Expectation-Maximization

Review: definition of machine learning

- Function approximation

  Problem setting:

  Set of instances (examples) $X = \{ x^1, \ldots, x^n\}$

  Unknown target function $f: X \rightarrow Y$

  Set of function hypothesis $H = \{h \mid h : X \rightarrow Y\}$, $h \approx f$

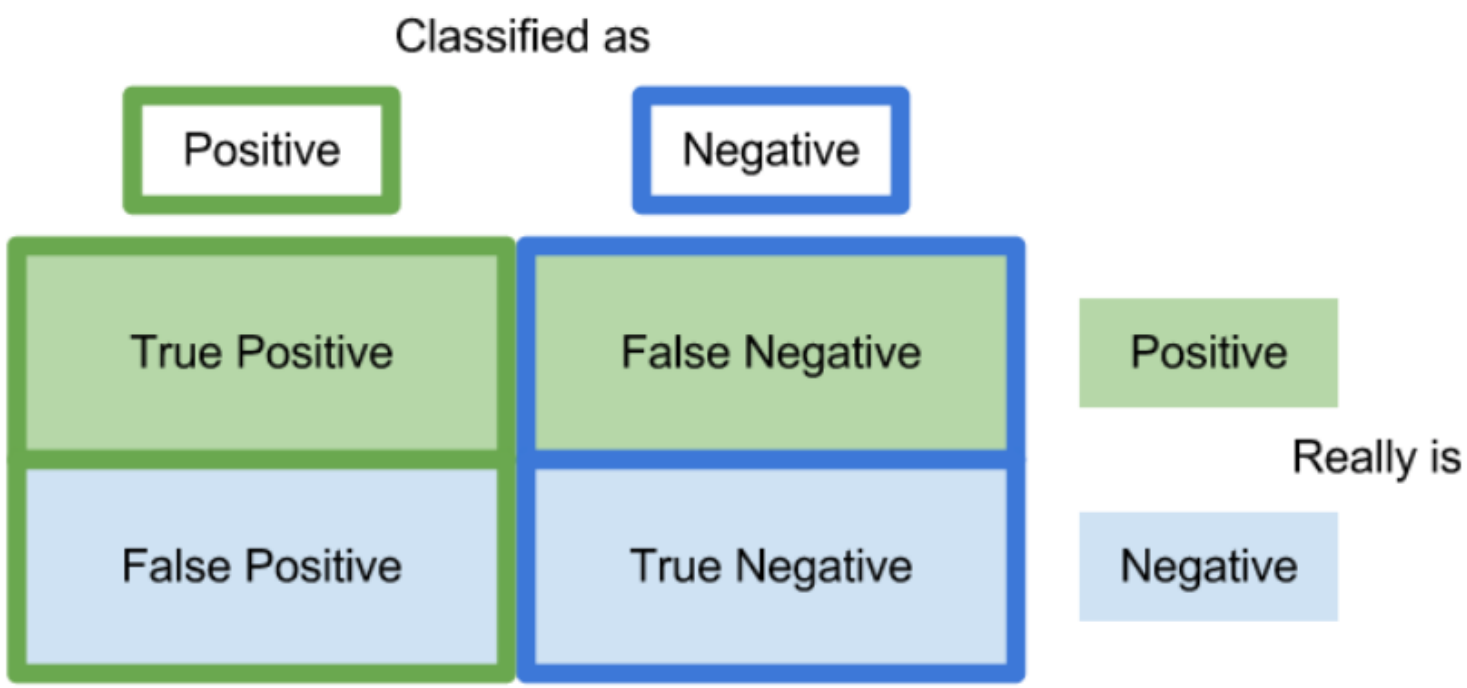
  Input:

  Training examples $\{ (x^i, y^i) \}$ of unknown target function f

  Output:

  Hypothesis $h \in H$ that best approximates target function f

Review: performance evaluation



$$\text{Precision} = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

$$\text{Recall} = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

Total data set: 100
Positive: 50, Negative:50

Program predicts
Positive:100

Precision = 50/(50+50) = 0.5
Recall = 50/50 = 1

Review: Maximum likelihood estimation (MLE)

- Task: rolling coins
- Data: observed set D of P(x = h) = θ and P(x = t) = 1- θ

Data D : h, t, t, h, h

$$P(D|\theta) = \theta(1-\theta)(1-\theta)\theta\theta = \theta^{a_1}(1-\theta)^{a_0}$$

- Learning (estimating) of θ by MLE

: choose θ that maximizes the probability of observed data

$$\begin{aligned}\widehat{\theta} &= \arg\max_{\theta} \; P(\mathcal{D} \mid \theta) \\ &= \arg\max_{\theta} \; \ln P(\mathcal{D} \mid \theta) \\ &= \arg\max_{\theta} \; \ln \theta^{a_1} (1-\theta)^{a_0}\end{aligned}$$

$$\frac{\partial}{\partial\theta}\Big(a_1 \ln\theta + a_0 \ln(1-\theta)\Big) = 0$$

$$a_1 \frac{1}{\theta} + a_0 \frac{-1}{1-\theta} = 0$$

$$\theta = \frac{a_1}{a_1 + a_0}$$

Binomial Distribution

Bernoulli trial

- two possible outcomes (S and P)

- constant probability Pr(S) = p

- independent trials

Binomial Distribution

- n Bernoulli trials

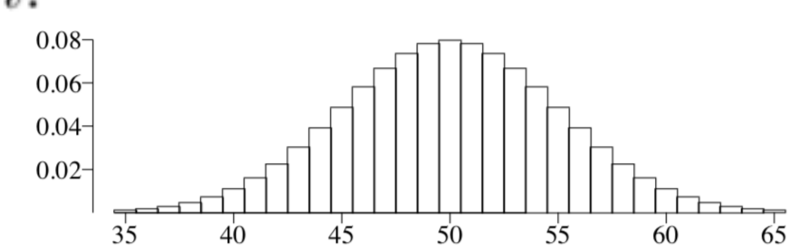
- Let X denote the total number of successes in the n trials

- The probability distribution of X is given as follows

$$P(X = x) = \frac{n!}{x!(n-x)!} p^x q^{n-x}, \text{ for } x = 0, 1, \ldots, n.$$

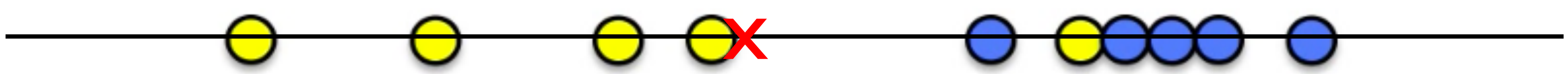
nCx

n=100, p = 0.5



Who am I?

Who am I?

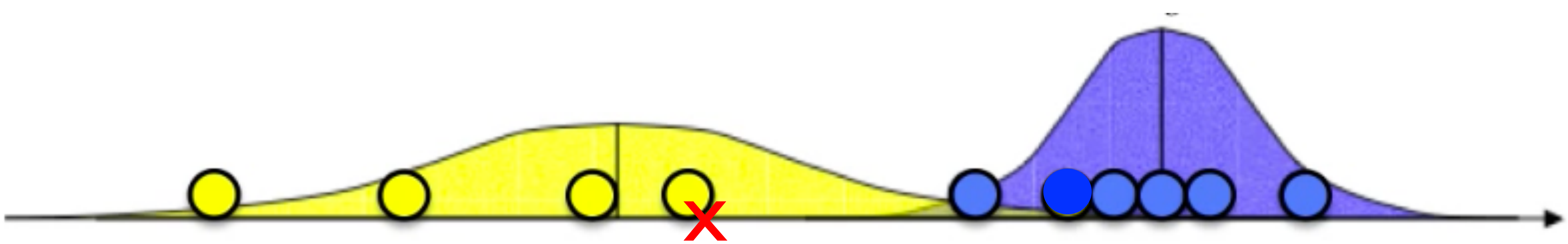


Observations $x_1 \ldots x_n$

What if we know the source of each observation?

