Expectation-Maximization

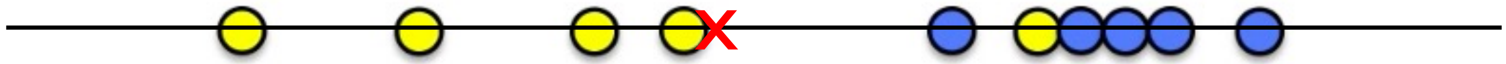
Who am I?

예측, 데이터 이해 → 2가지 방법 → 새 데이터 들어오면 어떻게 해야할까

① 비슷한 데이터끼리 잘 나뉘어 있을때 내 데이터가 어디 속할 것인가

② label이 없을때 데이터로부터 특성 찾아내기

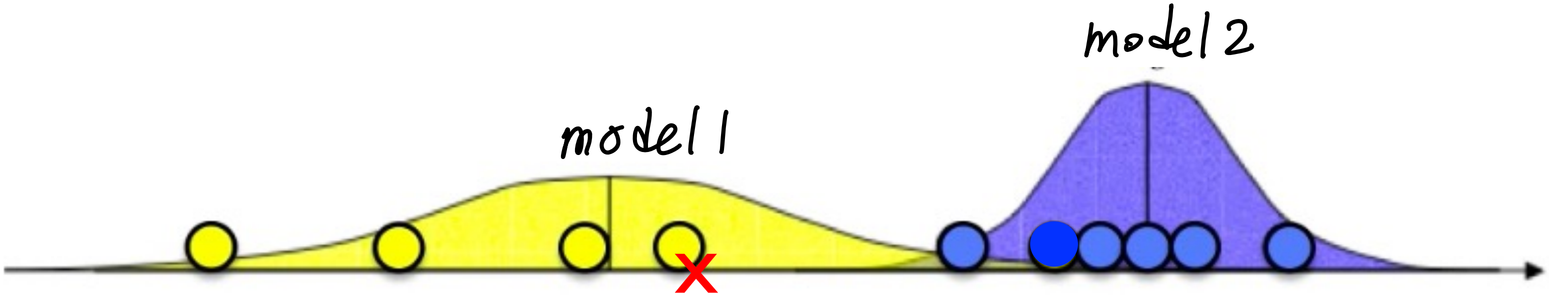
Who am I?



Observations $x_1 \ldots x_n$

What if we know the source of each observation? supervised

Who am I?



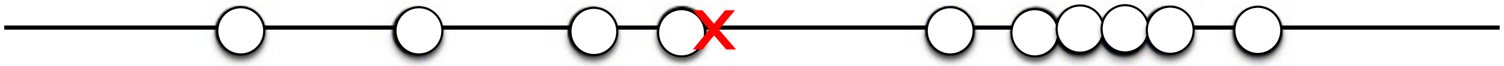
↓ supervised learning

어떤 모델에
속할 확률이 크지

data로 부터 classifier 만들수 있어 새로운 데이터 분류가능

Who am I?

unsupervised manner



Observations $x_1 \ldots x_n$

What if we know the source of each observation?

