathbb{R}^M$

Cluster centers

K-means Optimization

Computational complexity

$$\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_K, \mathbf{z} = \underset{\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_K, \mathbf{z}}{\operatorname{argmin}} \sum_{i=1}^N \|\mathbf{x}^{(i)} - \boldsymbol{\mu}_{\mathbf{z}^{(i)}}\|_2^2$$

K-means Optimization

Alternating minimization

$$\text{a) } \mathbf{z} = \underset{\mathbf{z}}{\operatorname{argmin}} \sum_{i=1}^N \left\| \mathbf{x}^{(i)} - \boldsymbol{\mu}_{\mathbf{z}(i)} \right\|_2^2$$

$$\text{b) } \boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_K = \underset{\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_K}{\operatorname{argmin}} \sum_{i=1}^N \left\| \mathbf{x}^{(i)} - \boldsymbol{\mu}_{\mathbf{z}(i)} \right\|_2^2$$

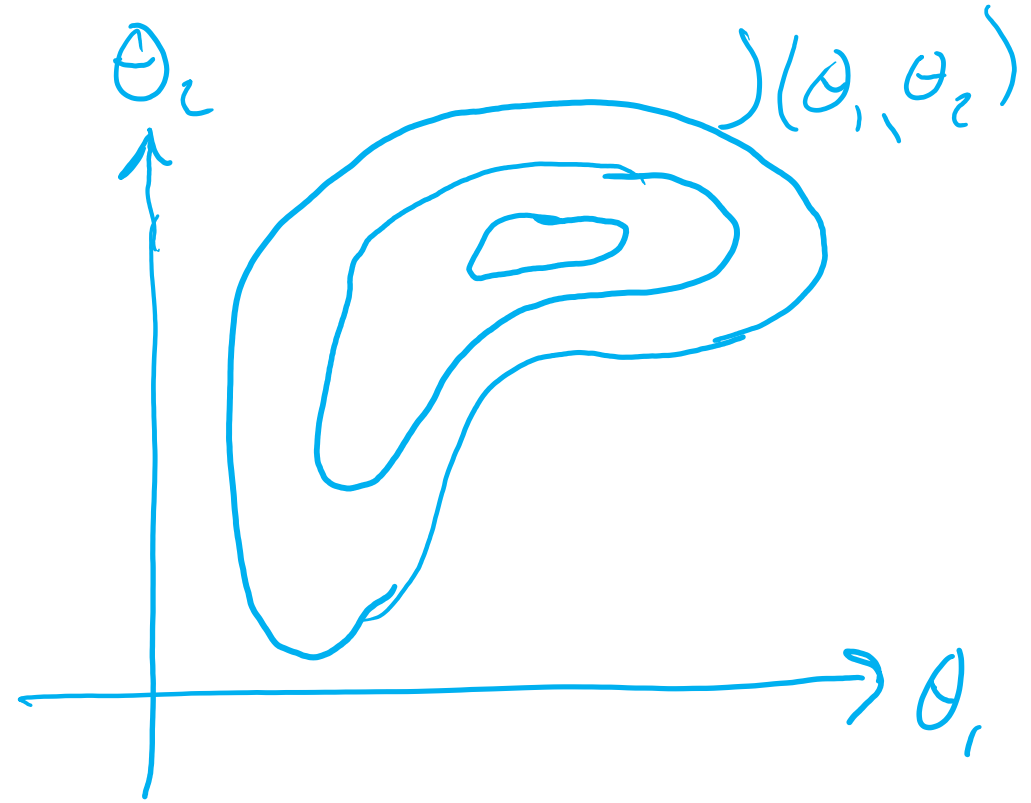
Alternating minimization

Coordinate descent

$$\min_{\theta_1, \theta_2} J(\theta_1, \theta_2)$$

Find minimum for one parameter

- $\theta_1 \leftarrow \underset{\theta_1}{\operatorname{argmin}} J(\theta_1, \theta_2)$
- $\theta_2 \leftarrow \underset{\theta_2}{\operatorname{argmin}} J(\theta_1, \theta_2)$



Alternating minimization

Block coordinate descent

$$\min_{\alpha, \beta} J(\alpha, \beta)$$

Find minimum for one set of parameter

a. $\alpha \leftarrow \operatorname{argmin}_{\alpha} J(\alpha, \beta)$

b. $\beta \leftarrow \operatorname{argmin}_{\beta} J(\alpha, \beta)$

Gaussian Mixture Models

Poll 1

Probability warm-up: Omega Pizzeria

Two pizzas with different distributions of toppings

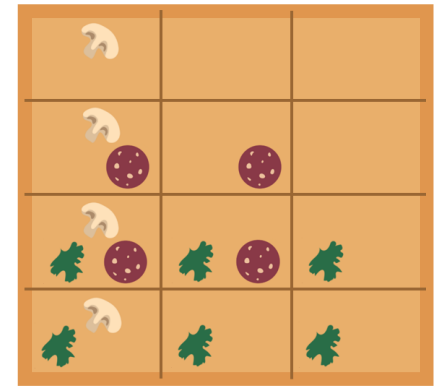
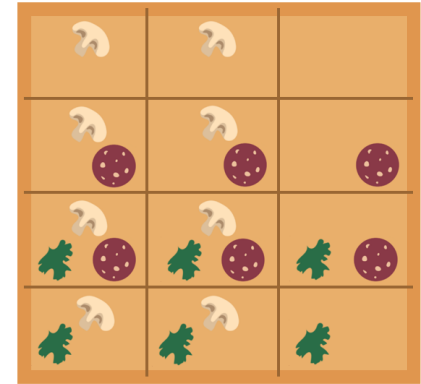
Let $Z=1$ represent the event of picking pizza 1 (rather than pizza 2)

Let $X=1$ represent the event of picking a slice with mushrooms on it

Given $P(Z = 1) = \frac{1}{4}$, what is the probability that I end up with a slice with mushrooms $P(X = 1)$?

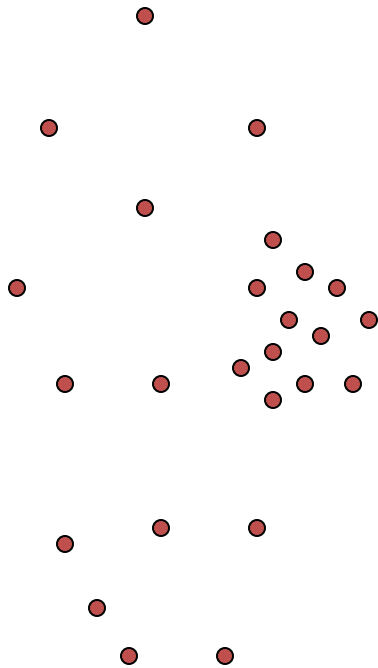
Give your answer as a (simplified) fraction.

$$P(X = 1 \mid Z = 1) = \frac{2}{3}$$



$$P(X = 1 \mid Z = 0) = \frac{1}{3}$$

(One) bad case for K-means



- Clusters may overlap
- Some clusters may be “wider” than others
- Clusters may not be linearly separable

Partitioning Algorithms

- K-means
 - **hard assignment**: each object belongs to only one cluster
- Mixture modeling
 - **soft assignment**: probability that an object belongs to a cluster

Generative approach

Gaussian Mixture Model

Mixture of K Gaussian distributions (multi-modal distribution)

Probability of point belonging to cluster k :

$$p(z_k = 1)$$

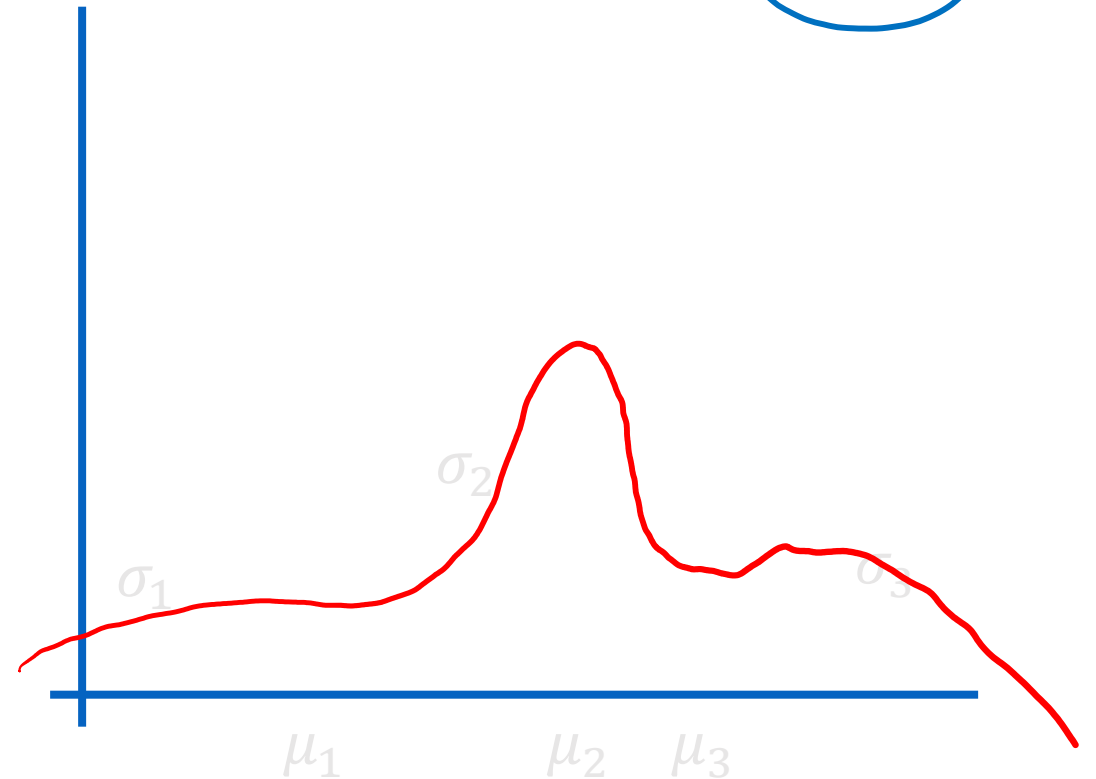
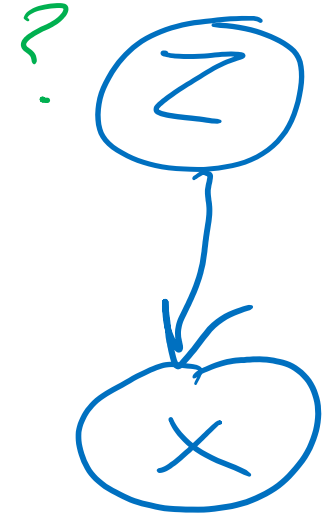
Density of point assuming cluster k :

$$p(x | z_k = 1) \sim \mathcal{N}(\mu_k, \sigma_k)$$

$$p(x) = \sum_k p(x | z_k = 1) p(z_k = 1)$$

Mixture
component

Mixture
proportion



Gaussian Mixture Model

Mixture of K Gaussian distributions (multi-modal distribution)

Probability of point belonging to cluster k :

$$p(z_k = 1)$$

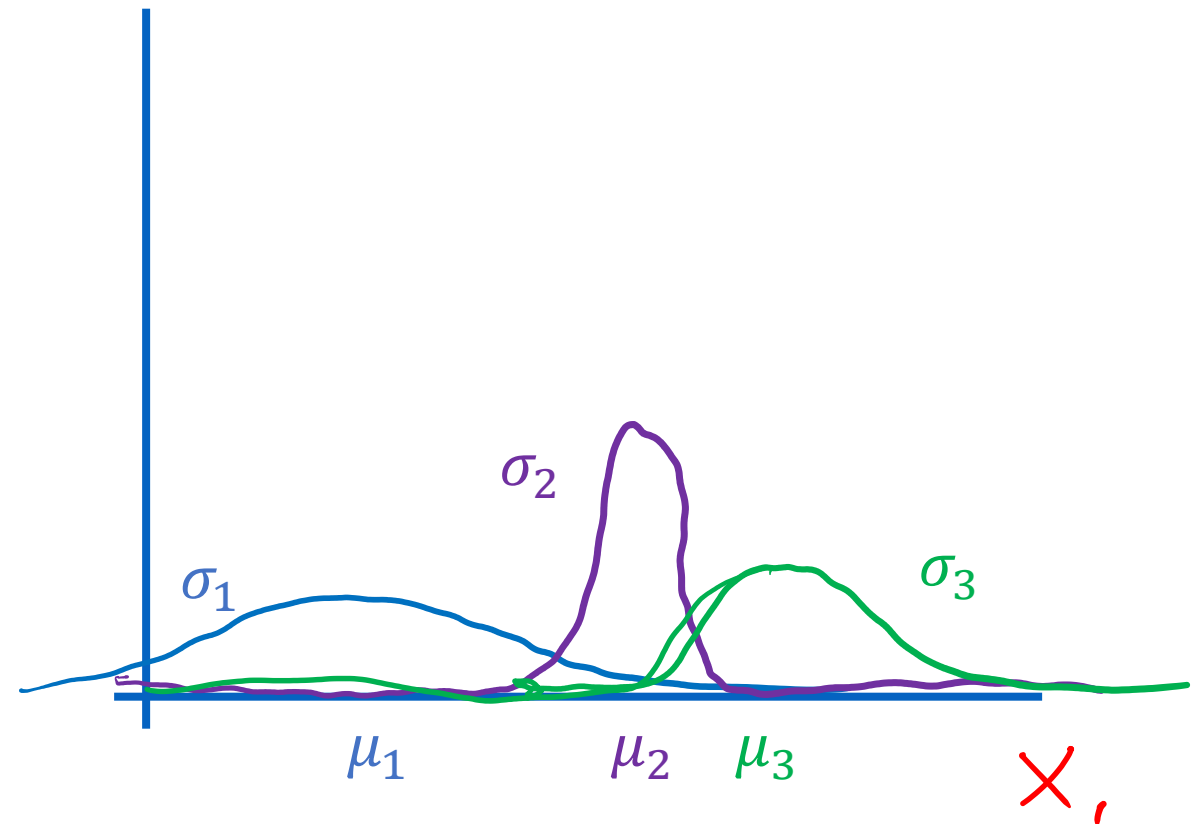
Density of point assuming cluster k :

$$p(x | z_k = 1) \sim \mathcal{N}(\mu_k, \sigma_k)$$

$$p(x) = \sum_k p(x | z_k = 1) p(z_k = 1)$$

Mixture
component

Mixture
proportion



Gaussian Mixture Model

Mixture of K Gaussian distributions (multi-modal distribution)