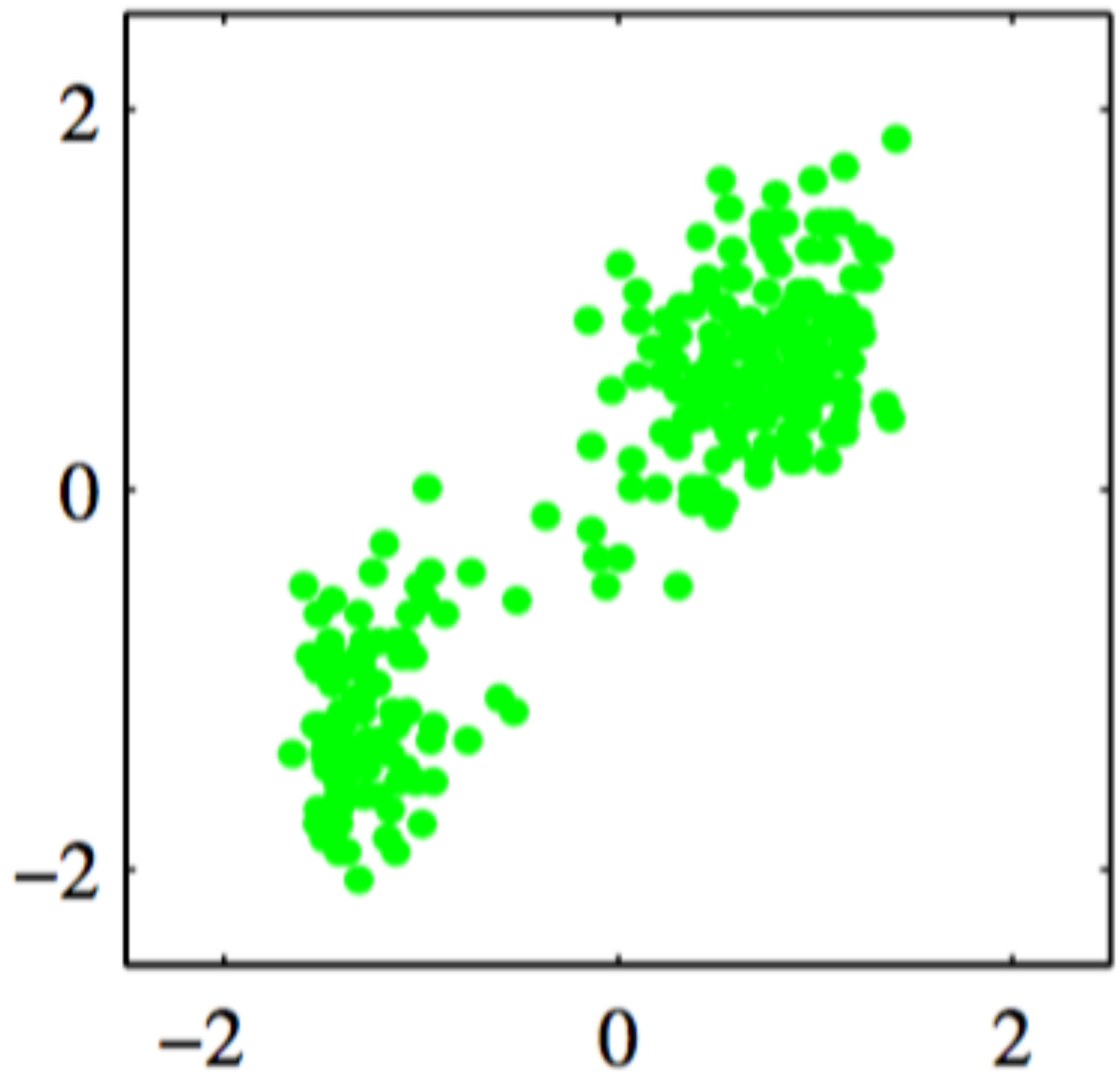
What if we don't know the source?

Who am I?

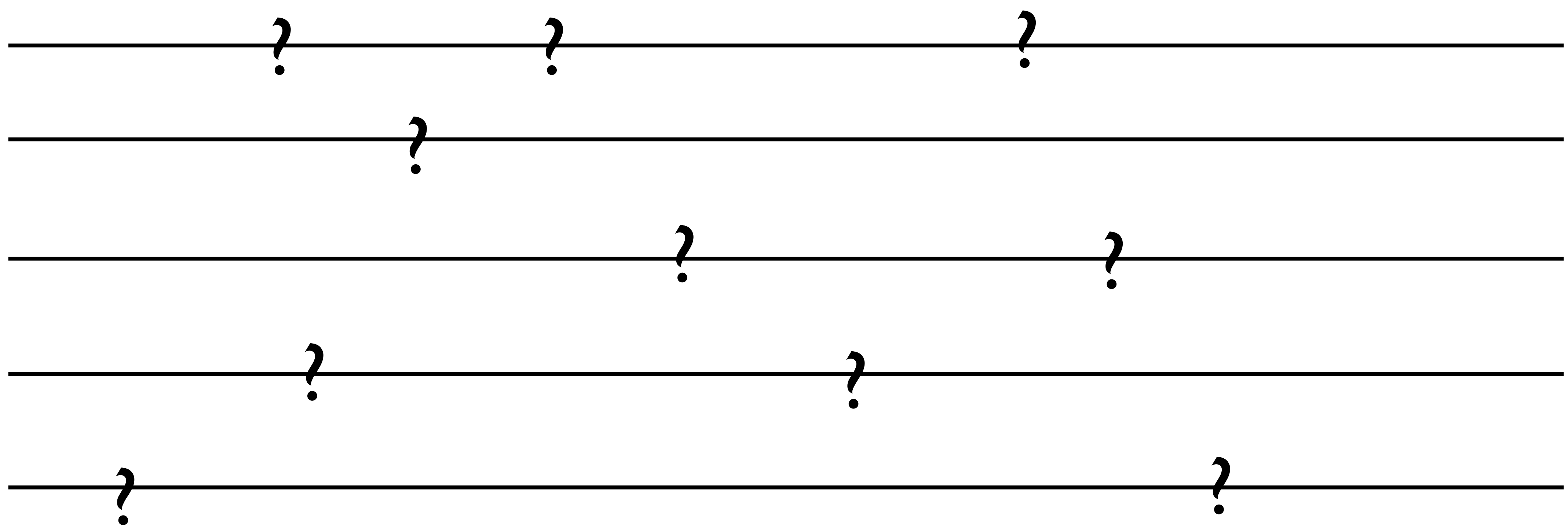


What if we know the model?

Who am I?



Who am I?



Maximum likelihood estimation (MLE)

H T T T H H T H T H

H H H H T H H H H H

H T H H H H H T H H

H T H T T T H H T T

T H H H T H H H T H

? H H H H T T H T H T

Maximum likelihood estimation (MLE)

	<H, T>
H T T T H H T H T H	<5, 5>
H H H H T H H H H H	<9, 1>
H T H H H H H T H H	<8, 2>
H T H T T T H H T T	<4, 6>
T H H H T H H H T H	<7, 3>

? H H H H T T H T H T

<6, 4>

Maximum likelihood estimation (MLE)

		<H, T>
B	H T T T H H T H T H	<5, 5>
A	H H H H T H H H H H	<9, 1>
A	H T H H H H H T H H	<8, 2>
B	H T H T T T H H T T	<4, 6>
A	T H H H T H H H T H	<7, 3>

? H H H H T T H T H T

<6, 4>

We need a model with parameter θ

Maximum likelihood estimation (MLE)

<H, T>

H T T T H H T H T H <5, 5>

H H H H T H H H H H <9, 1>

H T H H H H H T H H <8, 2>

H T H T T T H H T T <4, 6>

T H H H T H H H T H <7, 3>

? H H H H T T H T H T

<6, 4>

We need a model with parameter θ

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} P(D|\theta)$$

Maximum likelihood estimation (MLE)

		<H, T>
B	H T T T H H T H T H	<5, 5>
A	H H H H T H H H H H	<9, 1>
A	H T H H H H H T H H	<8, 2>
B	H T H T T T H H T T	<4, 6>
A	T H H H T H H H T H	<7, 3>