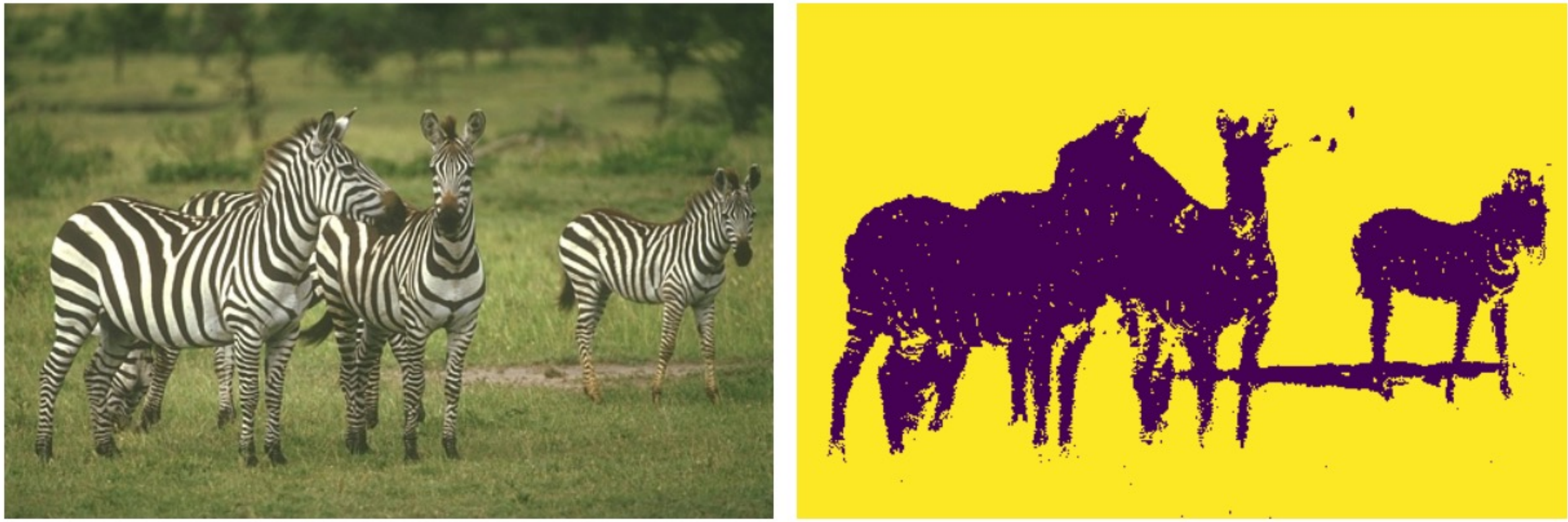
What if we don't know the source of each observation?