Probability of point belonging to cluster k :

$$p(z_k = 1)$$

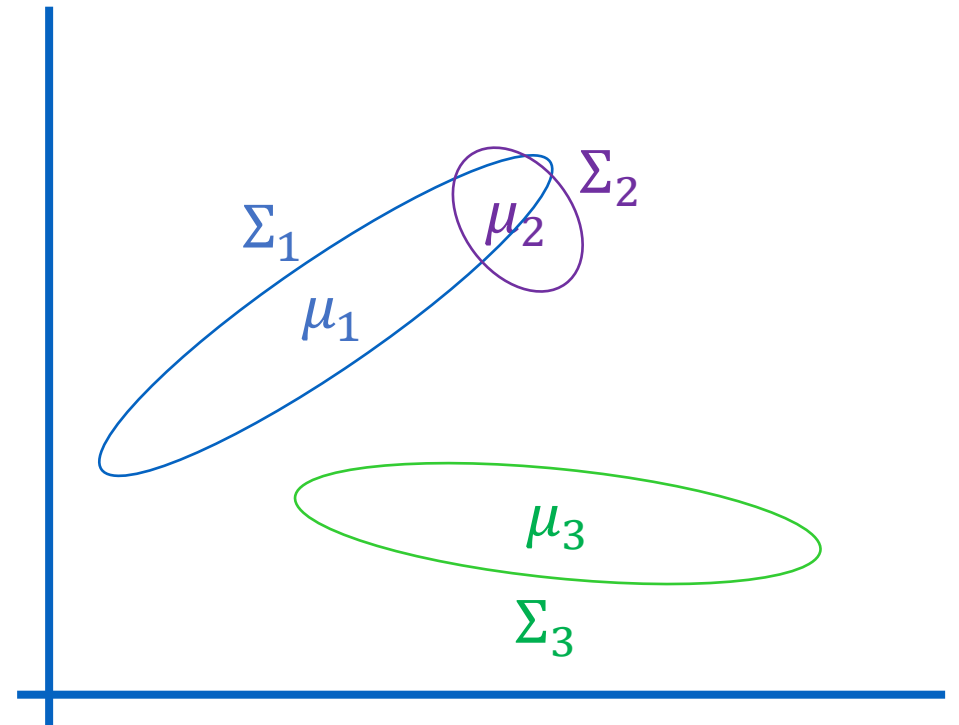
Density of point assuming cluster k :

$$p(\mathbf{x} \mid z_k = 1) \sim \mathcal{N}(\boldsymbol{\mu}_k, \Sigma_k)$$

$$p(\mathbf{x}) = \sum_k p(\mathbf{x} \mid z_k = 1) p(z_k = 1)$$

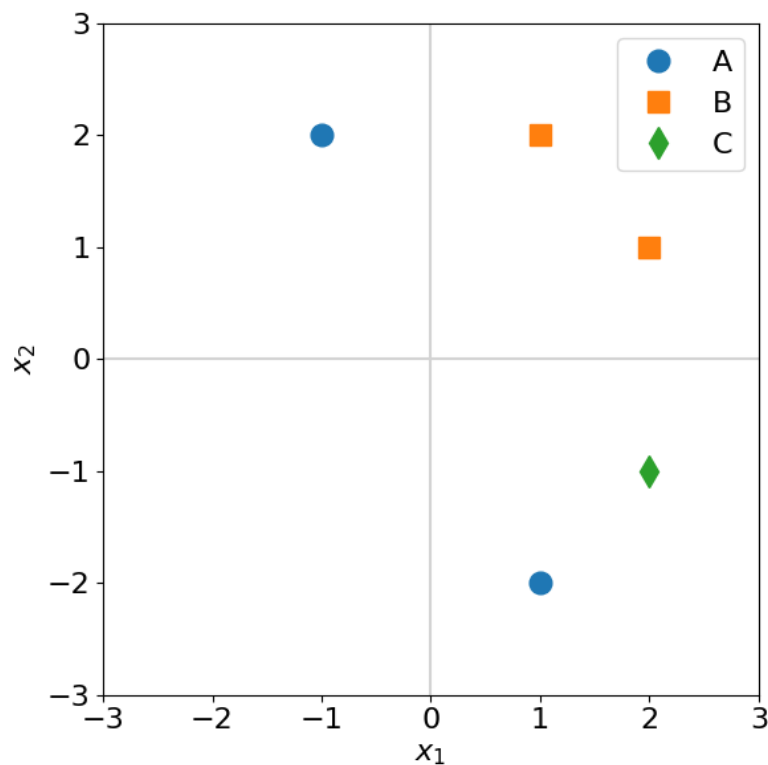
Mixture
component

Mixture
proportion



Generative Models: Supervised vs Unsupervised

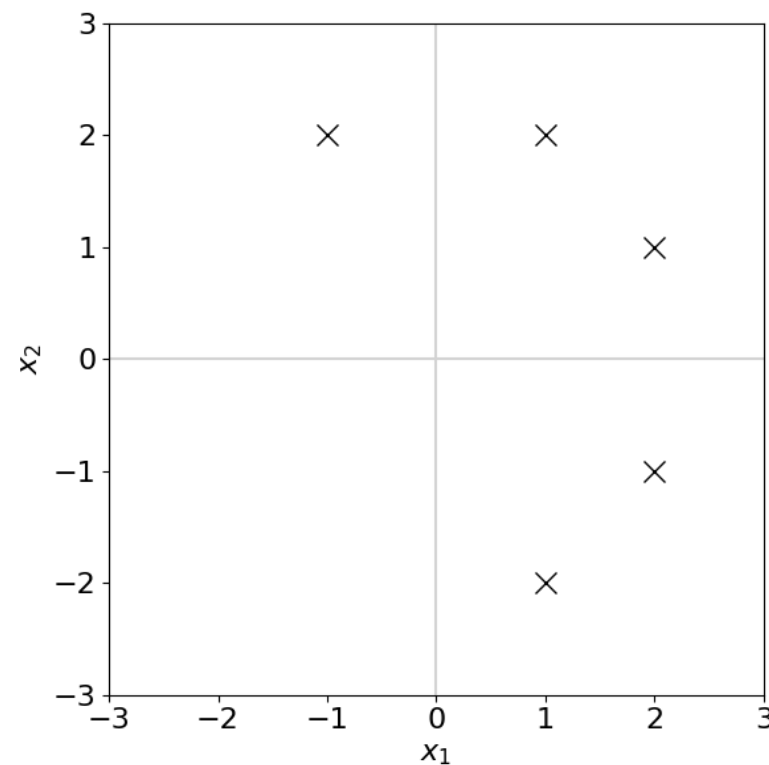
Gaussian discriminant analysis vs Gaussian mixture models



$$Y \sim \text{Categorical}(\pi_1, \pi_2, \pi_3)$$

$$X_{Y=k} \sim \mathcal{N}(\mu_k, \sigma_k^2)$$

$$\mathcal{D} = \{x^{(i)}, y^{(i)}\}_{i=1}^N$$



$$Z \sim \text{Categorical}(\pi_1, \pi_2, \pi_3)$$

$$X_{Z=k} \sim \mathcal{N}(\mu_k, \sigma_k^2)$$

$$\mathcal{D} = \{x^{(i)}\}_{i=1}^N$$

Soft K-means

Working our way up to GMMs

Soft K-means

Assume we have means, how could we estimate the probability a point belongs to each cluster?

Softmax of inverse squared distance

$$\begin{bmatrix} P(z_1 = 1 | \mathbf{x}) \\ P(z_2 = 1 | \mathbf{x}) \\ P(z_3 = 1 | \mathbf{x}) \end{bmatrix} = \begin{bmatrix} 0.307 \\ 0.379 \\ 0.314 \end{bmatrix} = \begin{bmatrix} e^{-\|x - \mu_1\|_2^2} \\ e^{-\|x - \mu_2\|_2^2} \\ e^{-\|x - \mu_3\|_2^2} \end{bmatrix} \frac{1}{\sum_{k=1}^3 e^{-\|x - \mu_k\|_2^2}}$$

Inverse distance (squared)

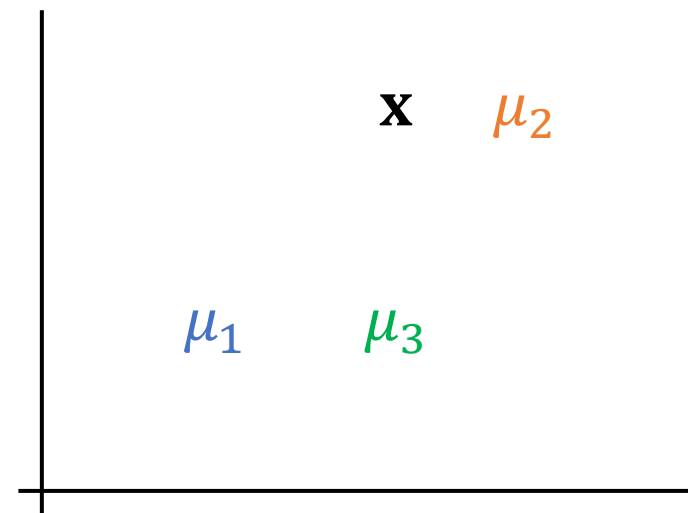
$$1/\|x - \mu_k\|_2^2$$

$$\begin{bmatrix} 1/25 \\ 1/4 \\ 1/16 \end{bmatrix}$$

Distance (squared)

$$\|x - \mu_k\|_2^2$$

$$\begin{bmatrix} 5^2 \\ 2^2 \\ 4^2 \end{bmatrix}$$



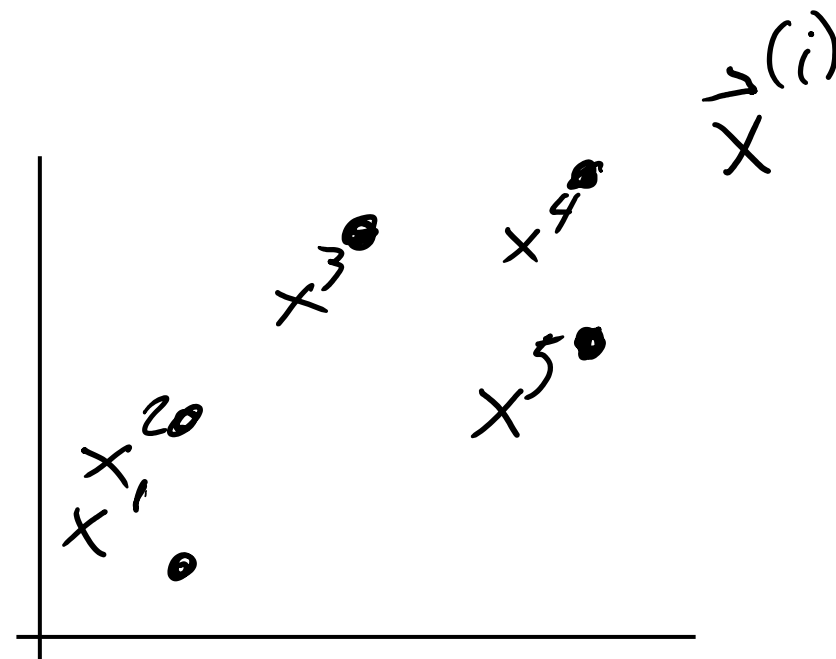
Soft K-means

Assume we have cluster probabilities for all points, how can we update the mean for cluster k ?

K-means

$$\mu_3 = \frac{1}{N_3} \sum x^{(i)} \mathbb{I}(z=3)$$

$$\mu_3 =$$



Soft K-means

Assume we have cluster probabilities for all points, how can we update the mean for cluster k ?