? H H H H T T H T H T

<6, 4>

We need a model with parameter θ

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} P(D|\theta)$$

$$P(D|\theta) = \theta^H (1-\theta)^T$$

θ = prob. of heads

We need a model for each class

θ_A = prob. of heads in coin type A

θ_B = prob. of heads in coin type B

Maximum likelihood estimation (MLE)

	Coin A	Coin B
H T T T H H T H T H		5 H, 5 T
H H H H T H H H H H	9 H, 1 T	
H T H H H H H T H H	8 H, 2 T	
H T H T T T H H T T		4 H, 6 T
T H H H T H H H T H	7 H, 3 T	
	24 H, 6 T	9 H, 11 T

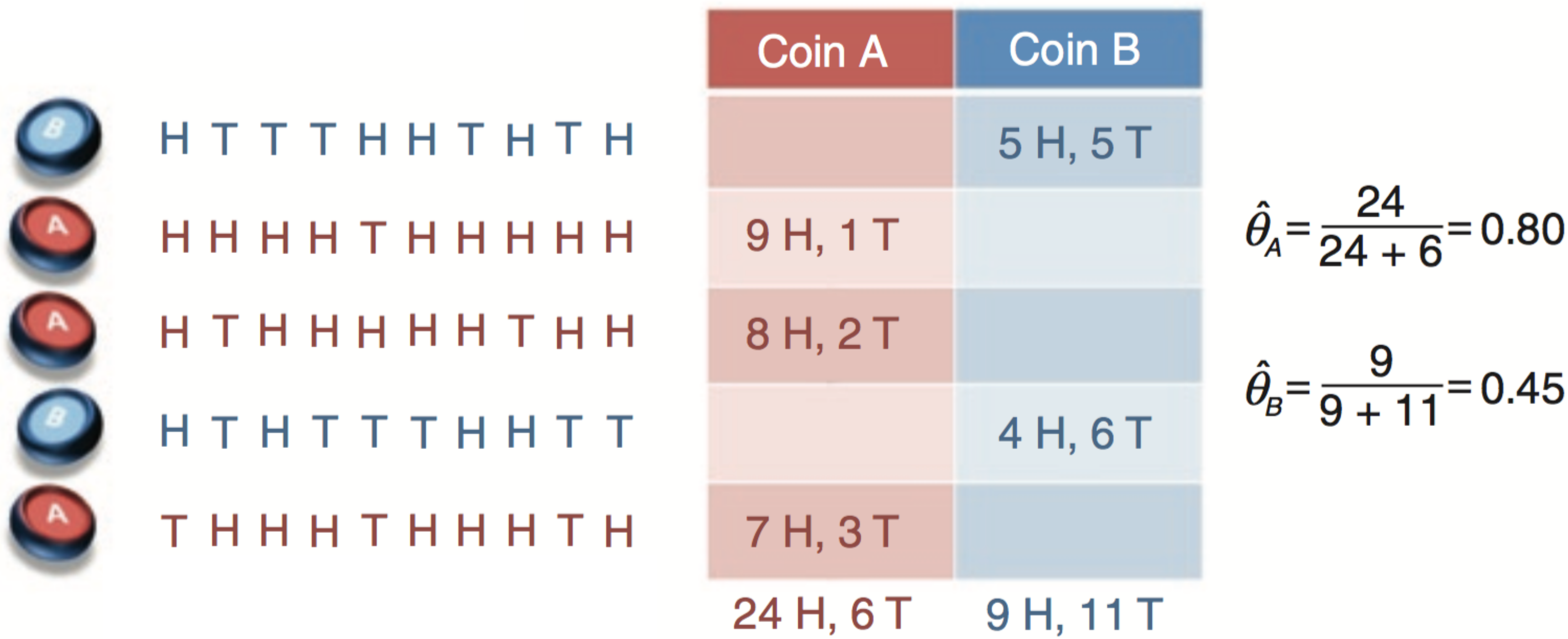
Maximum likelihood estimation (MLE)

	Coin A	Coin B
H T T T H H T H T H		5 H, 5 T
H H H H T H H H H H	9 H, 1 T	
H T H H H H H T H H	8 H, 2 T	
H T H T T T H H T T		4 H, 6 T
T H H H T H H H T H	7 H, 3 T	
	24 H, 6 T	9 H, 11 T

$$\hat{\theta}_A = \frac{24}{24+6} = 0.80$$

$$\hat{\theta}_B = \frac{9}{9+11} = 0.45$$

Maximum likelihood estimation (MLE)



$$\hat{y} = \hat{f}(\mathbf{x}) = \operatorname*{argmax}_{c=1}^{C} p(y = c|\mathbf{x}, \mathcal{D})$$

Expectation-Maximization (EM) vs. MLE

? H T T T H H T H T H

? H H H H T H H H H H

? H T H H H H H T H H

? H T H T T T H H T T

? T H H H T H H H T H

$\hat{\theta}_A = \ ?$

$\hat{\theta}_B = \ ?$

? H H H H T T H T H T

Expectation-Maximization (EM) vs. MLE

? H T T T H H T H T H

? H H H H T H H H H H

? H T H H H H H T H H

? H T H T T T H H T T

? T H H H T H H H T H

$\hat{\theta}_A =$?

$\hat{\theta}_B =$?

? H H H H T T H T H T

→ need to estimate hidden (latent, unobserved) variables and parameters

Expectation-Maximization (EM)

EM is a procedure for learning hidden variables from partially observed data

X: observed variable

Z: hidden variable

θ : parameters for model

assign arbitrary values for parameters θ

iterate until convergence

E step: estimate the values of hidden variable Z by using θ and X

$$Z = \text{argmax } P(Z \mid X, \theta)$$

M step: obtain more accurate parameters θ using observed variable X and estimated Z

(use MLE for parameters)

$$\theta = \text{argmax } P(D \mid \theta_k)$$

EM: coin example

? H T T T H H T H T H

? H H H H T H H H H H

? H T H H H H H T H H

? H T H T T T H H T T

? T H H H T H H H T H

$\hat{\theta}_A = ?$

$\hat{\theta}_B = ?$

X = $\{x^1, x^2, x^3, x^4, x^5\}$ is the number of heads observed,

where $x^i \in \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$

For example, $x^1 = 5$, $x^2 = 9$, $x^3 = 8$, $x^4 = 4$, $x^5 = 7$

Z = $\{z^1, z^2, z^3, z^4, z^5\}$ is the type of coin, where $z^i \in \{A, B\}$,

θ is the probability of heads

$$\hat{\theta}_A = \frac{\text{\# of heads using coin A}}{\text{total \# of flips using coin A}}$$

EM: coin example

1	H	T	T	T	H	H	T	H	T	H
2	H	H	H	H	H	T	H	H	H	H
3	H	T	H	H	H	H	H	T	H	H
4	H	T	H	T	T	T	H	H	T	T
5	T	H	H	H	T	H	H	H	T	H

Is the first toss from A or B? z^{I} =A or B when x^{I} = 5?

→ Is the first toss more likely from the distribution of A or B?

→ $P(z^{I} = A \mid x^{I}) > P(z^{I} = B \mid x^{I})$?