Image segmentation

무슨 object가 어떻게 있는지를 구별

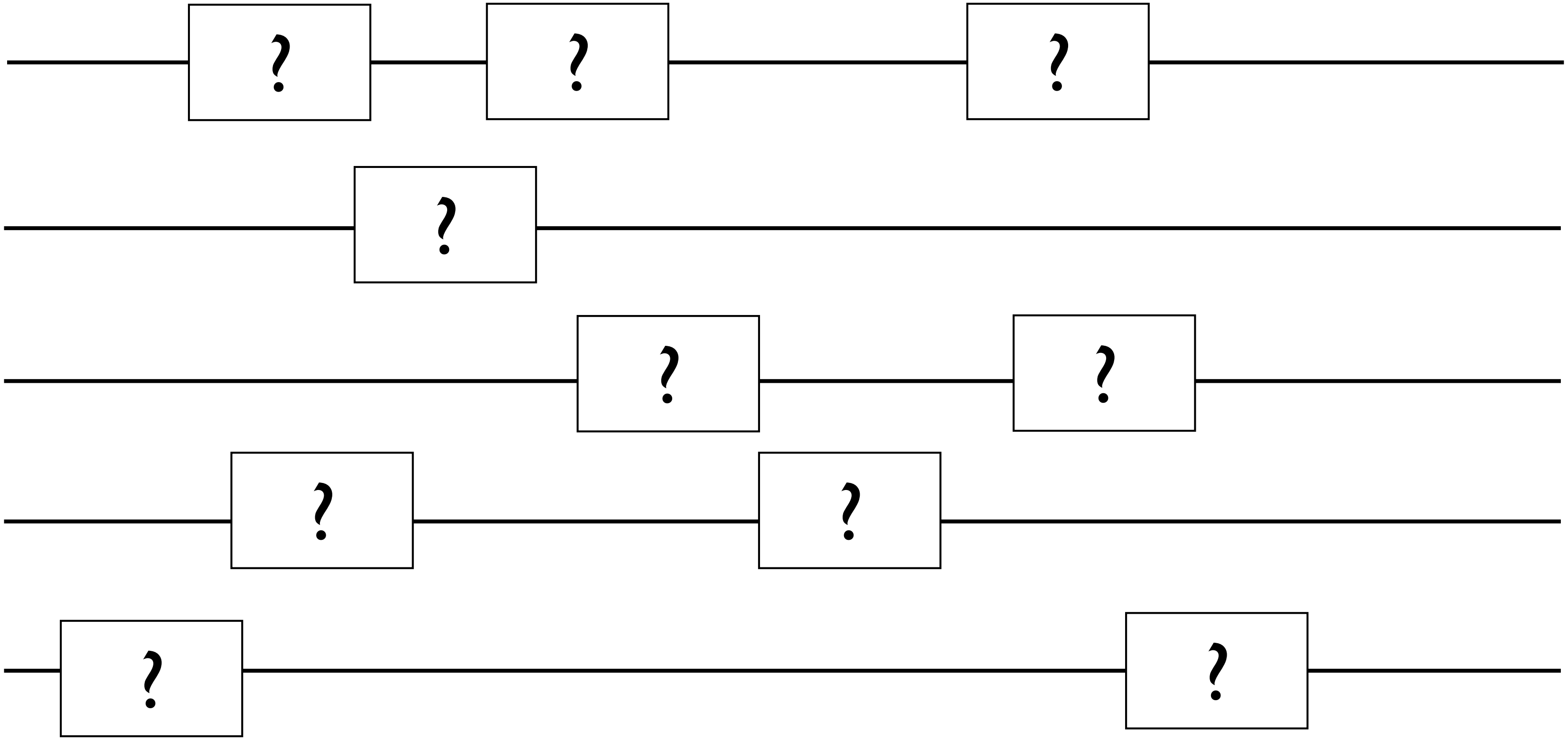
object와 background 구분

→ data에서 clue를 찾아서 나눠야함



Motif finding

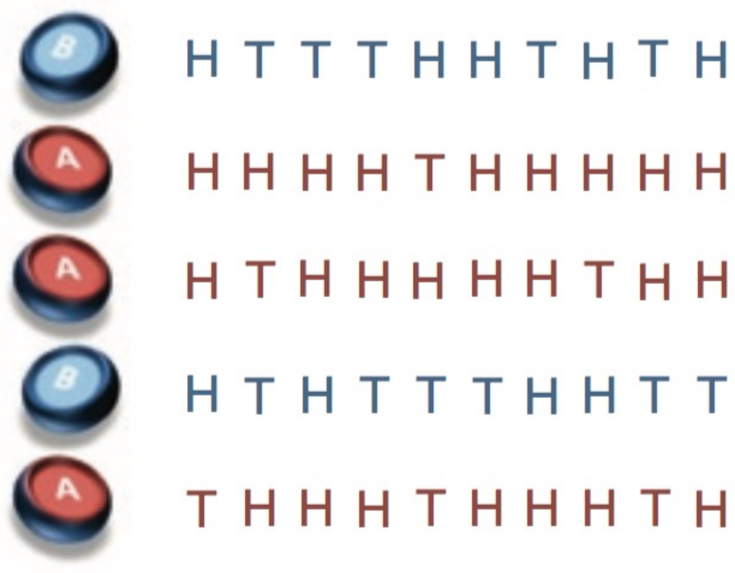
특정 바이러스, 약물에 바인딩 하는데, 뭐에 하는지 어떻게 하는지



Maximum likelihood estimation (MLE)

<동전 던지기>

동전 Type : A, B



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

A, B 각각에 해당하는 모델을 만들면 새로 동전을 던졌을때 A인지 B인지 알수있음

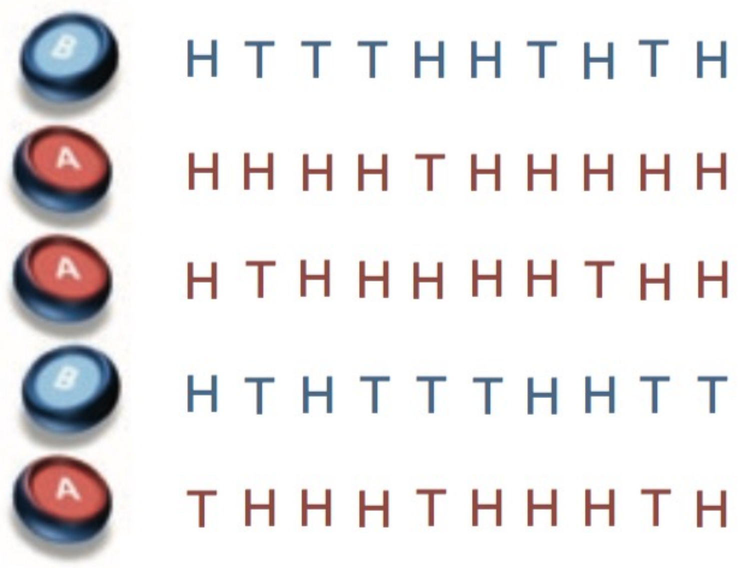
Maximum likelihood estimation (MLE)

observation

<H, T> 동전은 independent

→ 순서 별로 중요X

→ < 앞면이 나올 개수, 뒷면이 나올 개수 > 로 요약됨



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>

model estimate

$\hat{}$

(parameter)