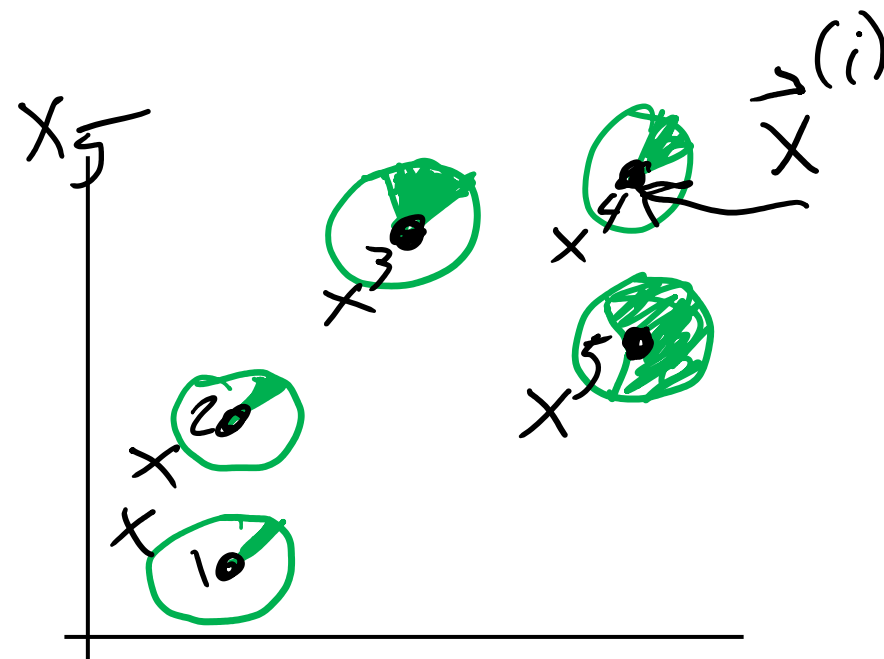
$$\mu_3 = \frac{x^{(3)} + x^{(5)}}{1 + 1}$$

Hard K-means

$$\mu_3 =$$

K-means

$$\mu_3 = \frac{1}{N_3} \sum x^{(i)} \mathbb{I}(z=3)$$



Soft K-means

Assume we have cluster probabilities for all points, how can we update the mean for cluster k ?

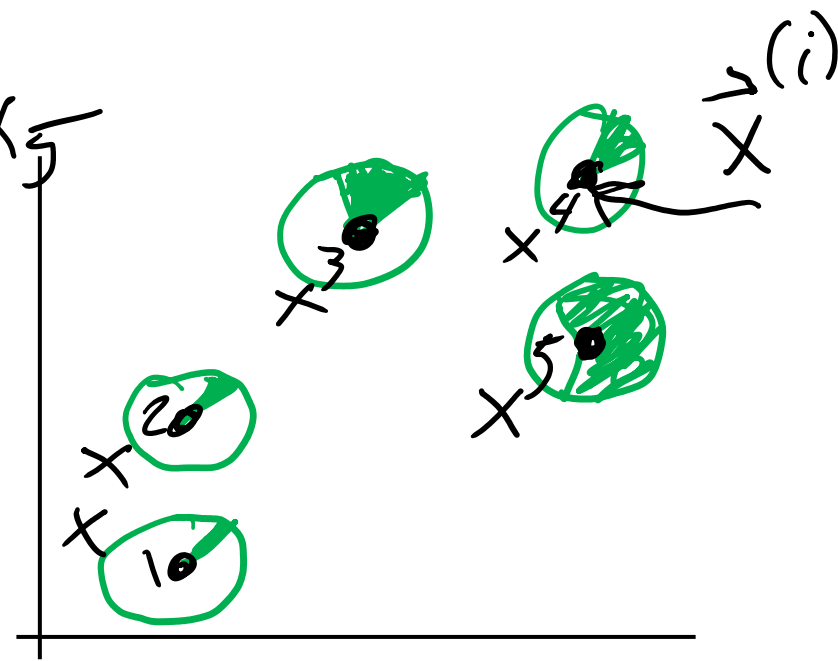
K -means
$$\mu_3 = \frac{1}{N_3} \sum x^{(i)} \mathbb{I}(z=3)$$

$$\mu_3 = \frac{x_3 + x_5}{1 + 1}$$

Hard K-means

μ_3 ?

$$\mu_3 = \frac{x_1 + x_2 + x_3 + x_4 + x_5}{1 + 1 + 1 + 1 + 1}$$



Soft K-means

0) Initialize clusters

Repeat until convergence

1) Update cluster probabilities

2) Update cluster means

Back to (full) GMMs

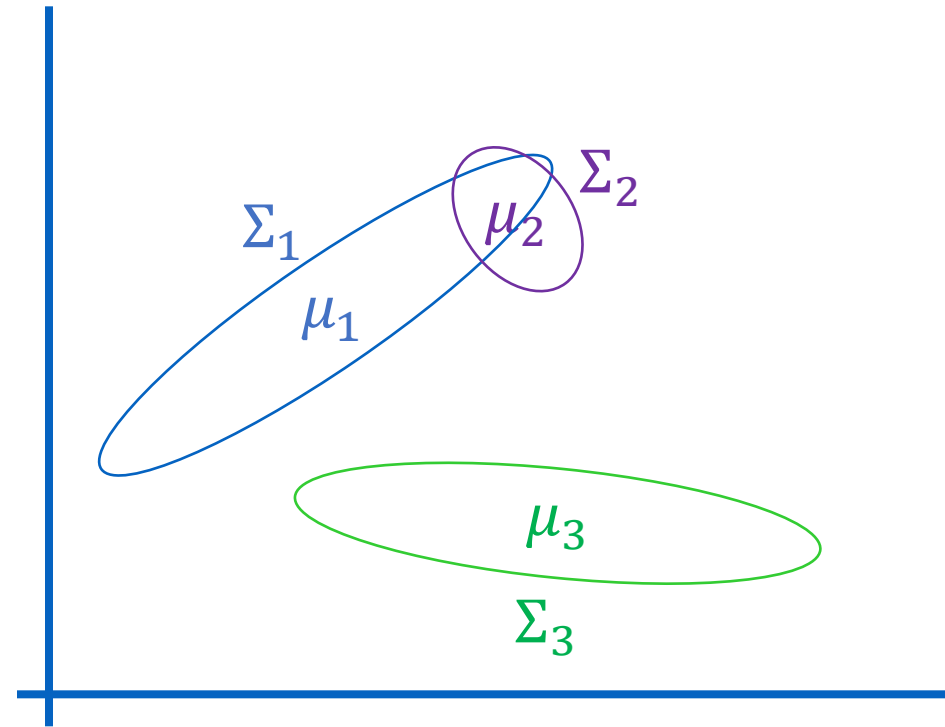
Gaussian Mixture Model

Mixture of K Gaussian distributions (multi-modal distribution)

- There are K components
- Component k generates data from a Gaussian with mean vector μ_k and covariance matrix Σ_k

Each data point is generated according to the follow recipe:

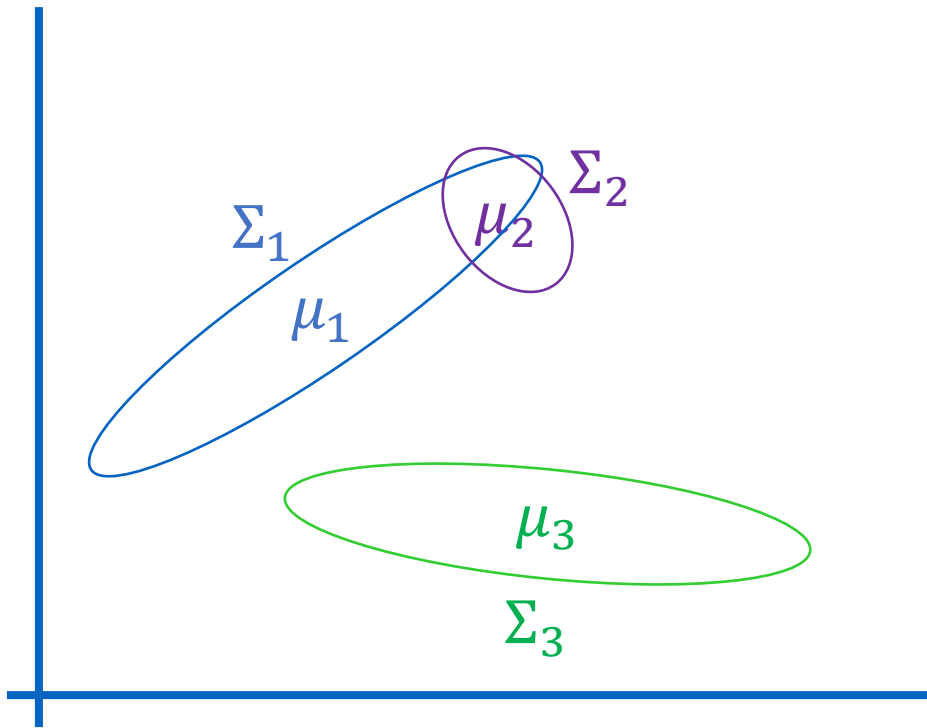
- 1) Pick a component at random:
Choose component k with probability $p(z_k = 1)$
- 2) Data point $\mathbf{x} \sim \mathcal{N}(\mu_k, \Sigma_k)$



Learning General GMM

Mixture of K Gaussian distributions (multi-modal distribution)

$$x_1, \dots, x_M \sim p(\mathbf{x}) = \sum_{k=1}^K p(\mathbf{x} \mid z_k = 1) p(z_k = 1)$$



Mixture: $\pi_k \stackrel{\text{def}}{=} p(z_k = 1)$

Gaussian components:

$$p(\mathbf{x} \mid z_k = 1) \sim \mathcal{N}(\boldsymbol{\mu}_k, \Sigma_k)$$

Parameters:

$$\theta \stackrel{\text{def}}{=} \{\pi_k, \boldsymbol{\mu}_k, \Sigma_k\}_{k=1}^K$$

How to estimate parameters? Can we do MLE even without labels $\mathbf{z}$?

Learning General GMM

Maximize marginal likelihood:

$$\begin{aligned} & \operatorname{argmax}_{\theta} \prod_i^N p(\mathbf{x}^{(i)} \mid \theta) \\ &= \operatorname{argmax}_{\theta} \prod_i^N \sum_{k=1}^K p(\mathbf{x}^{(i)}, z_k^{(i)} = 1 \mid \theta) \\ &= \operatorname{argmax}_{\theta} \prod_i^N \sum_{k=1}^K p(z_k^{(i)} = 1) p(\mathbf{x}^{(i)} \mid z_k^{(i)} = 1) \\ &= \operatorname{argmax}_{\theta} \prod_i^N \sum_{k=1}^K \pi_k |\Sigma_k|^{-\frac{1}{2}} e^{-\frac{1}{2}(\mathbf{x}^{(i)} - \boldsymbol{\mu}_k)^T \Sigma_k^{-1} (\mathbf{x}^{(i)} - \boldsymbol{\mu}_k)} \end{aligned}$$

Learning General GMM

Maximize marginal likelihood:

$$\begin{aligned} & \operatorname{argmax}_{\theta} \prod_i^N p(\mathbf{x}^{(i)} \mid \theta) \\ &= \operatorname{argmax}_{\theta} \prod_i^N \sum_{k=1}^K \pi_k |\Sigma_k|^{-\frac{1}{2}} e^{-\frac{1}{2}(\mathbf{x}^{(i)} - \mu_k)^T \Sigma_k^{-1} (\mathbf{x}^{(i)} - \mu_k)} \end{aligned}$$

How do we find the π_k, μ_k, Σ_k which give the max. marginal likelihood?

- a) Set $\frac{\partial}{\partial \mu_k} \ell(\theta; \mathcal{D}) = 0$ and solve for μ_k , etc. ? **No closed-form solution**
- b) Use gradient descent? **Doable, but complicated, often slow, and need to consider constraints on parameters**

Log (Marginal) Likelihood for Missing Data

Marginalize over missing data, $\mathbf{z}^{(i)}$

$$\ell(\theta \mid \mathcal{D}) = \log \prod_i^N p(\mathbf{x}^{(i)} \mid \theta)$$

Log (Marginal) Likelihood for Missing Data

Marginalize over missing data, $\mathbf{z}^{(i)}$

$$\begin{aligned}\ell(\theta \mid \mathcal{D}) &= \log \prod_i^N p(\mathbf{x}^{(i)} \mid \theta) \\ &= \sum_{i=1}^N \log p(x^i \mid \theta) \\ &= \sum_i \log \sum_{k=1}^K p(x^i, z^i \mid \theta) \\ &= \sum_{i=1}^N \log \sum_{k=1}^K \pi_k f_{\mathcal{N}}(x^i \mid \mu_k, \Sigma_k)\end{aligned}$$


GMM vs K-means

Maximize marginal likelihood:

$$\begin{aligned} & \operatorname{argmax}_{\theta} \prod_{i=1}^N p(\mathbf{x}^{(i)} \mid \theta) \\ &= \operatorname{argmax}_{\theta} \prod_{i=1}^N \sum_{k=1}^K p(z^{(i)} = k) p(\mathbf{x}^{(i)} \mid z^{(i)} = k) \end{aligned}$$

What happens if we assume a **hard-assignment**?