EM: coin example

1	H	T	T	T	H	H	T	H	T	H
2	H	H	H	H	H	T	H	H	H	H
3	H	T	H	H	H	H	H	T	H	H
4	H	T	H	T	T	T	H	H	T	T
5	T	H	H	H	T	H	H	H	T	H

$\theta_A = 0.6$, $\theta_B = 0.5$ (when parameters are given initially)

calculate the likelihood for $P(z^i = A| d^i)$ by using $P(d^i \mid \theta_A)$ and $P(d^i \mid \theta_B)$

→whether coin A or B is more likely to generate the given result from tossing

EM: coin example

1	H	T	T	T	H	H	T	H	T	H
2	H	H	H	H	H	T	H	H	H	H
3	H	T	H	H	H	H	H	T	H	H
4	H	T	H	T	T	T	H	H	T	T
5	T	H	H	H	T	H	H	H	T	H

$\theta_A = 0.6$, $\theta_B = 0.5$ (when parameters are given initially)

calculate the likelihood for $P(z^i = A | d^i)$ by using $P(d^i | \theta_A)$ and $P(d^i | \theta_B)$

→whether coin A or B is more likely to generate the given result from tossing

$$P(z^1 = A \mid d^1) \approx \frac{P(d^1 \mid \theta_A)}{P(d^1 \mid \theta_A) + P(d^1 \mid \theta_B)}$$

$P(d) = nCk\ \theta^k (1-\theta)^{n-k}$

k is the number of heads-up

θ is the probability of heads-up

$P(d_1 \mid \theta_A) = {}_{10}C_5\ 0.6^5\ 0.4^5$

$P(d_1 \mid \theta_B) = {}_{10}C_5\ 0.5^5\ 0.5^5$

$P(z^1 = A \mid d_1) = 0.45$

$P(z^1 = B \mid d_1) = 0.55$

EM: coin example

1	H	T	T	T	H	H	T	H	T	H
2	H	H	H	H	H	T	H	H	H	H
3	H	T	H	H	H	H	H	T	H	H
4	H	T	H	T	T	T	H	H	T	T
5	T	H	H	H	T	H	H	H	T	H

randomly assigned for the first iteration

$\theta_A^{(0)} = 0.6, \quad \theta_B^{(0)} = 0.5$

$P_A = P(z^i = A \mid d^i)$

d	**X**	P_A	P_B	**Z**
1	5	0.45	0.55	B
2	9	0.80	0.20	A
3	8	0.73	0.27	A
4	4	0.35	0.65	B
5	7	0.65	0.35	A

x is the number of heads

z is the type of coin

E-step: assign the expected values to the hidden variable based on the given model

EM: coin example

randomly assigned for the first iteration

$\theta_A^{(0)} = 0.6, \quad \theta_B^{(0)} = 0.5$

	X	P_A	P_B	**Z**
1	5	0.45	0.55	B
2	9	0.80	0.20	A
3	8	0.73	0.27	A
4	4	0.35	0.65	B
5	7	0.65	0.35	A

x is the number of heads

z is the type of coin

	A	B
1		5H5T
2	9H1T	
3	8H2T	
4		4H6T
5	7H3T	

$\theta_A^{(1)} = 24 / (24+6) = 0.8$

$\theta_B^{(1)} = 9 / (9+11) = 0.45$

E-step: assign the expected values to the hidden variable based on the given model

M-step: update the parameters that maximize the probability

EM: coin example

$\theta_A^{(1)} = 0.8, \quad \theta_B^{(1)} = 0.45$

	X	A	B	Z
1	5	0.1	0.9	B
2	9			
3	8			
4	4			
5	7			

$$P(d_1 \mid \theta_A^{(1)}) = {}_{10}C_5 \; 0.8^5 \, 0.2^5 = 0.026$$

$$P(d_1 \mid \theta_B^{(1)}) = {}_{10}C_5 \; 0.45^5 \, 0.55^5 = 0.234$$

$$P(z^1 = A \mid d_1) = \frac{P(d_1 \mid \theta_A^{(1)})}{P(d_1 \mid \theta_A^{(1)}) + P(d_1 \mid \theta_B^{(1)})} = 0.1$$

E-step: assign the expected values to the hidden variable

M-step: update the parameters that maximize the probability

EM: coin example

$\theta_A^{(1)} = 0.8, \quad \theta_B^{(1)} = 0.45$

	X	A	B	**Z**
1	5	0.1	0.9	B
2	9	0.98	0.02	A
3	8			
4	4			
5	7			

$$P(d_2 \mid \theta_A^{(1)}) = {}_{10}C_9 \;\; 0.8^9\, 0.2^1 = 0.268$$

$$P(d_2 \mid \theta_B^{(1)}) = {}_{10}C_9 \;\; 0.45^9\, 0.55^1 = 0.004$$

$$P(z^1 = A \mid d_2) = \frac{P(d_2 \mid \theta_A^{(1)})}{P(d_2 \mid \theta_A^{(1)}) + P(d_2 \mid \theta_B^{(1)})} = 0.98$$

$$P(d_1 \mid \theta_A^{(1)}) = {}_{10}C_5 \;\; 0.8^5\, 0.2^5 = 0.026$$

$$P(d_1 \mid \theta_B^{(1)}) = {}_{10}C_5 \;\; 0.45^5\, 0.55^5 = 0.234$$

$$P(z^1 = A \mid d_1) = \frac{P(d_1 \mid \theta_A^{(1)})}{P(d_1 \mid \theta_A^{(1)}) + P(d_1 \mid \theta_B^{(1)})} = 0.1$$

E-step: assign the expected values to the hidden variable

M-step: update the parameters that maximize the probability

EM: coin example

$\theta_A^{(0)} = 0.6, \quad \theta_B^{(0)} = 0.5$

$\theta_A^{(1)} = 0.8, \quad \theta_B^{(1)} = 0.45$

	X	A	B	**Z**
1	5	0.1	0.9	B
2	9	0.98	0.02	A
3	8			A
4	4			A
5	7			A

	A	B
1		5H5T
2	9H1T	
3	8H2T	
4	4H6T	
5	7H3T	

$P(d_1 \mid \theta_A^{(1)}) = {}_{10}C_5 \;\; 0.8^5 \, 0.2^5 = 0.026$

$P(d_1 \mid \theta_B^{(1)}) = {}_{10}C_5 \;\; 0.45^5 \, 0.55^5 = 0.234$

$$P(z^1 = A \mid d_1) = \frac{P(d_1 \mid \theta_A^{(1)})}{P(d_1 \mid \theta_A^{(1)}) + P(d_1 \mid \theta_B^{(1)})} = 0.1$$

$\theta_A^{(2)} = 28 / (28+12) = 0.7$

$\theta_B^{(2)} = 5 / (5+5) = 0.5$

E-step: assign the expected values to the hidden variable

M-step: update the parameters that maximize the probability

Expectation-Maximization (EM)

EM is a procedure for learning hidden variables from partially observed data

X: observed variable

Z: hidden variable

θ : parameters for model

assign arbitrary values for parameters θ

iterate until convergence

E step: estimate the values of hidden variable Z by using θ and X

$$Z = \text{argmax } P(Z \mid X, \theta)$$

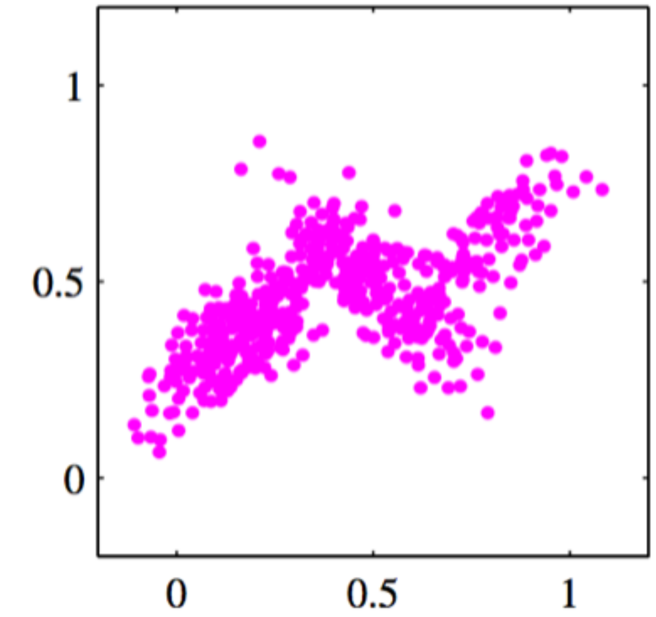
M step: obtain more accurate parameters θ using observed variable X and estimated Z

(use MLE for parameters)

$$\theta = \text{argmax } P(D \mid \theta_k)$$

Types of assignments

- hard clustering
    - clusters do not overlap
    - element either belongs to a specific cluster or not
- soft clustering
    - clusters may overlap
    - the degree of association between clusters and instances