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>

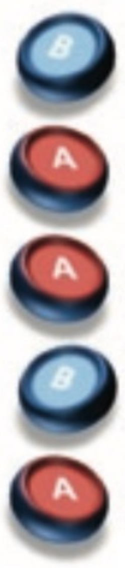
We need a model with parameter θ

data의 probability를 maximize 하는 θ를 찾는것 = model learning → "MLE"라고 함

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

data의 probability를 maximize 할수있는 θ 값을 찾아서 그 모델의 parameter로 주겠다

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>

$P\ (D|\ \theta) = \theta^h\ (1 - \theta)^t$

θ = prob. of heads

h = # of head

t = # of tail

We need a model with parameter θ

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

data의 probability를 최대로 하는 θ

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)

$$\hat{\theta} = argmax\, lnP(D|\theta)$$

$$= argmax\, ln(\theta^{h} (1-\theta)^{t})$$

$$= argmax(\, ln\theta^{h} + ln(1-\theta)^{t})$$

→ max 값 찾기

→ 미분해서 0이 되는 θ값 찾기

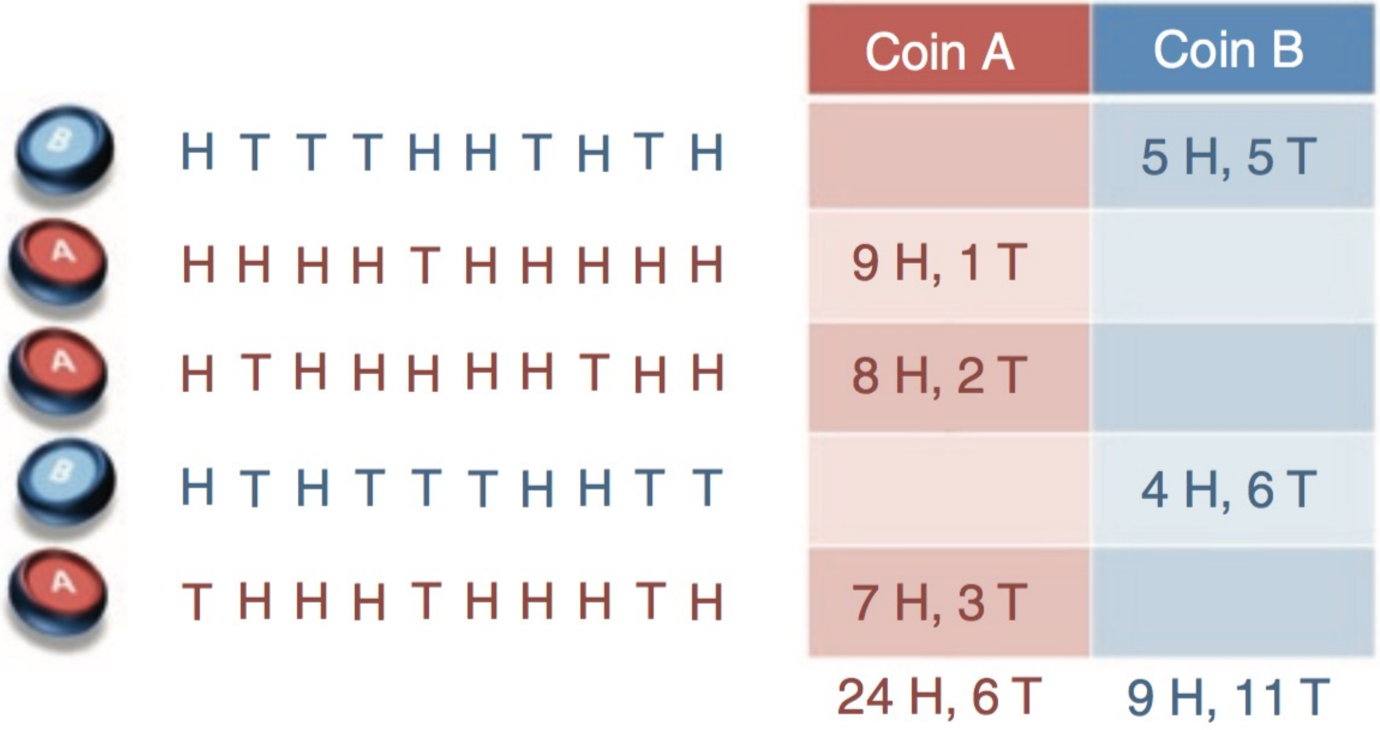
$$\frac{\partial(\, ln\theta^{h} + ln(1-\theta)^{t})}{\partial\theta} = 0$$