$p(z^{(i)} = k) = 1$ if point i belongs to the k -th cluster

(and assume variances are all the same)

$$\begin{aligned} & \operatorname{argmax}_{\theta} \prod_{i=1}^N \sum_{k=1}^K p(z^{(i)} = k) p(\mathbf{x}^{(i)} \mid z^{(i)} = k) \\ &= \operatorname{argmax}_{\theta} \prod_{i=1}^N e^{-\frac{1}{2} \|\mathbf{x}^{(i)} - \mu_{z^{(i)}}\|_2^2} \\ &= \operatorname{argmin}_{\theta} \sum_{i=1}^N \|\mathbf{x}^{(i)} - \mu_{z^{(i)}}\|_2^2 \end{aligned}$$

Same as K-means!

K-means Optimization

Alternating minimization

$$\text{a) } \mathbf{z} = \underset{\mathbf{z}}{\operatorname{argmin}} \sum_{i=1}^N \left\| \mathbf{x}^{(i)} - \boldsymbol{\mu}_{\mathbf{z}^{(i)}} \right\|_2^2$$

$$\text{b) } \boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_K = \underset{\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_K}{\operatorname{argmin}} \sum_{i=1}^N \left\| \mathbf{x}^{(i)} - \boldsymbol{\mu}_{\mathbf{z}^{(i)}} \right\|_2^2$$

We're going to use alternating minimization to solve GMMs as well!

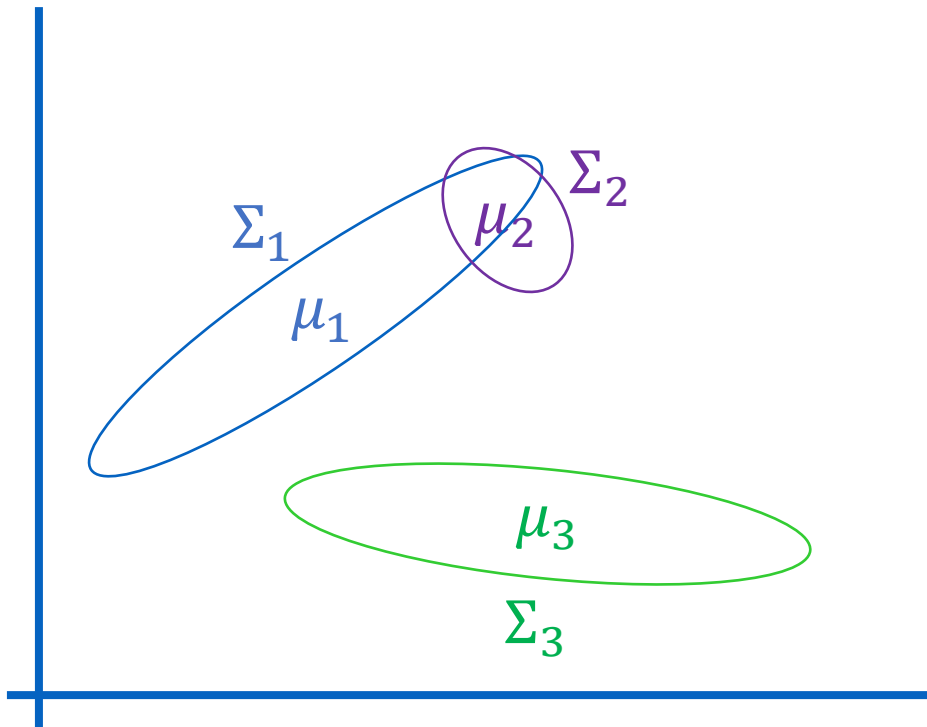
EM Algorithm for GMMs

We'll get to what EM means later

Learning General GMM

Mixture of K Gaussian distributions (multi-modal distribution)

$$x_1, \dots, x_M \sim p(\mathbf{x}) = \sum_{k=1}^K p(\mathbf{x} \mid Z_k = 1) p(Z_k = 1)$$



Mixture: $\pi_k \stackrel{\text{def}}{=} p(Z_k = 1)$

Gaussian components:

$$p(\mathbf{x} \mid Z_k = 1) \sim \mathcal{N}(\boldsymbol{\mu}_k, \Sigma_k)$$

Parameters:

$$\boldsymbol{\theta} \stackrel{\text{def}}{=} \{\pi_k, \boldsymbol{\mu}_k, \Sigma_k\}_{k=1}^K$$

How to estimate parameters? Can we do MLE even without labels $\mathbf{z}$?

EM for GMMS

Initialize parameters

For $t = 0$, $\pi_k^{(0)}$, $\boldsymbol{\mu}_k^{(0)}$, $\Sigma_k^{(0)}$

$$P(Z_k^{(i)} = 1) \quad Z = \begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_k \end{bmatrix}$$

↑
categorical

E-step

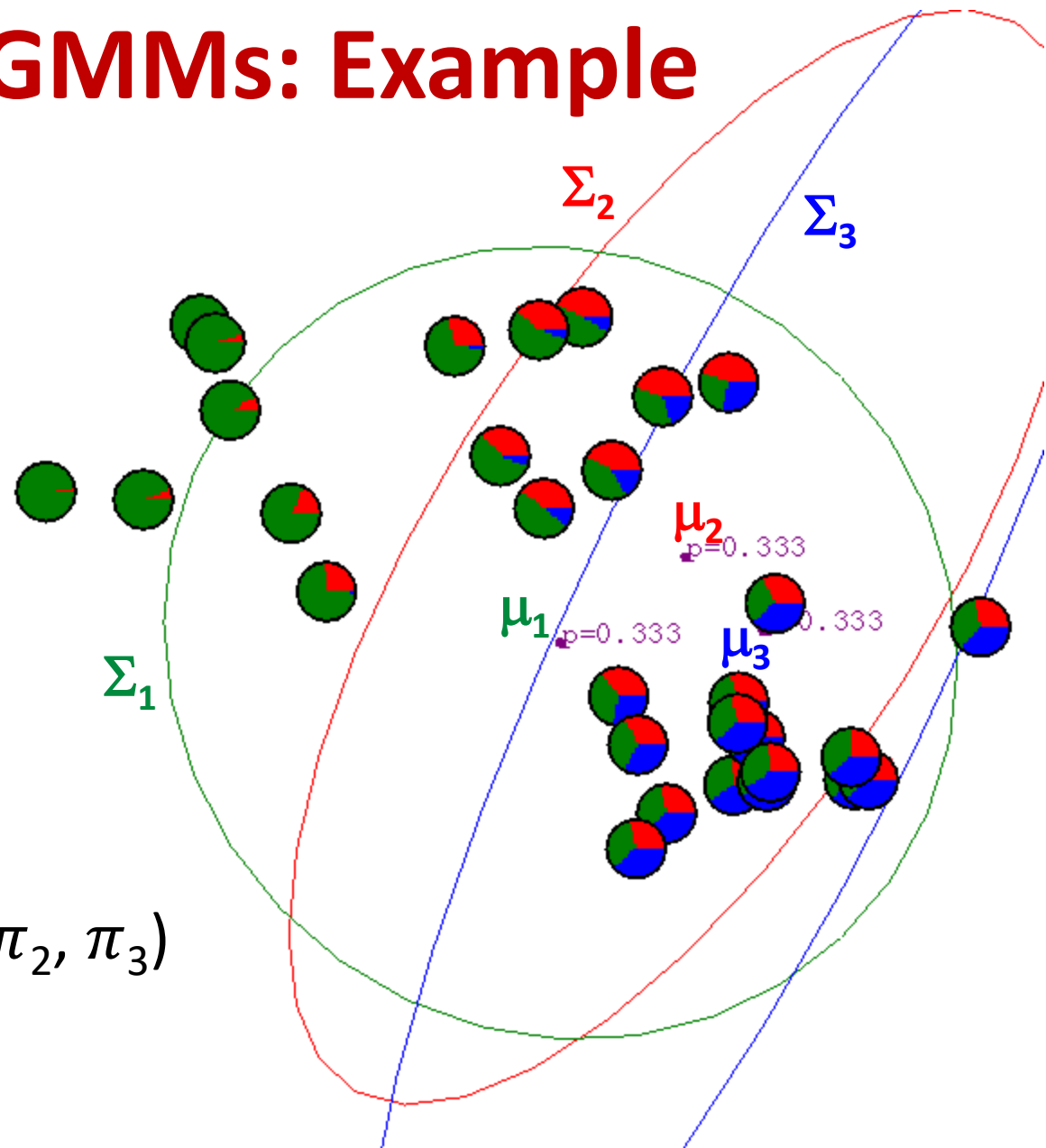
For a fixed set of Gaussian mixture model parameters, $\theta^{(t)}$, update the probability that each point, $\mathbf{x}^{(i)}$, belongs to cluster k , $p\left(Z_k^{(i)} = 1 \mid \mathbf{x}^{(i)}, \theta^{(t)}\right)$

M-step

For a fixed $p\left(Z_k^{(i)} = 1 \mid \mathbf{x}^{(i)}, \theta^{(t)}\right)$, update the estimate for each parameter, $\pi_k^{(t+1)}$, $\boldsymbol{\mu}_k^{(t+1)}$, $\Sigma_k^{(t+1)}$

Iterate between E and M steps

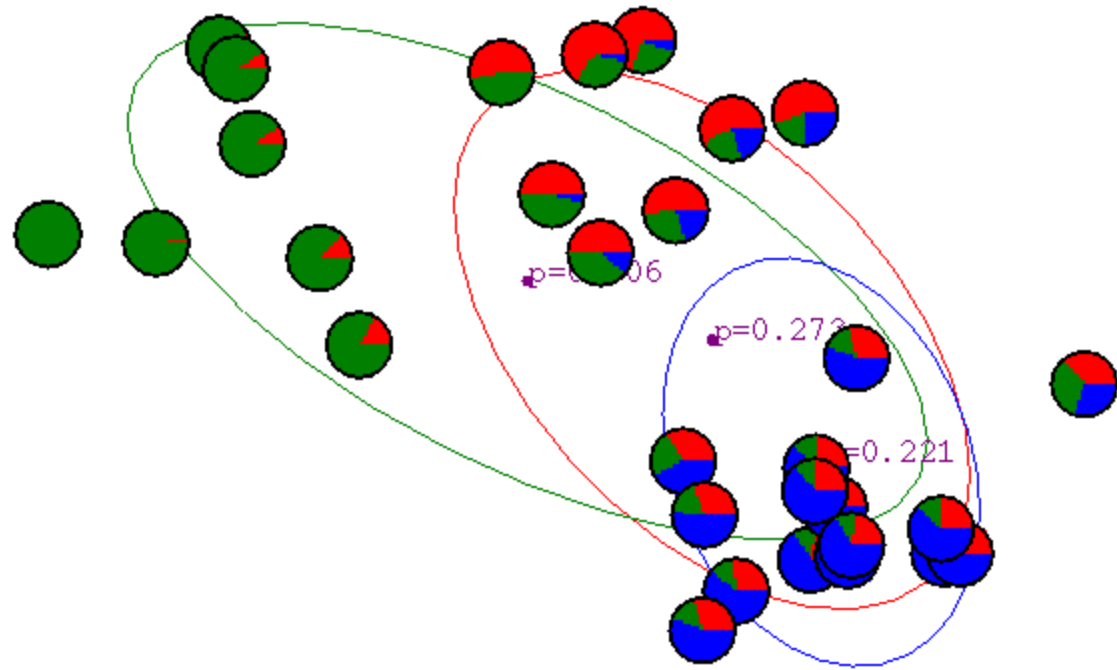
EM for GMMs: Example



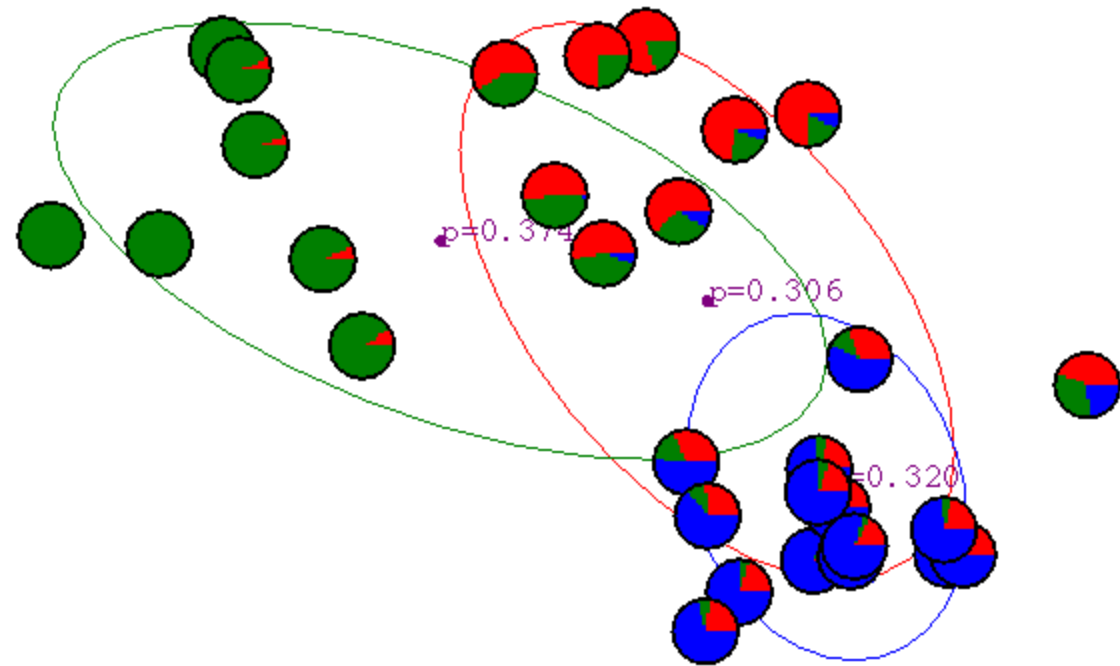
$$P(z_2 = 1 \mid x_j, \mu_1, \mu_2, \mu_3, \Sigma_1, \Sigma_2, \Sigma_3, \pi_1, \pi_2, \pi_3)$$