EM: coin example for soft assignment

Z
B
A
A
B
A

	A	B
1		5H5T
2	9H1T	
3	8H2T	
4		4H6T
5	7H3T	

randomly assigned for the first iteration

$\theta_A^{(0)} = 0.6, \quad \theta_B^{(0)} = 0.5$

	X	P_A	P_B	**Z**
1	5	0.45	0.55	
2	9	0.80	0.20	
3	8	0.73	0.27	
4	4	0.35	0.65	
5	7	0.65	0.35	

x is the number of heads

z is the type of coin

	A	B	
1	2.2H 2.2T	2.8H 2.8H	5H5T
2	7.2H 0.8T	1.8H 0.2T	9H1T
3	5.9H 1.5T	2.1H 0.5T	8H2T
4	1.4H 2.1H	2.6H 3.9T	4H6T
5	4.5H 1.9T	2.5H 1.1T	7H3T

$\theta_A^{(1)} = 21.3 / (21.3+8.6) = 0.71$

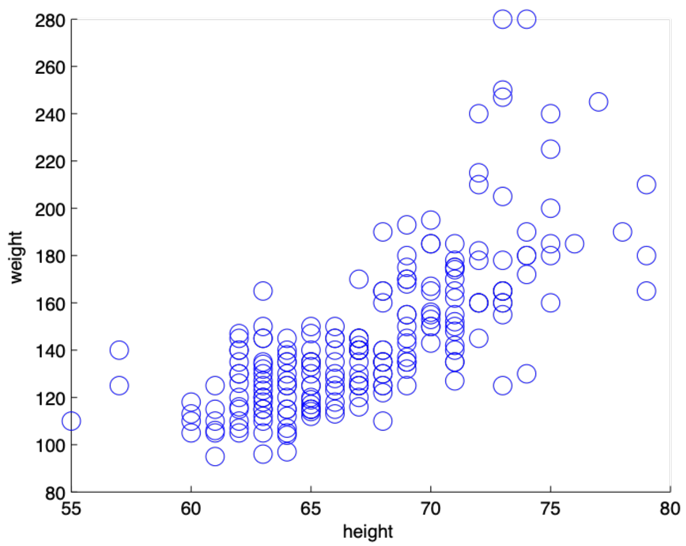
$\theta_B^{(1)} = 11.7 / (11.7+8.4) = 0.58$

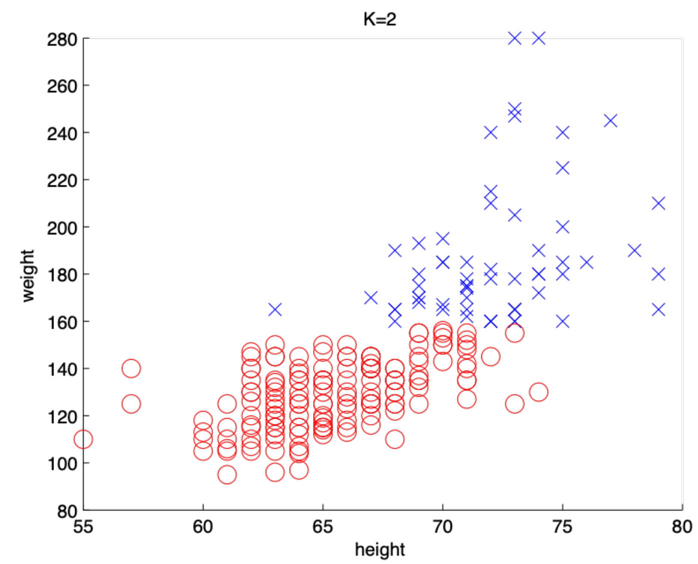
E-step: assign the expected values to the hidden variable based on the given model

M-step: update the parameters that maximize the probability

Unsupervised learning

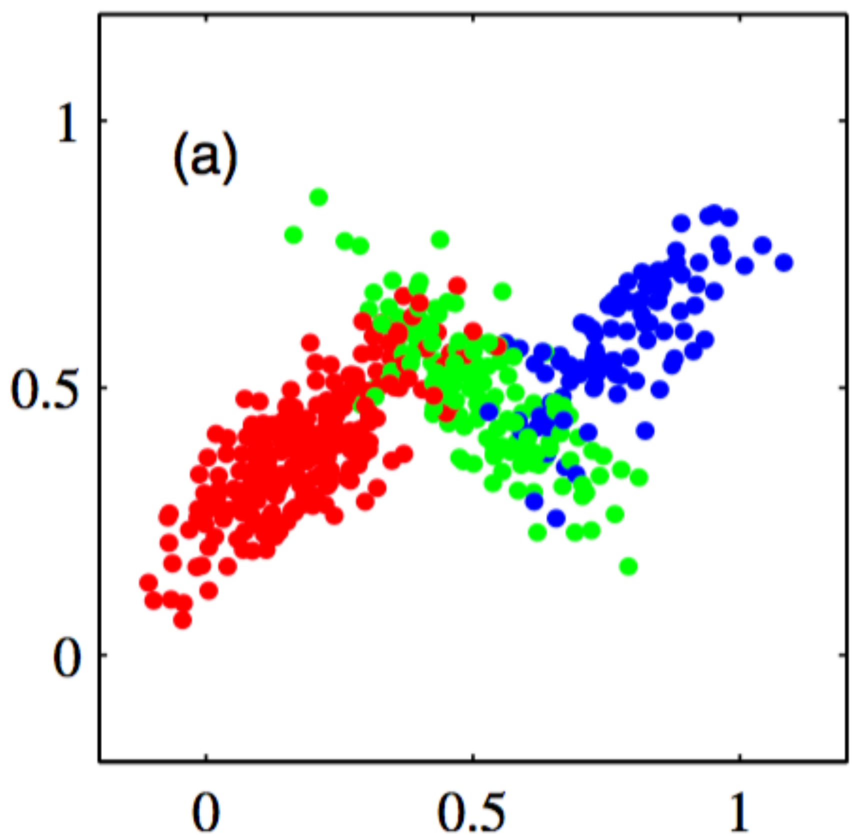
- Discovering clusters

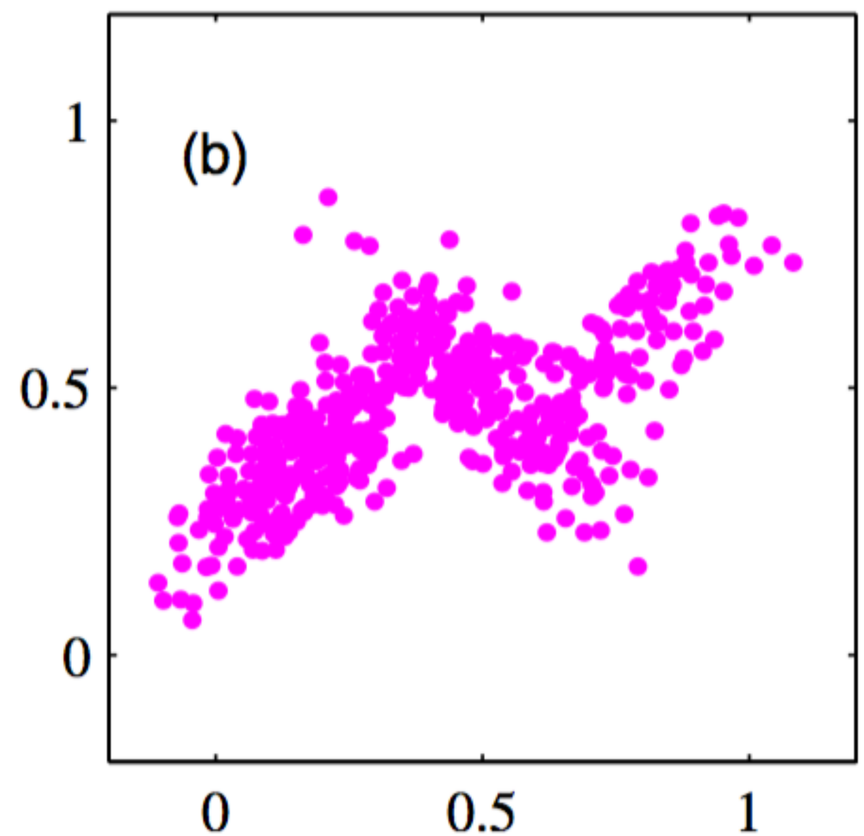




$$z_i^* = \mathrm{argmax}_k \, p(z_i = k | \mathbf{x}_i, \mathcal{D})$$ Latent variable