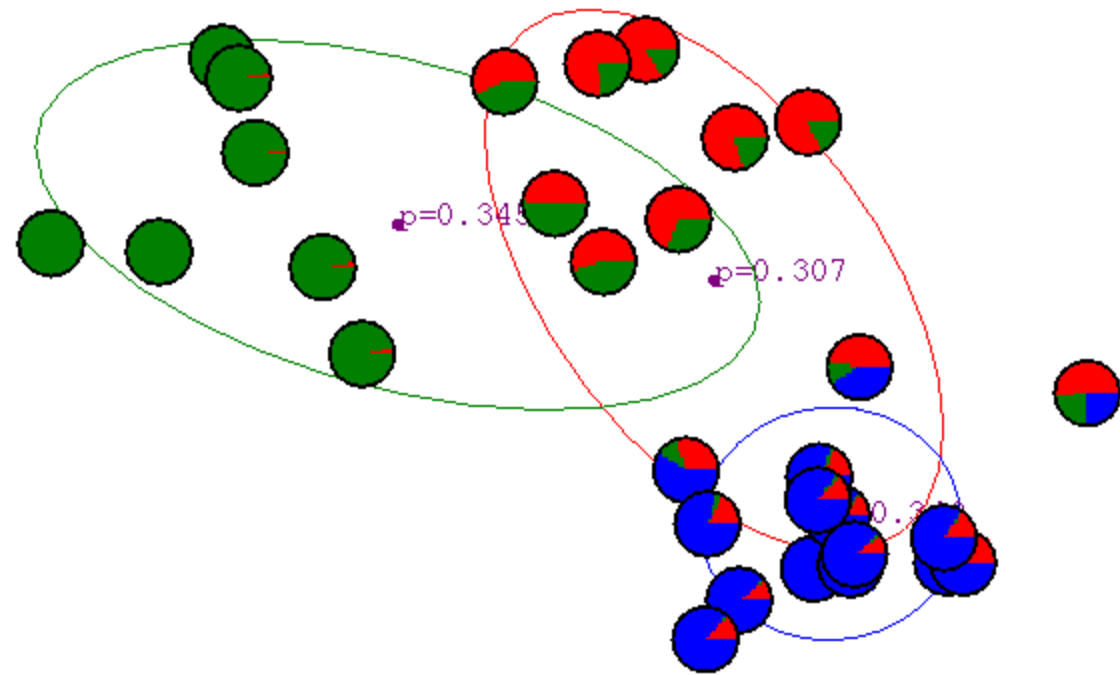
After 1st iteration



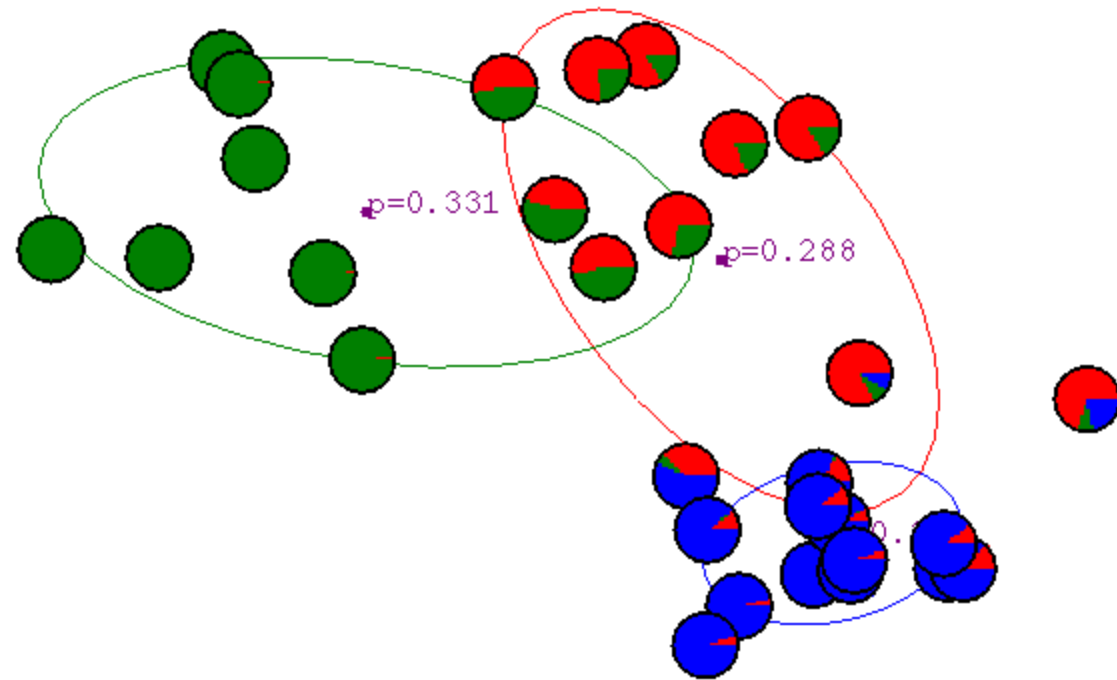
After 2nd iteration



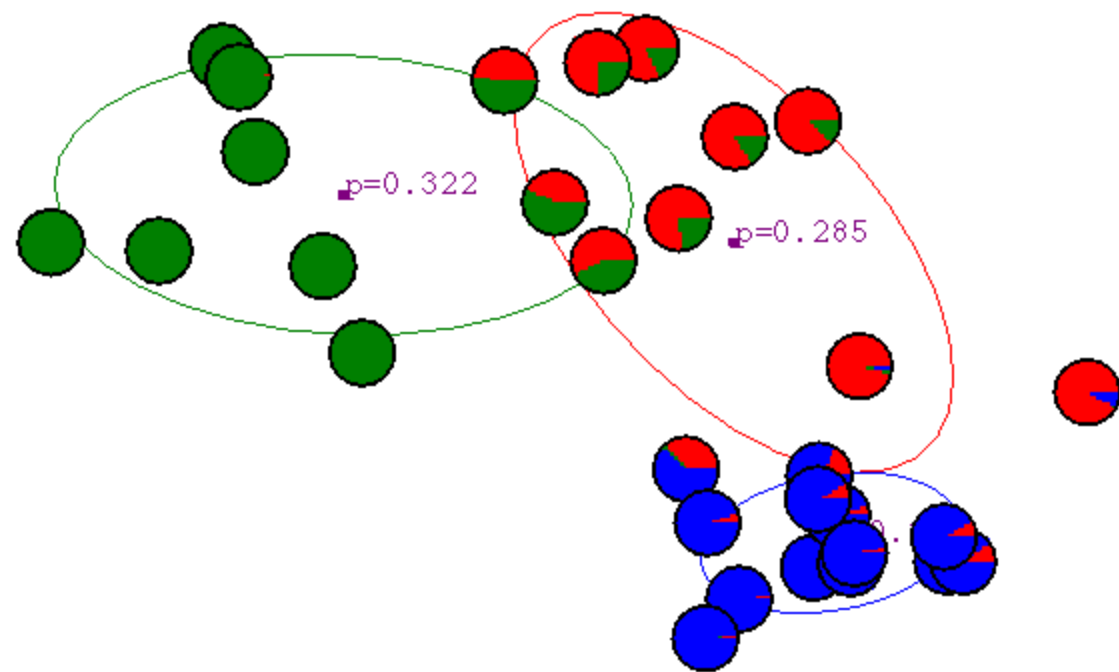
After 3rd iteration



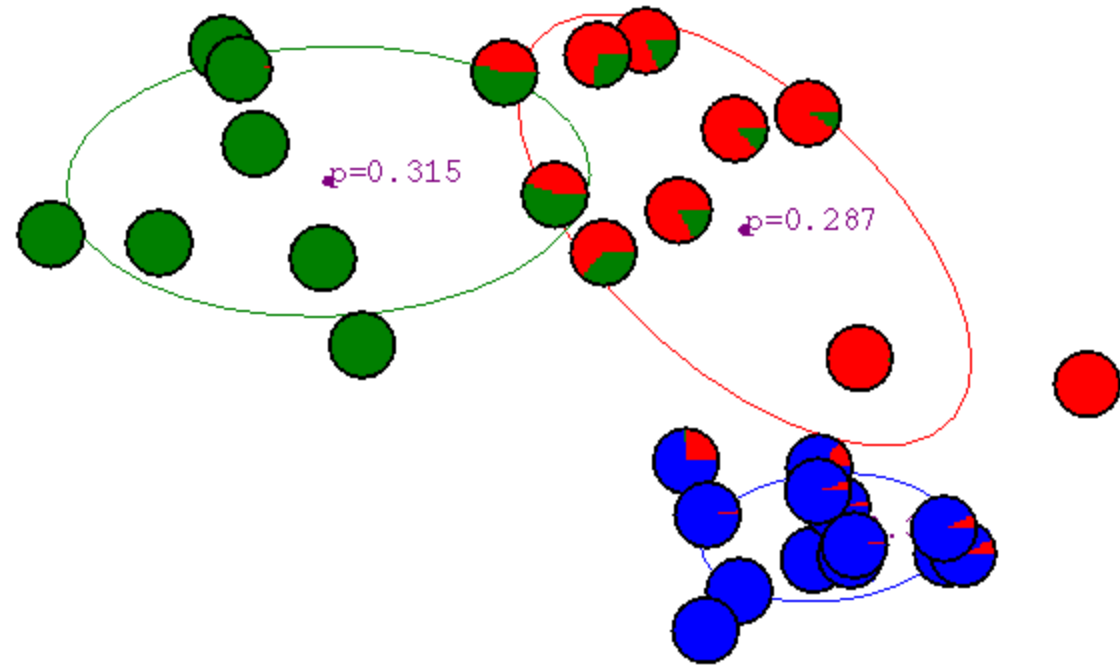
After 4th iteration



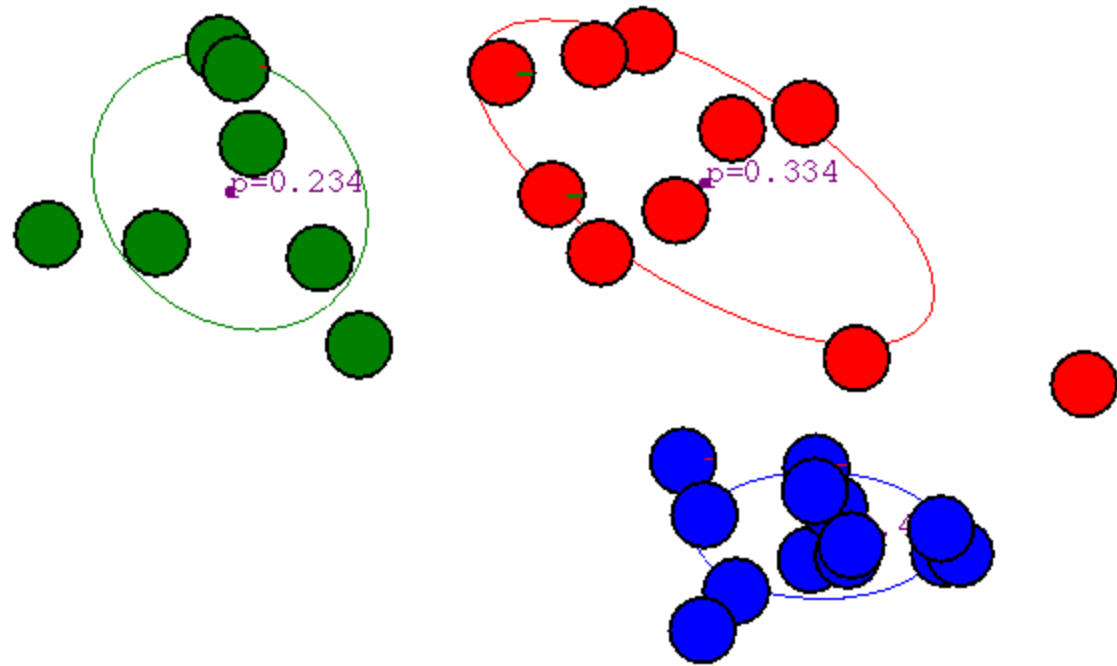
After 5th iteration



After 6th iteration



After 20th iteration



EM for GMMS

E-step

Estimated probabilistic cluster assignments $p \left(Z_k^{(i)} = 1 \mid \mathbf{x}^{(i)}, \boldsymbol{\theta}^{(t)} \right)$

Formulate **expected** value over Z of complete log likelihood

M-step

$$\left. \begin{array}{l} \pi_k^{(t+1)} \\ \boldsymbol{\mu}_k^{(t+1)} \\ \Sigma_k^{(t+1)} \end{array} \right\} = \operatorname{argmax}_{\boldsymbol{\theta}} E_{Z|X, \boldsymbol{\theta}^{(t)}} [\ell_c(\boldsymbol{\theta} \mid \mathcal{D}_c)]$$

$$\begin{aligned} \text{Complete Log likelihood } \mathcal{D} = \{\mathbf{x}^{(i)}, \mathbf{z}^{(i)}\} \\ \ell_c(\boldsymbol{\theta} \mid \mathcal{D}_c) = \log \prod_i^N p(\mathbf{x}^{(i)}, \mathbf{z}^{(i)} \mid \boldsymbol{\theta}) \end{aligned}$$

EM for GMMS

E-step


$$\underline{p\left(z_k^{(i)} = 1 \mid \mathbf{x}^{(i)}, \boldsymbol{\theta}^{(t)}\right)} \leftarrow \frac{\pi_k^{(t)} \mathcal{N}\left(\mathbf{x}^{(i)}; \boldsymbol{\mu}_k^{(t)}, \boldsymbol{\Sigma}_k^{(t)}\right)}{\sum_{j=1}^K \pi_j^{(t)} \mathcal{N}\left(\mathbf{x}^{(i)}; \boldsymbol{\mu}_j^{(t)}, \boldsymbol{\Sigma}_j^{(t)}\right)}, \forall i, k$$

M-step

$$\pi_k^{(t+1)} \leftarrow \frac{\sum_{i=1}^N P\left(z_k^{(i)} = 1 \mid \mathbf{x}^{(i)}, \boldsymbol{\theta}^{(t)}\right)}{N}, \forall k$$

$$\boldsymbol{\mu}_k^{(t+1)} \leftarrow \frac{\sum_{i=1}^N P\left(z_k^{(i)} = 1 \mid \mathbf{x}^{(i)}, \boldsymbol{\theta}^{(t)}\right) \mathbf{x}^{(i)}}{\sum_{i=1}^N P\left(z_k^{(i)} = 1 \mid \mathbf{x}^{(i)}, \boldsymbol{\theta}^{(t)}\right)}, \forall k$$

$$\boldsymbol{\Sigma}_k^{(t+1)} \leftarrow \frac{\sum_{i=1}^N P\left(z_k^{(i)} = 1 \mid \mathbf{x}^{(i)}, \boldsymbol{\theta}^{(t)}\right) \left(\mathbf{x}^{(i)} - \boldsymbol{\mu}_k^{(t+1)}\right) \left(\mathbf{x}^{(i)} - \boldsymbol{\mu}_k^{(t+1)}\right)^T}{\sum_{i=1}^N P\left(z_k^{(i)} = 1 \mid \mathbf{x}^{(i)}, \boldsymbol{\theta}^{(t)}\right)}, \forall k$$

Optimization formulation

Supervised

vs

Unsupervised

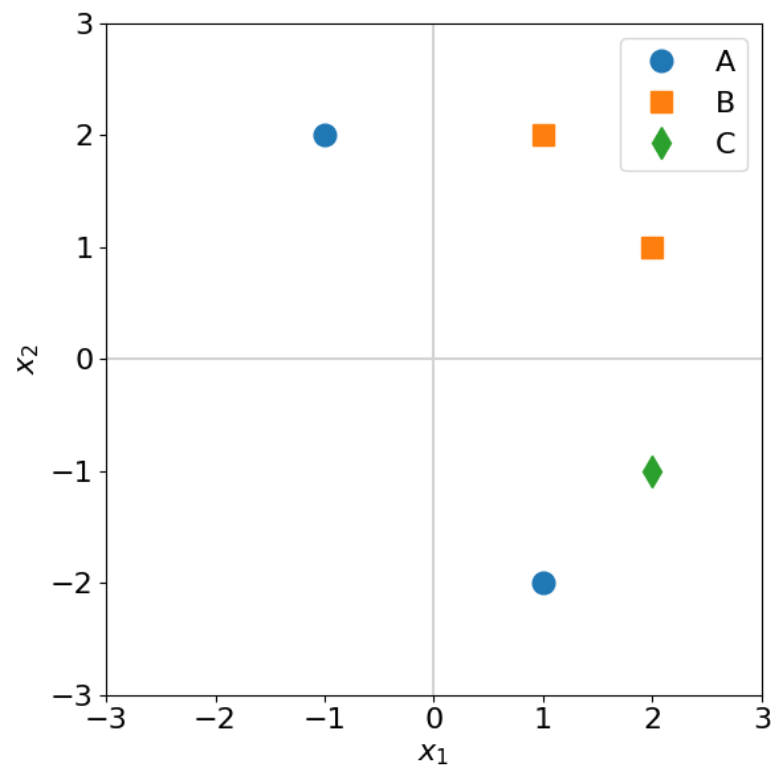
Gaussian discriminant analysis

vs

Gaussian Mixture Models

Generative Models: Supervised vs Unsupervised

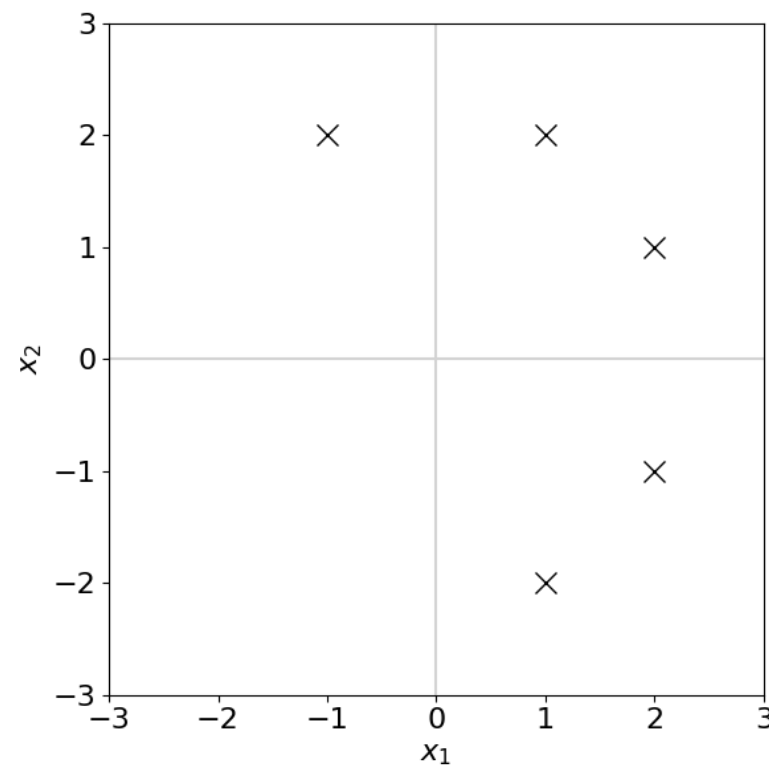
Gaussian discriminant analysis vs Gaussian mixture models



$$Y \sim \text{Categorical}(\pi_1, \pi_2, \pi_3)$$

$$X_{Y=k} \sim \mathcal{N}(\mu_k, \sigma_k^2)$$

$$\mathcal{D} = \{x^{(i)}, y^{(i)}\}_{i=1}^N$$



$$Z \sim \text{Categorical}(\pi_1, \pi_2, \pi_3)$$

$$X_{Z=k} \sim \mathcal{N}(\mu_k, \sigma_k^2)$$

$$\mathcal{D} = \{x^{(i)}\}_{i=1}^N$$

Generative Model: Supervised

MLE: Gaussian discriminant analysis

$$\operatorname{argmax}_{\theta} \prod_i^N p(\mathbf{x}^{(i)}, \mathbf{y}^{(i)})$$

Reminder

$\mathbf{y}^{(i)}$
↑
categorical

$$= \begin{bmatrix} y_1^{(i)} \\ y_2^{(i)} \\ \vdots \\ y_K^{(i)} \end{bmatrix}$$

$$\underline{P(Y_k^{(i)} = 1)}$$

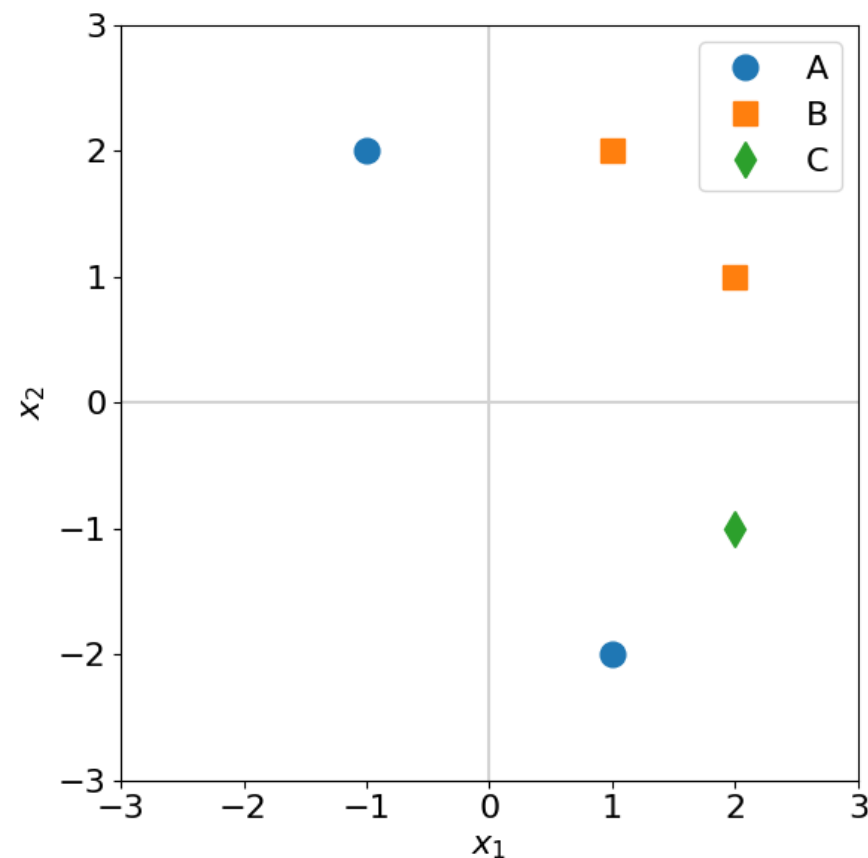
$$P(Y_{k=1}^{(i)} = 1) \triangleq \pi_1$$

$$P(Y_{k=2}^{(i)} = 1) \triangleq \pi_2$$
$$\vdots$$

$$Y \sim \text{Categorical}(\pi_1, \pi_2, \pi_3)$$

$$X_{Y=k} \sim \mathcal{N}(\mu_k, \sigma_k^2).$$

$$\mathcal{D} = \{x^{(i)}, y^{(i)}\}_{i=1}^N$$



Generative Model: Supervised

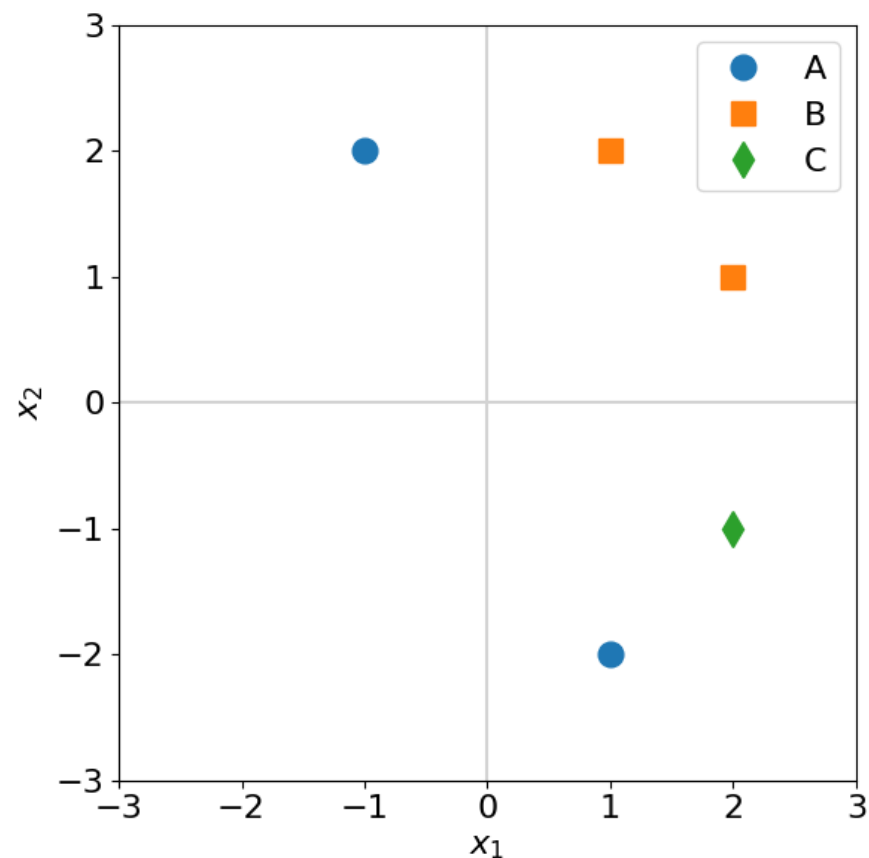
MLE: Gaussian discriminant analysis

$$\begin{aligned} & \operatorname{argmax}_{\theta} \prod_i^N p(\mathbf{x}^{(i)}, \mathbf{y}^{(i)}) \\ &= \operatorname{argmax}_{\theta} \prod_i^N \prod_k^K p(\mathbf{x}^{(i)}, Y_k^{(i)} = 1)^{y_k^{(i)}} \\ &= \operatorname{argmax}_{\theta} \prod_i^N \prod_k^K \left(p(Y_k^{(i)} = 1) p(\mathbf{x}^{(i)} | Y_k^{(i)} = 1) \right)^{y_k^{(i)}} \\ &= \operatorname{argmax}_{\theta} \prod_i^N \prod_k^K \left(\pi_k |\Sigma_k|^{-\frac{1}{2}} e^{-\frac{1}{2}(\mathbf{x}^{(i)} - \boldsymbol{\mu}_k)^T \Sigma_k^{-1} (\mathbf{x}^{(i)} - \boldsymbol{\mu}_k)} \right)^{y_k^{(i)}} \\ &= \operatorname{argmax}_{\theta} \log \prod_i^N \prod_k^K \left(\pi_k |\Sigma_k|^{-\frac{1}{2}} e^{-\frac{1}{2}(\mathbf{x}^{(i)} - \boldsymbol{\mu}_k)^T \Sigma_k^{-1} (\mathbf{x}^{(i)} - \boldsymbol{\mu}_k)} \right)^{y_k^{(i)}} \\ &= \operatorname{argmax}_{\theta} \sum_i^N \sum_k^K y_k^{(i)} \log \left(\pi_k |\Sigma_k|^{-\frac{1}{2}} e^{-\frac{1}{2}(\mathbf{x}^{(i)} - \boldsymbol{\mu}_k)^T \Sigma_k^{-1} (\mathbf{x}^{(i)} - \boldsymbol{\mu}_k)} \right) \end{aligned}$$