Unsupervised learning

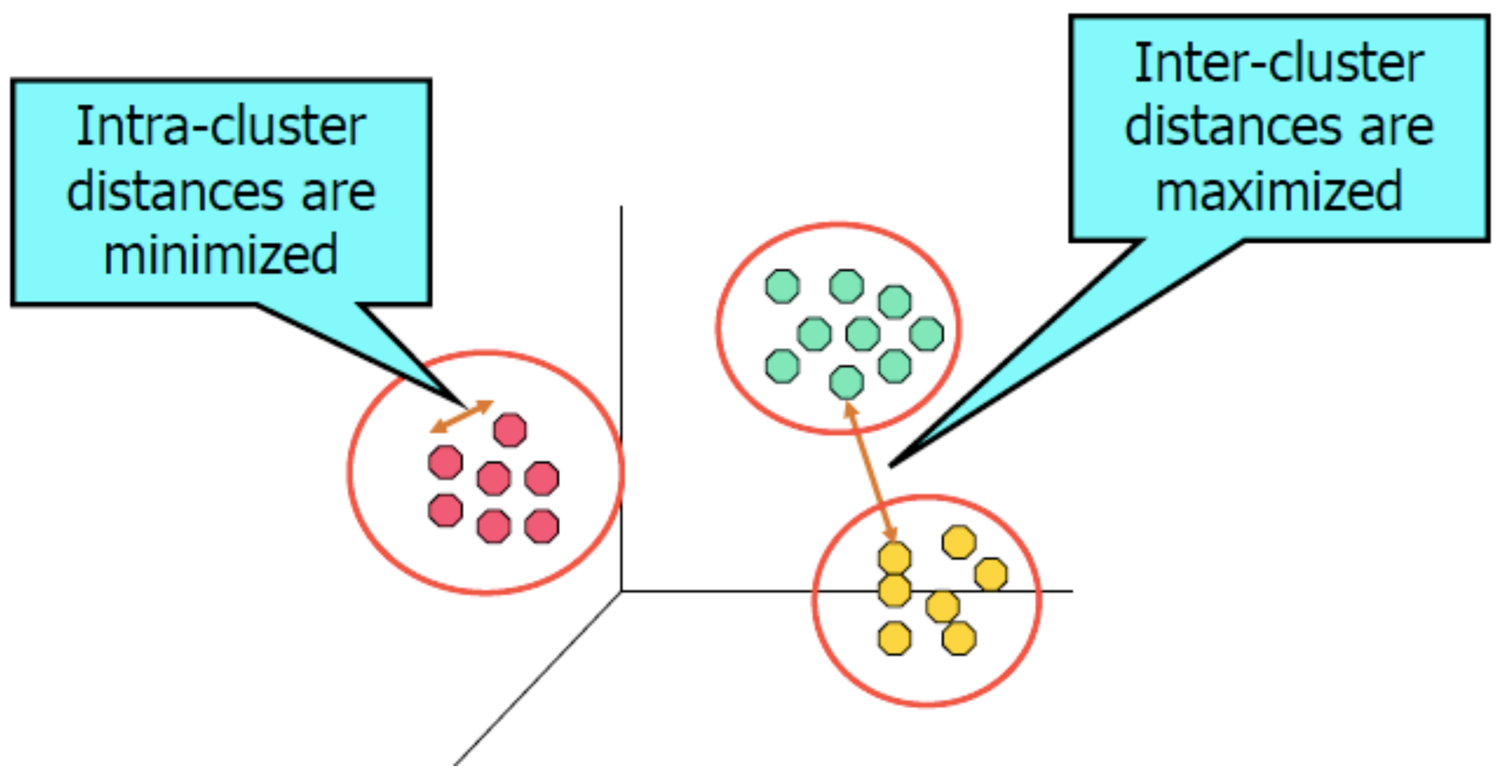




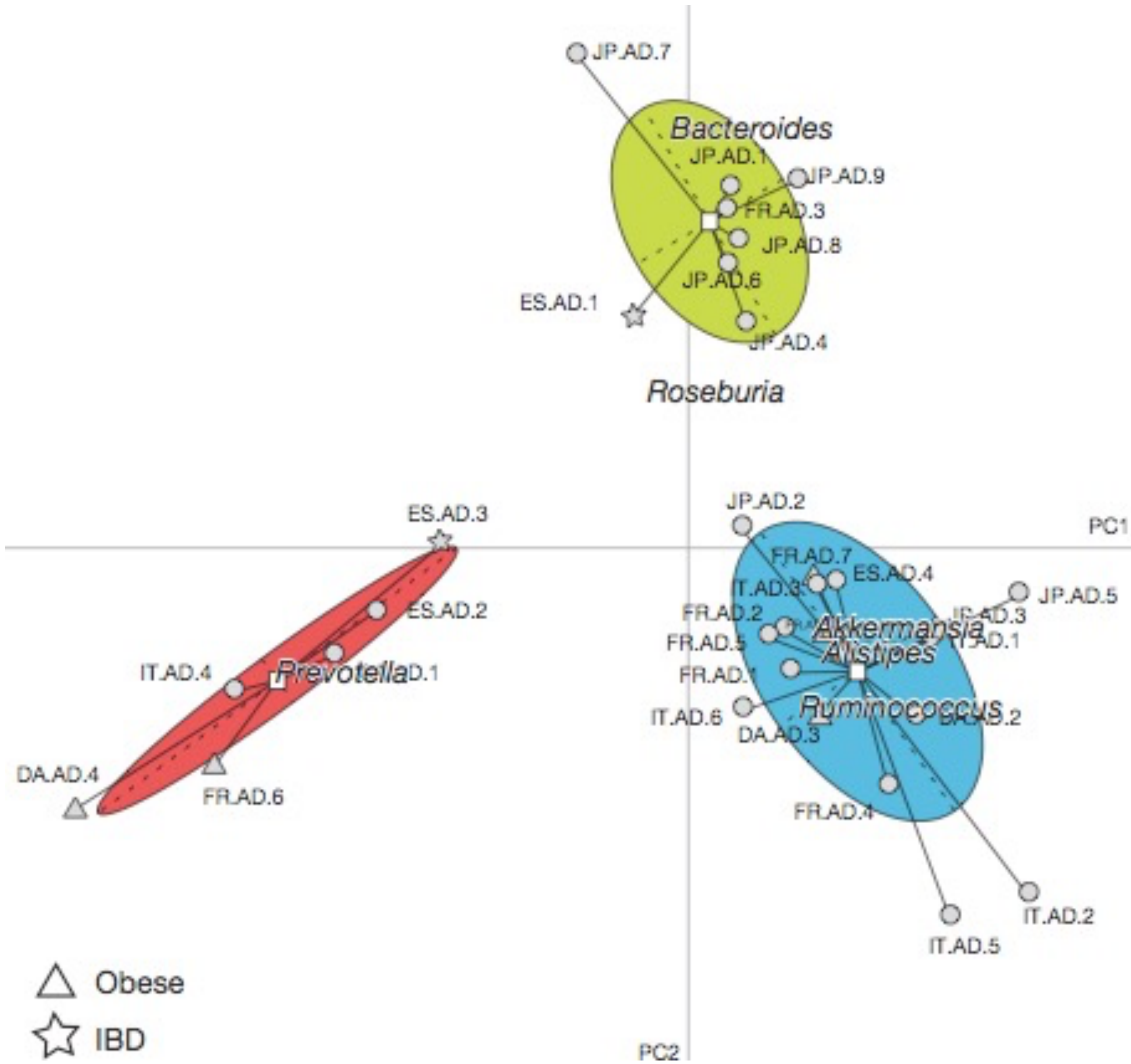
γ

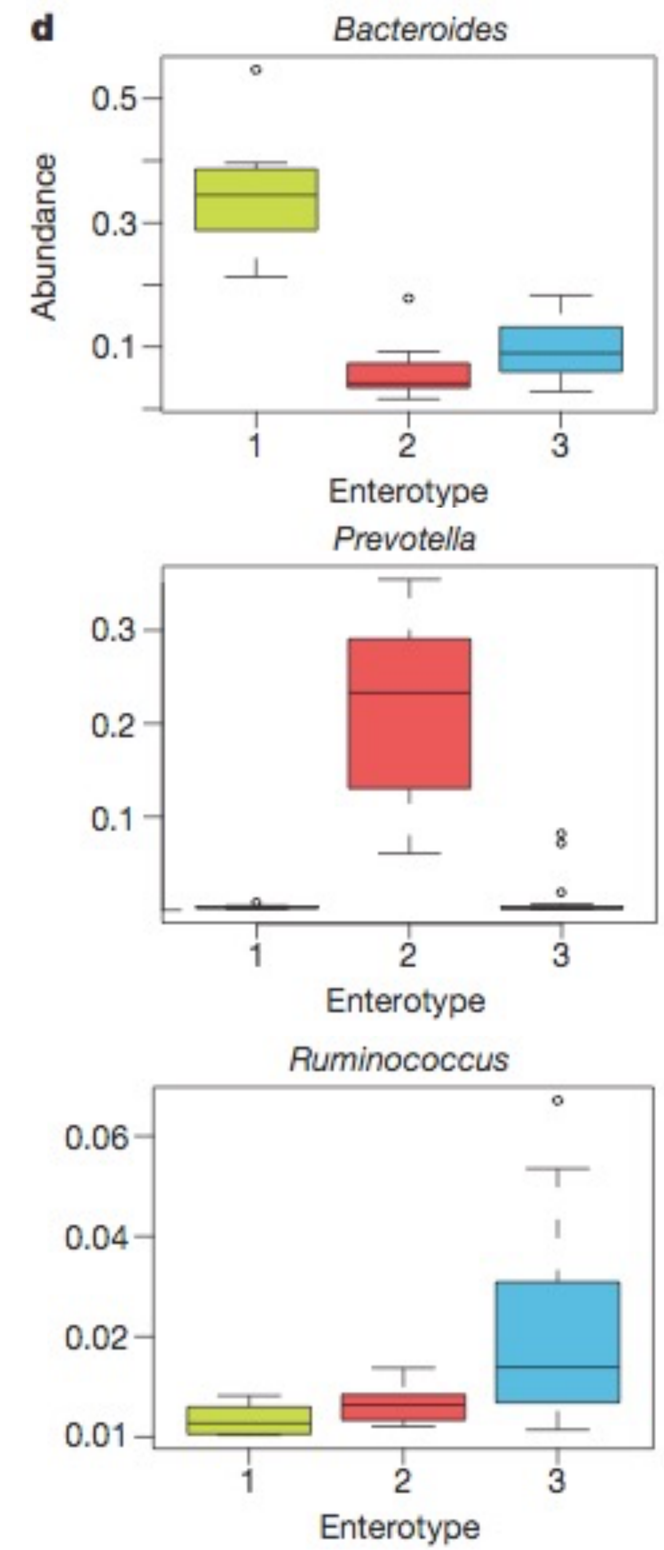
Clustering

- Clustering is a problem of identifying clusters of data points in a multidimensional space
- Considering a cluster as comprising a group of data points whose inter-point distances are small compared with the distance to the points outside of the cluster
- Optimal assignment to the latent cluster



Clustering in biomedical data





	sample1	sample2	sample3	sample4
Bacteria A				
Bacteria B				
Bacteria C				

Arumugam, M. et al. *Nature*, (2011)

K-means clustering

- When given a set of data $\{x^1, x^2, x^3, \ldots, x^N\}$, which is N examples of a D-dimensional variable x, partition the data set into K clusters

  → Finding assignment of examples to clusters $\{r_{nk}\}$ and a set of vectors $\{\mu_k\}$, such that the sum of the squares of the distances of each data point to its closest vector μ_k is minimum

  - μ_k: prototype associated with the k^{th} cluster, which represent the center of the cluster
  - $r_{nk} = 1$ if a data point x^n is assigned to cluster k

    $r_{nj} = 0$ for $j \neq k$

$$(r_{nk}) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}$$

$$\sum_k r_{nk} = 1$$

objective function