$$Y \sim \text{Categorical}(\pi_1, \pi_2, \pi_3)$$

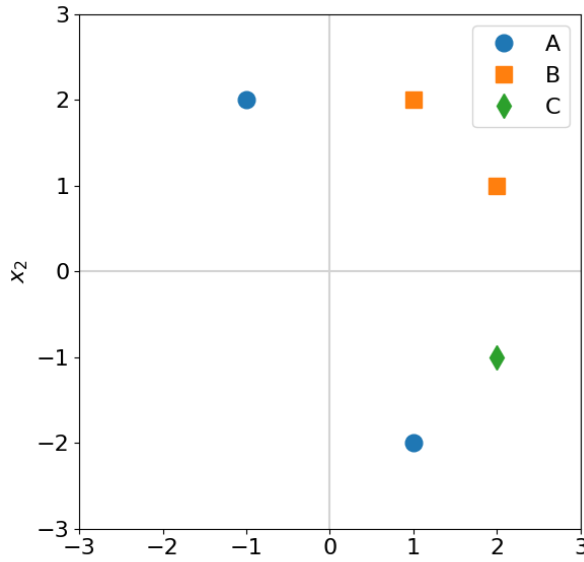
$$X_{Y=k} \sim \mathcal{N}(\mu_k, \sigma_k^2).$$

$$\mathcal{D} = \{x^{(i)}, y^{(i)}\}_{i=1}^N$$

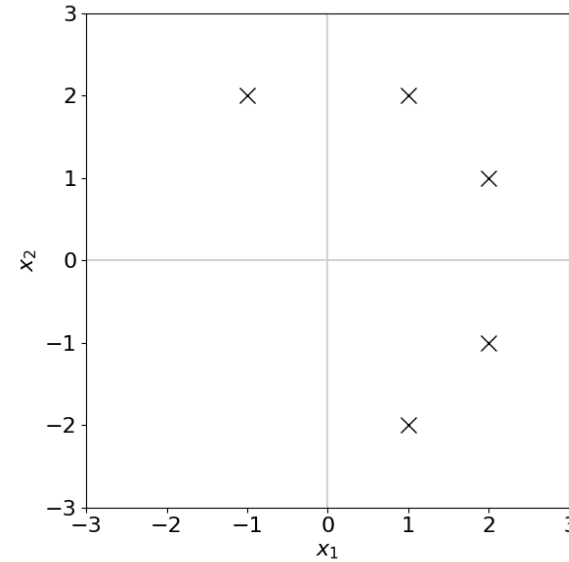


Generative Models: Supervised vs Unsupervised

Discriminant analysis vs Gaussian mixture models



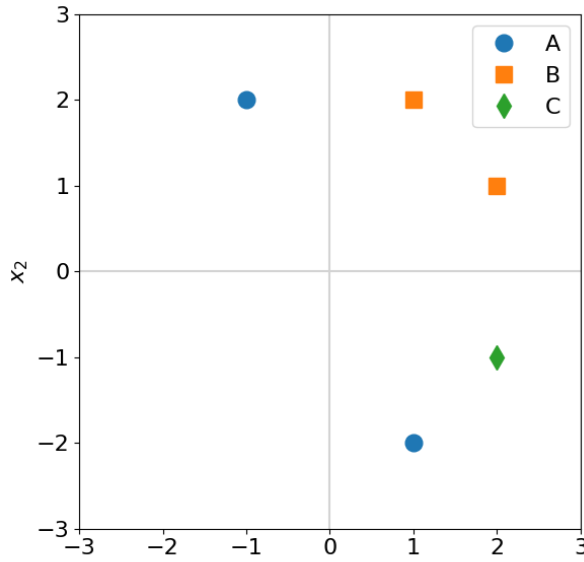
$$\begin{aligned}
 & \operatorname{argmax}_{\theta} \prod_i^{N^{x_1}} p(\mathbf{x}^{(i)}, \mathbf{y}^{(i)}) \\
 &= \operatorname{argmax}_{\theta} \prod_i^N \prod_k^K p\left(\mathbf{x}^{(i)}, Y_k^{(i)} = 1\right)^{y_k^{(i)}} \\
 &= \operatorname{argmax}_{\theta} \prod_i^N \prod_k^K \left(p\left(Y_k^{(i)} = 1\right) p\left(\mathbf{x}^{(i)} \mid Y_k^{(i)} = 1\right) \right)^{y_k^{(i)}}
 \end{aligned}$$



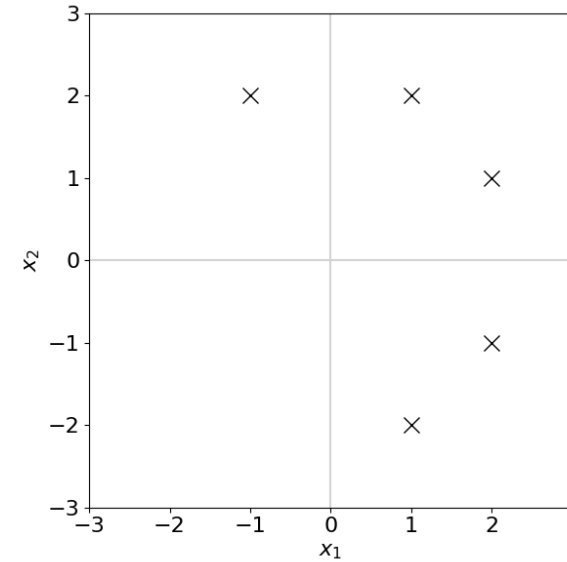
$$\begin{aligned}
 & \operatorname{argmax}_{\theta} \prod_i^N p(\mathbf{x}^{(i)}) \\
 &= \operatorname{argmax}_{\theta} \prod_i^N
 \end{aligned}$$

Generative Models: Supervised vs Unsupervised

Discriminant analysis vs Gaussian mixture models



$$\begin{aligned}
 & \operatorname{argmax}_{\theta} \prod_i^{N^{x_1}} p(\mathbf{x}^{(i)}, \mathbf{y}^{(i)}) \\
 &= \operatorname{argmax}_{\theta} \prod_i^N \prod_k^K p(\mathbf{x}^{(i)}, Y_k^{(i)} = 1) \quad y_k^{(i)} \\
 &= \operatorname{argmax}_{\theta} \prod_i^N \prod_k^K \left(p(Y_k^{(i)} = 1) p(\mathbf{x}^{(i)} \mid Y_k^{(i)} = 1) \right)^{y_k^{(i)}}
 \end{aligned}$$



$$\begin{aligned}
 & \operatorname{argmax}_{\theta} \prod_i^N p(\mathbf{x}^{(i)}) \\
 &= \operatorname{argmax}_{\theta} \prod_i^N \sum_{k=1}^K p(\mathbf{x}^{(i)}, Z_k^{(i)} = 1) \\
 &= \operatorname{argmax}_{\theta} \prod_i^N \sum_{k=1}^K p(Z_k^{(i)} = 1) p(\mathbf{x}^{(i)} \mid Z_k^{(i)} = 1)
 \end{aligned}$$

Generative Models: Supervised vs Unsupervised

Discriminant analysis vs Gaussian mixture models

$$\begin{aligned} & \operatorname{argmax}_{\theta} \prod_i^N p(\mathbf{x}^{(i)}, \mathbf{y}^{(i)} | \theta) \\ &= \operatorname{argmax}_{\theta} \prod_i^N \prod_k^K p(\mathbf{x}^{(i)}, Y_k^{(i)} = 1 | \theta)^{y_k^{(i)}} \end{aligned}$$

prod w/ indicator trick

$$= \operatorname{argmax}_{\theta} \prod_i^N \prod_k^K \left(p(Y_k^{(i)} = 1) p(\mathbf{x}^{(i)} | Y_k^{(i)} = 1) \right)^{y_k^{(i)}}$$

$$= \operatorname{argmax}_{\theta} \prod_i^N \prod_k^K \left(\pi_k |\Sigma_k|^{-\frac{1}{2}} e^{-\frac{1}{2}(\mathbf{x}^{(i)} - \boldsymbol{\mu}_k)^T \Sigma_k^{-1} (\mathbf{x}^{(i)} - \boldsymbol{\mu}_k)} \right)^{y_k^{(i)}}$$

$$= \operatorname{argmax}_{\theta} \log \prod_i^N \prod_k^K \left(\pi_k |\Sigma_k|^{-\frac{1}{2}} e^{-\frac{1}{2}(\mathbf{x}^{(i)} - \boldsymbol{\mu}_k)^T \Sigma_k^{-1} (\mathbf{x}^{(i)} - \boldsymbol{\mu}_k)} \right)^{y_k^{(i)}}$$

$$= \operatorname{argmax}_{\theta} \sum_i^N \sum_k^K y_k^{(i)} \log \left(\pi_k |\Sigma_k|^{-\frac{1}{2}} e^{-\frac{1}{2}(\mathbf{x}^{(i)} - \boldsymbol{\mu}_k)^T \Sigma_k^{-1} (\mathbf{x}^{(i)} - \boldsymbol{\mu}_k)} \right)$$

$$\begin{aligned} & \operatorname{argmax}_{\theta} \prod_i^N p(\mathbf{x}^{(i)} | \theta) \\ &= \operatorname{argmax}_{\theta} \prod_i^N \sum_{k=1}^K p(\mathbf{x}^{(i)}, Z_k^{(i)} = 1 | \theta) \end{aligned}$$

sum *no z*

$$= \operatorname{argmax}_{\theta} \prod_i^N \sum_{k=1}^K p(Z_k^{(i)} = 1) p(\mathbf{x}^{(i)} | Z_k^{(i)} = 1)$$

$$= \operatorname{argmax}_{\theta} \prod_i^N \sum_{k=1}^K \pi_k |\Sigma_k|^{-\frac{1}{2}} e^{-\frac{1}{2}(\mathbf{x}^{(i)} - \boldsymbol{\mu}_k)^T \Sigma_k^{-1} (\mathbf{x}^{(i)} - \boldsymbol{\mu}_k)}$$

$$= \operatorname{argmax}_{\theta} \log \prod_i^N \sum_{k=1}^K \pi_k |\Sigma_k|^{-\frac{1}{2}} e^{-\frac{1}{2}(\mathbf{x}^{(i)} - \boldsymbol{\mu}_k)^T \Sigma_k^{-1} (\mathbf{x}^{(i)} - \boldsymbol{\mu}_k)}$$

$$= \operatorname{argmax}_{\theta} \sum_i^N \log \sum_{k=1}^K \pi_k |\Sigma_k|^{-\frac{1}{2}} e^{-\frac{1}{2}(\mathbf{x}^{(i)} - \boldsymbol{\mu}_k)^T \Sigma_k^{-1} (\mathbf{x}^{(i)} - \boldsymbol{\mu}_k)}$$

Expectation-Maximization for GMM

Supervised vs Unsupervised Log Likelihood

Supervised $\mathcal{D}_{supervised} = \{\mathbf{x}^{(i)}, \mathbf{y}^{(i)}\}$

$$\ell(\theta \mid \mathcal{D}_{supervised})$$

$$= \log \prod_i^N p(\mathbf{x}^{(i)}, \mathbf{y}^{(i)})$$

$$= \log \prod_i^N \prod_k^K p(\mathbf{x}^{(i)}, Y_k^{(i)} = 1)^{y_k^{(i)}}$$

$$= \log \prod_i^N \prod_k^K \left(p(Y_k^{(i)} = 1) p(\mathbf{x}^{(i)} \mid Y_k^{(i)} = 1) \right)^{y_k^{(i)}}$$

$$= \sum_i^N \sum_k^K y_k^{(i)} \log \left(p(Y_k^{(i)} = 1) p(\mathbf{x}^{(i)} \mid Y_k^{(i)} = 1) \right)$$

Log likelihood $\mathcal{D} = \{\mathbf{x}^{(i)}\}$

$$\ell(\theta \mid \mathcal{D})$$

$$= \log \prod_i^N p(\mathbf{x}^{(i)})$$

$$= \log \prod_i^N \sum_k^K p(\mathbf{x}^{(i)}, Z_k^{(i)} = 1)$$

$$= \log \prod_i^N \sum_k^K p(Z_k^{(i)} = 1) p(\mathbf{x}^{(i)} \mid Z_k^{(i)} = 1)$$

$$= \sum_i^N \log \sum_k^K p(Z_k^{(i)} = 1) p(\mathbf{x}^{(i)} \mid Z_k^{(i)} = 1)$$

Unsupervised vs Supervised

Log is stuck outside of the summation over z ☹️

DATASET

$$\ell(\theta \mid \textit{DATASET})$$

Unsupervised

$$\mathcal{D} = \{\mathbf{x}^{(i)}\}$$

$$\sum_i^N \log \sum_{k=1}^K p\left(z_k^{(i)} = 1\right) p\left(\mathbf{x}^{(i)} \mid z_k^{(i)} = 1\right)$$

Supervised

$$\mathcal{D}_{\textit{supervised}} = \{\mathbf{x}^{(i)}, y^{(i)}\}$$

$$\sum_i^N \sum_k^K y_k^{(i)} \log \left(p\left(Y_k^{(i)} = 1\right) p\left(\mathbf{x}^{(i)} \mid Y_k^{(i)} = 1\right) \right)$$

Unsupervised vs Supervised vs Complete Log Likelihood

Complete log likelihood: Pretend that we have the cluster labels in the dataset

DATASET

$$\ell(\theta \mid \textit{DATASET})$$

Unsupervised

$$\mathcal{D} = \{\mathbf{x}^{(i)}\}$$

$$\sum_i^N \log \sum_{k=1}^K p\left(z_k^{(i)} = 1\right) p\left(\mathbf{x}^{(i)} \mid z_k^{(i)} = 1\right)$$

Supervised

$$\mathcal{D}_{\textit{supervised}} = \{\mathbf{x}^{(i)}, y^{(i)}\}$$

$$\sum_i^N \sum_k^K y_k^{(i)} \log \left(p\left(Y_k^{(i)} = 1\right) p\left(\mathbf{x}^{(i)} \mid Y_k^{(i)} = 1\right) \right)$$

Complete

$$\mathcal{D}_{\textit{complete}} = \{\mathbf{x}^{(i)}, z^{(i)}\}$$

$$\sum_i^N \sum_k^K z_k^{(i)} \log \left(p\left(z_k^{(i)} = 1\right) p\left(\mathbf{x}^{(i)} \mid z_k^{(i)} = 1\right) \right)$$

Expected Value of Complete Log Likelihood

Supervised $\mathcal{D}_{supervised} = \{\mathbf{x}^{(i)}, \mathbf{y}^{(i)}\}$ Complete Log likelihood $\mathcal{D}_c = \{\mathbf{x}^{(i)}, \mathbf{z}^{(i)}\}$

$$\ell(\theta \mid \mathcal{D}_{supervised})$$

$$= \log \prod_i^N p(\mathbf{x}^{(i)}, \mathbf{y}^{(i)})$$

$$= \sum_i^N \sum_k^K y_k^{(i)} \log \left(p(Y_k^{(i)} = 1) p(\mathbf{x}^{(i)} \mid Y_k^{(i)} = 1) \right)$$

$$\mathbb{E}_{\mathbf{Z} \mid \mathbf{X}, \theta} [\ell_c(\theta \mid \mathcal{D}_c)]$$

$$= \mathbb{E}_{\mathbf{Z} \mid \mathbf{X}, \theta} \left[\log \prod_i^N p(\mathbf{x}^{(i)}, \mathbf{z}^{(i)}) \right]$$

$$= \sum_i^N \sum_k^K p(Z_k^{(i)} = 1 \mid \mathbf{x}^{(i)}) \log \left(p(Z_k^{(i)} = 1) p(\mathbf{x}^{(i)} \mid Z_k^{(i)} = 1) \right)$$

Details: Expected Value of Complete Log Likelihood

Supervised $\mathcal{D}_{supervised} = \{\mathbf{x}^{(i)}, \mathbf{y}^{(i)}\}$ Complete Log likelihood $\mathcal{D}_c = \{\mathbf{x}^{(i)}, \mathbf{z}^{(i)}\}$

$$\ell(\theta \mid \mathcal{D}_{supervised})$$

$$= \log \prod_i^N p(\mathbf{x}^{(i)}, \mathbf{y}^{(i)})$$

$$= \log \prod_i^N \prod_k^K p(\mathbf{x}^{(i)}, Y_k^{(i)} = 1)^{y_k^{(i)}}$$

$$= \log \prod_i^N \prod_k^K \left(p(Y_k^{(i)} = 1) p(\mathbf{x}^{(i)} \mid Y_k^{(i)} = 1) \right)^{y_k^{(i)}}$$

$$= \sum_i^N \sum_k^K y_k^{(i)} \log \left(p(Y_k^{(i)} = 1) p(\mathbf{x}^{(i)} \mid Y_k^{(i)} = 1) \right)$$

$$\mathbb{E}_{\mathbf{Z} \mid \mathbf{X}, \theta} [\ell_c(\theta \mid \mathcal{D}_c)]$$

$$= \mathbb{E}_{\mathbf{Z} \mid \mathbf{X}, \theta} \left[\log \prod_i^N p(\mathbf{x}^{(i)}, \mathbf{z}^{(i)}) \right]$$

$$= \mathbb{E}_{\mathbf{Z} \mid \mathbf{X}, \theta} \left[\log \prod_i^N \prod_k^K p(\mathbf{x}^{(i)}, Z_k^{(i)} = 1)^{z_k^{(i)}} \right]$$

$$= \mathbb{E}_{\mathbf{Z} \mid \mathbf{X}, \theta} \left[\log \prod_i^N \prod_k^K \left(p(Z_k^{(i)} = 1) p(\mathbf{x}^{(i)} \mid Z_k^{(i)} = 1) \right)^{z_k^{(i)}} \right]$$

$$= \mathbb{E}_{\mathbf{Z} \mid \mathbf{X}, \theta} \left[\sum_i^N \sum_k^K z_k^{(i)} \log \left(p(Z_k^{(i)} = 1) p(\mathbf{x}^{(i)} \mid Z_k^{(i)} = 1) \right) \right]$$

$$= \sum_i^N \sum_k^K \mathbb{E}_{\mathbf{Z} \mid \mathbf{X}, \theta} [Z_k^{(i)}] \log \left(p(Z_k^{(i)} = 1) p(\mathbf{x}^{(i)} \mid Z_k^{(i)} = 1) \right)$$

$$= \sum_i^N \sum_k^K p(Z_k^{(i)} = 1 \mid \mathbf{x}^{(i)}) \log \left(p(Z_k^{(i)} = 1) p(\mathbf{x}^{(i)} \mid Z_k^{(i)} = 1) \right)$$

Details: Expected Value of Complete Log Likelihood

Even more details

$$\begin{aligned}\mathbb{E}_{Z|X,\theta} [Z_k^{(i)}] &= \sum_{z_k \in \{0,1\}} z_k p(Z_k^{(i)} = z_k \mid \mathbf{x}^{(i)}) \\ &= 0 + 1 \cdot p(Z_k^{(i)} = z_k \mid \mathbf{x}^{(i)}) \\ &= p(Z_k^{(i)} = z_k \mid \mathbf{x}^{(i)})\end{aligned}$$

Complete Log likelihood $\mathcal{D}_c = \{\mathbf{x}^{(i)}, \mathbf{z}^{(i)}\}$

$$\begin{aligned}\mathbb{E}_{Z|X,\theta} [\ell_c(\theta \mid \mathcal{D}_c)] &= \mathbb{E}_{Z|X,\theta} \left[\log \prod_i^N p(\mathbf{x}^{(i)}, \mathbf{z}^{(i)}) \right] \\ &= \mathbb{E}_{Z|X,\theta} \left[\log \prod_i^N \prod_k^K p(\mathbf{x}^{(i)}, z_k^{(i)} = 1)^{z_k^{(i)}} \right] \\ &= \mathbb{E}_{Z|X,\theta} \left[\log \prod_i^N \prod_k^K \left(p(z_k^{(i)} = 1) p(\mathbf{x}^{(i)} \mid z_k^{(i)} = 1) \right)^{z_k^{(i)}} \right] \\ &= \mathbb{E}_{Z|X,\theta} \left[\sum_i^N \sum_k^K z_k^{(i)} \log \left(p(z_k^{(i)} = 1) p(\mathbf{x}^{(i)} \mid z_k^{(i)} = 1) \right) \right] \\ &= \sum_i^N \sum_k^K \mathbb{E}_{Z|X,\theta} [Z_k^{(i)}] \log \left(p(z_k^{(i)} = 1) p(\mathbf{x}^{(i)} \mid z_k^{(i)} = 1) \right) \\ &= \sum_i^N \sum_k^K p(z_k^{(i)} = 1 \mid \mathbf{x}^{(i)}) \log \left(p(z_k^{(i)} = 1) p(\mathbf{x}^{(i)} \mid z_k^{(i)} = 1) \right)\end{aligned}$$

EM for GMMS

E-step

Estimated probabilistic cluster assignments $p \left(Z_k^{(i)} = 1 \mid \mathbf{x}^{(i)}, \boldsymbol{\theta}^{(t)} \right)$

Formulate **expected** value over Z of complete log likelihood

M-step

$$\left. \begin{array}{l} \pi_k^{(t+1)} \\ \boldsymbol{\mu}_k^{(t+1)} \\ \Sigma_k^{(t+1)} \end{array} \right\} = \operatorname{argmax}_{\boldsymbol{\theta}} E_{Z|X, \boldsymbol{\theta}^{(t)}} [\ell_c(\boldsymbol{\theta} \mid \mathcal{D}_c)]$$

$$\begin{aligned} \text{Complete Log likelihood } \mathcal{D} = \{\mathbf{x}^{(i)}, \mathbf{z}^{(i)}\} \\ \ell_c(\boldsymbol{\theta} \mid \mathcal{D}_c) = \log \prod_i^N p(\mathbf{x}^{(i)}, \mathbf{z}^{(i)} \mid \boldsymbol{\theta}) \end{aligned}$$

EM for GMMS

Initialize parameters

For $t = 0$, $\pi_k^{(0)}$, $\boldsymbol{\mu}_k^{(0)}$, $\Sigma_k^{(0)}$

E-step

For a fixed set of Gaussian mixture model parameters, $\theta^{(t)}$, update the probability that each point, $\mathbf{x}^{(i)}$, belongs to cluster k , $p\left(Z_k^{(i)} = 1 \mid \mathbf{x}^{(i)}, \boldsymbol{\theta}^{(t)}\right)$

M-step

For a fixed $p\left(Z_k^{(i)} = 1 \mid \mathbf{x}^{(i)}, \boldsymbol{\theta}^{(t)}\right)$, update the estimate for each parameter, $\pi_k^{(t+1)}$, $\boldsymbol{\mu}_k^{(t+1)}$, $\Sigma_k^{(t+1)}$

Iterate between E and M steps

EM for GMMS

E-step

$$p\left(z_k^{(i)} = 1 \mid \mathbf{x}^{(i)}, \boldsymbol{\theta}^{(t)}\right) \leftarrow \frac{\pi_k^{(t)} \mathcal{N}\left(\mathbf{x}^{(i)}; \boldsymbol{\mu}_k^{(t)}, \boldsymbol{\Sigma}_k^{(t)}\right)}{\sum_{j=1}^K \pi_j^{(t)} \mathcal{N}\left(\mathbf{x}^{(i)}; \boldsymbol{\mu}_j^{(t)}, \boldsymbol{\Sigma}_j^{(t)}\right)}, \forall i, k$$

M-step

$$\pi_k^{(t+1)} \leftarrow \frac{\sum_{i=1}^N P\left(z_k^{(i)} = 1 \mid \mathbf{x}^{(i)}, \boldsymbol{\theta}^{(t)}\right)}{N}, \forall k$$

$$\boldsymbol{\mu}_k^{(t+1)} \leftarrow \frac{\sum_{i=1}^N P\left(z_k^{(i)} = 1 \mid \mathbf{x}^{(i)}, \boldsymbol{\theta}^{(t)}\right) \mathbf{x}^{(i)}}{\sum_{i=1}^N P\left(z_k^{(i)} = 1 \mid \mathbf{x}^{(i)}, \boldsymbol{\theta}^{(t)}\right)}, \forall k$$

$$\boldsymbol{\Sigma}_k^{(t+1)} \leftarrow \frac{\sum_{i=1}^N P\left(z_k^{(i)} = 1 \mid \mathbf{x}^{(i)}, \boldsymbol{\theta}^{(t)}\right) \left(\mathbf{x}^{(i)} - \boldsymbol{\mu}_k^{(t+1)}\right) \left(\mathbf{x}^{(i)} - \boldsymbol{\mu}_k^{(t+1)}\right)^T}{\sum_{i=1}^N P\left(z_k^{(i)} = 1 \mid \mathbf{x}^{(i)}, \boldsymbol{\theta}^{(t)}\right)}, \forall k$$