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \|\mathbf{x}_n - \boldsymbol{\mu}_k\|^2$$

Review: Expectation-Maximization (EM)

EM is a procedure for learning hidden variables from partially observed data

X: observed variable

Z: hidden variable

θ : parameters for model

assign arbitrary values for parameters θ

iterate until convergence

E step: estimate the values of hidden variable Z by using θ and X

$$Z = \text{argmax } P(Z \mid X, \theta)$$

M step: obtain more accurate parameters θ using observed variable X and estimated Z

calculate MLE of parameters

$$\theta = \text{argmax } P(D \mid \theta_k)$$

K-means clustering

- K-means clustering uses EM approach
  - choose an initial values for μ_k
  - repeat two steps
    - E-step: assign each example to the nearest prototype by minimizing J;

      → determine r_{nk}

$$r_{nk} = \begin{cases} 1 & \text{if } k = \arg\min_j \|\mathbf{x}_n - \boldsymbol{\mu}_j\|^2 \\ 0 & \text{otherwise.} \end{cases}$$

    - M-step: update the prototypes with the data points assigned;

      → determine μ_k with the new r_{nk}

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \|\mathbf{x}_n - \boldsymbol{\mu}_k\|^2$$

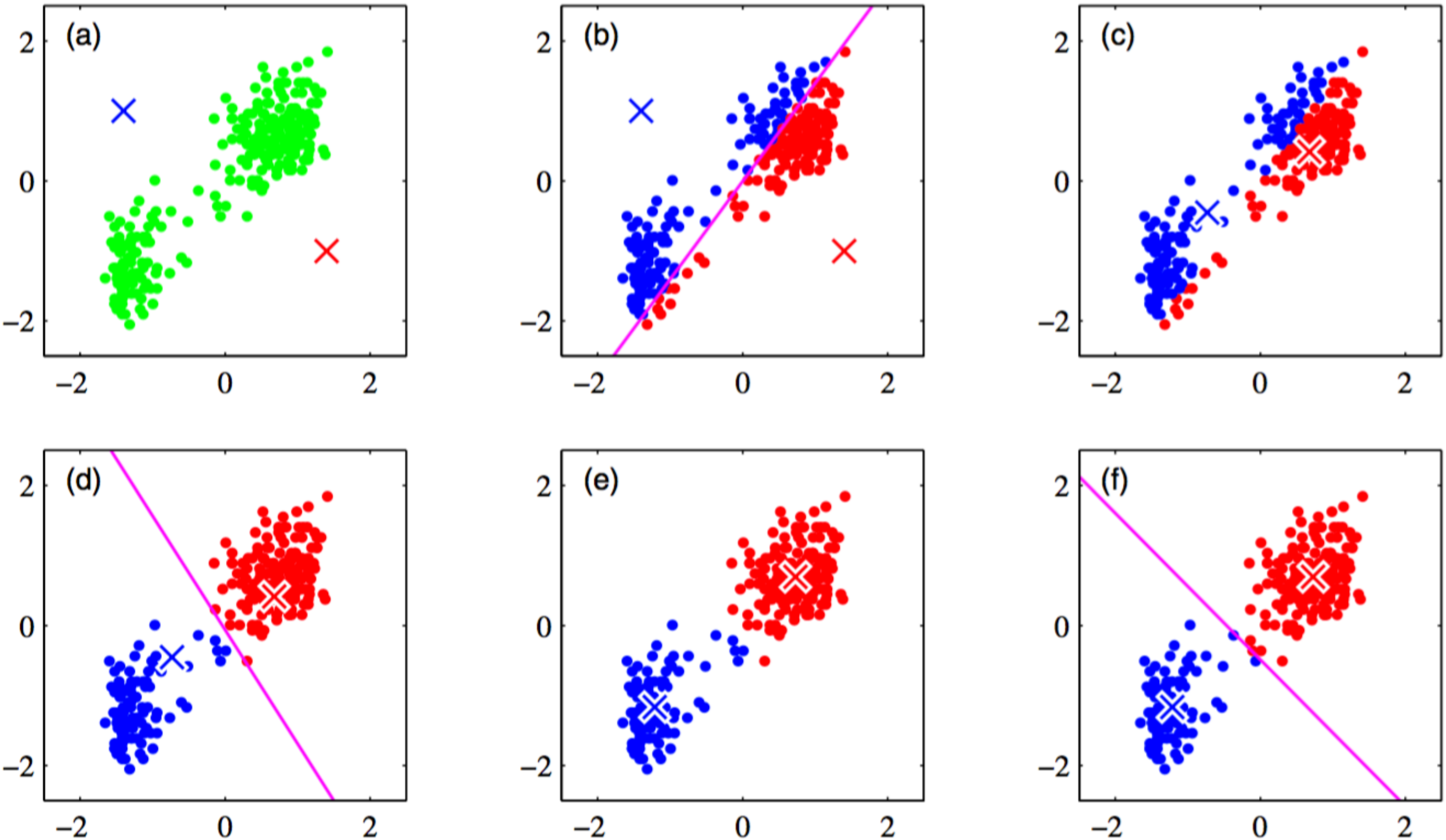
$$2 \sum_{n=1}^{N} r_{nk}(\mathbf{x}_n - \boldsymbol{\mu}_k) = 0$$

For each k,

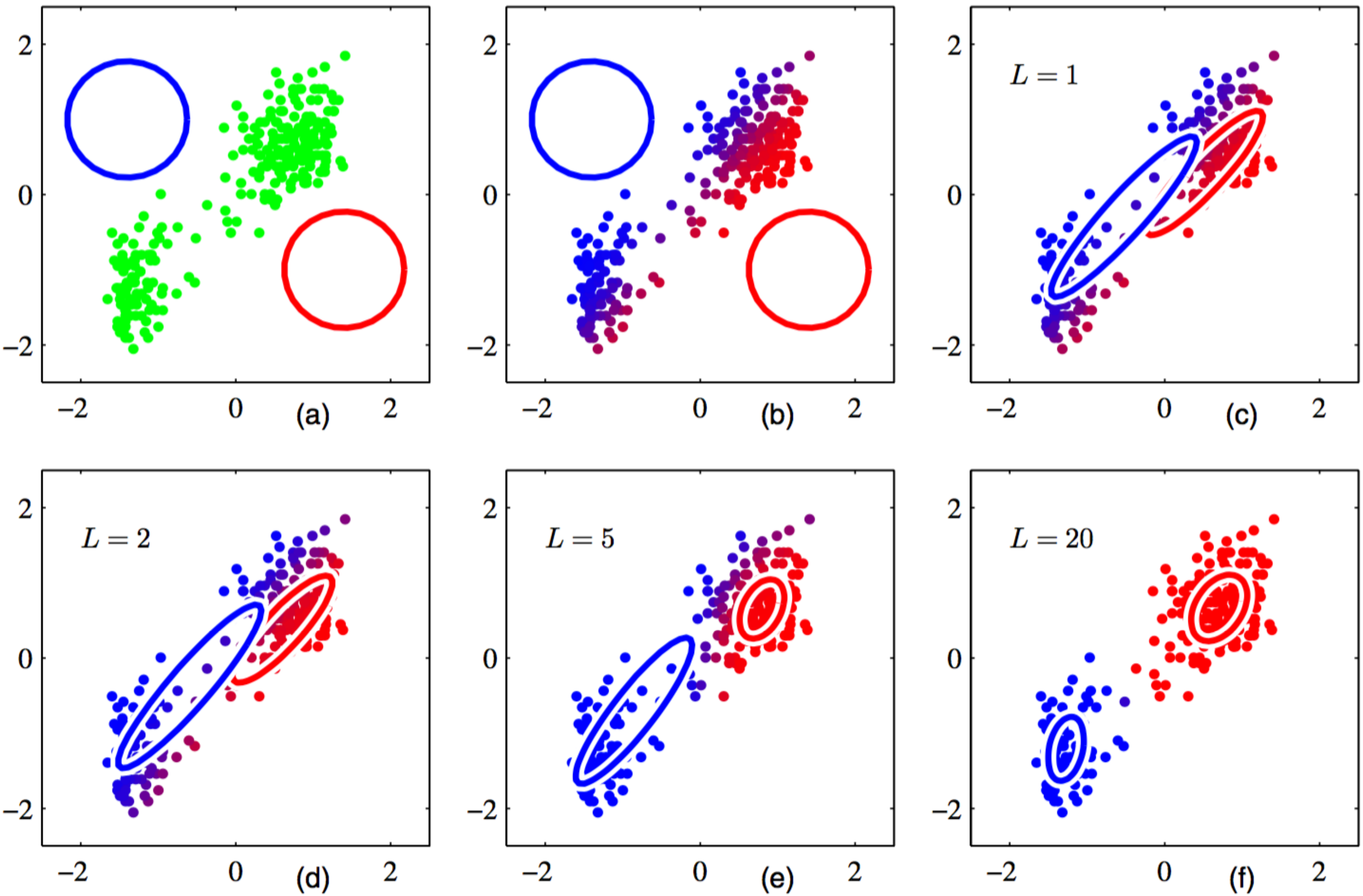
set the derivative of J to 0 with respect to μ_k

$$\boldsymbol{\mu}_k = \frac{\sum_n r_{nk} \mathbf{x}_n}{\sum_n r_{nk}}$$

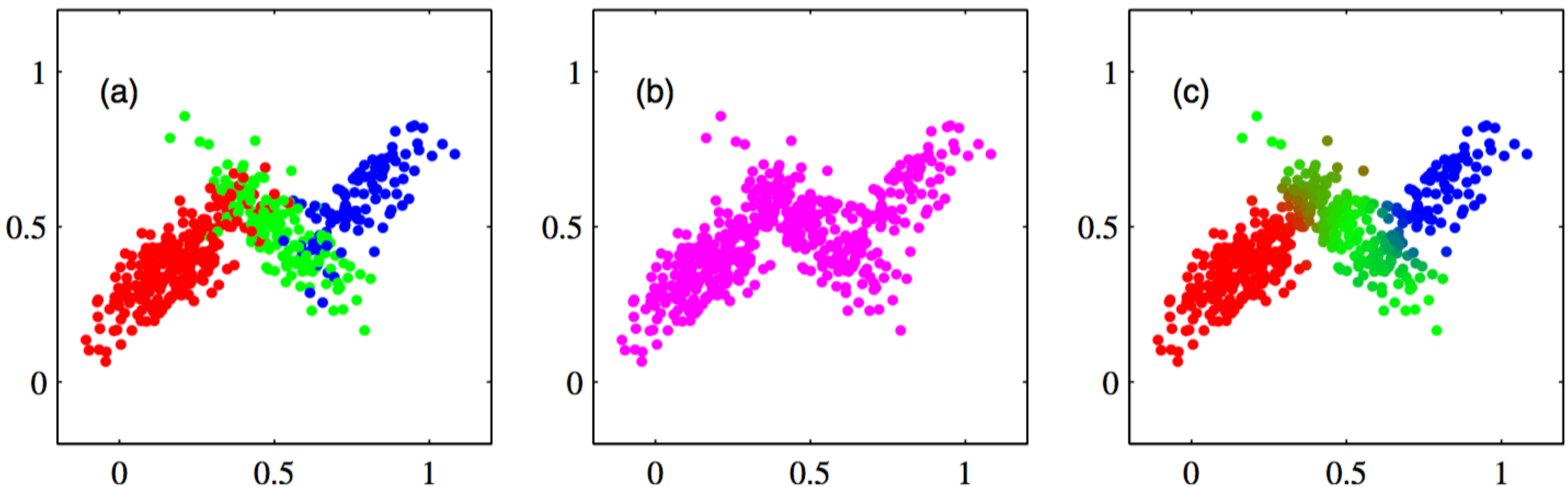
K-means clustering



EM for Gaussian mixture



EM for Gaussian mixture



(a) example of 500 data points drawn from 3 Gaussian models
(b) plotting only x values
(c) the color represent the value of the responsibility $\gamma(z_{nk})$ associated with data point x^n