←

$$h\frac{1}{\theta} + t\frac{-1}{1-\theta} = 0$$

$$\theta = \frac{h}{h+t} = \frac{\text{앞면 개수}}{\text{전체 시행}}$$ = 앞면이 나올 확률 값

Maximum likelihood estimation (MLE)

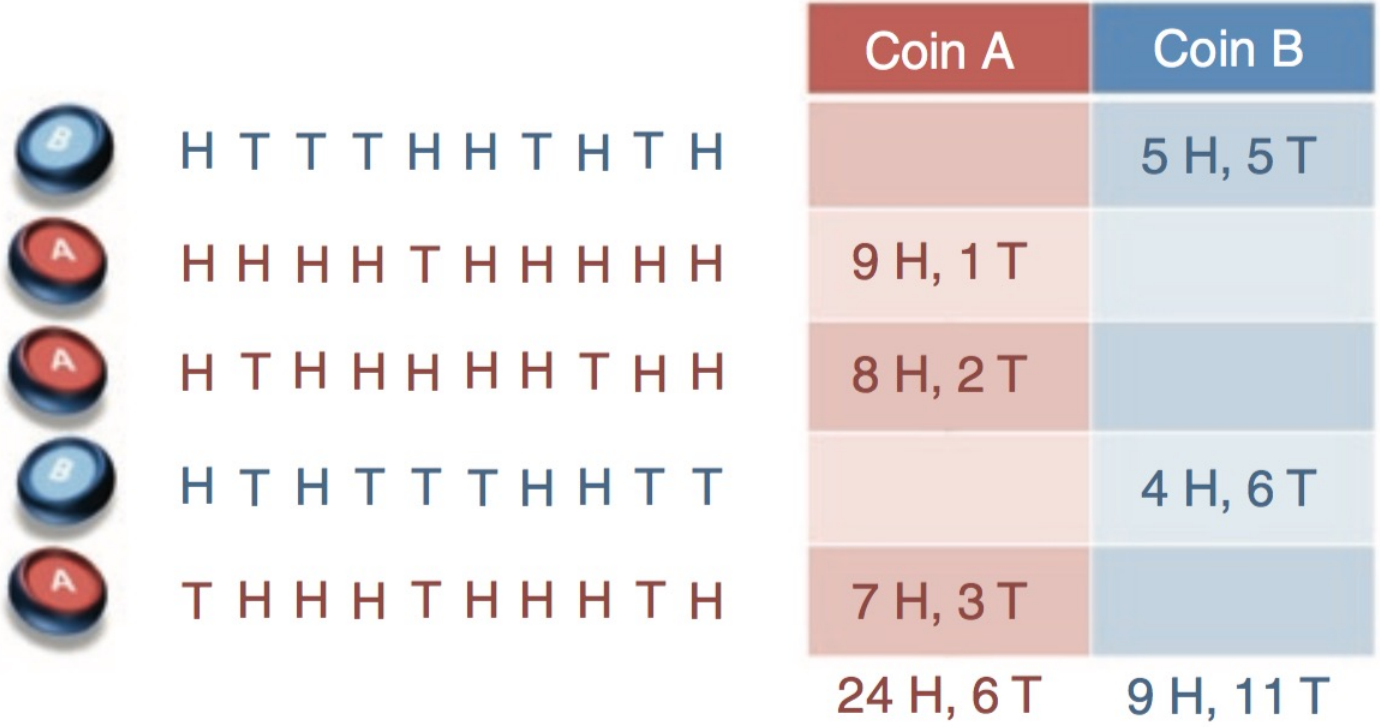


		Coin A	Coin B
B	H T T T H H T H T H		5 H, 5 T
A	H H H H T H H H H H	9 H, 1 T	
A	H T H H H H H T H H	8 H, 2 T	
B	H T H T T T H H T T		4 H, 6 T
A	T H H H T H H H T H	7 H, 3 T	
		24 H, 6 T	9 H, 11 T

Maximum likelihood estimation (MLE)

estimation

$\theta = \frac{h}{h+t}$



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

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

앞면이 나올 확률

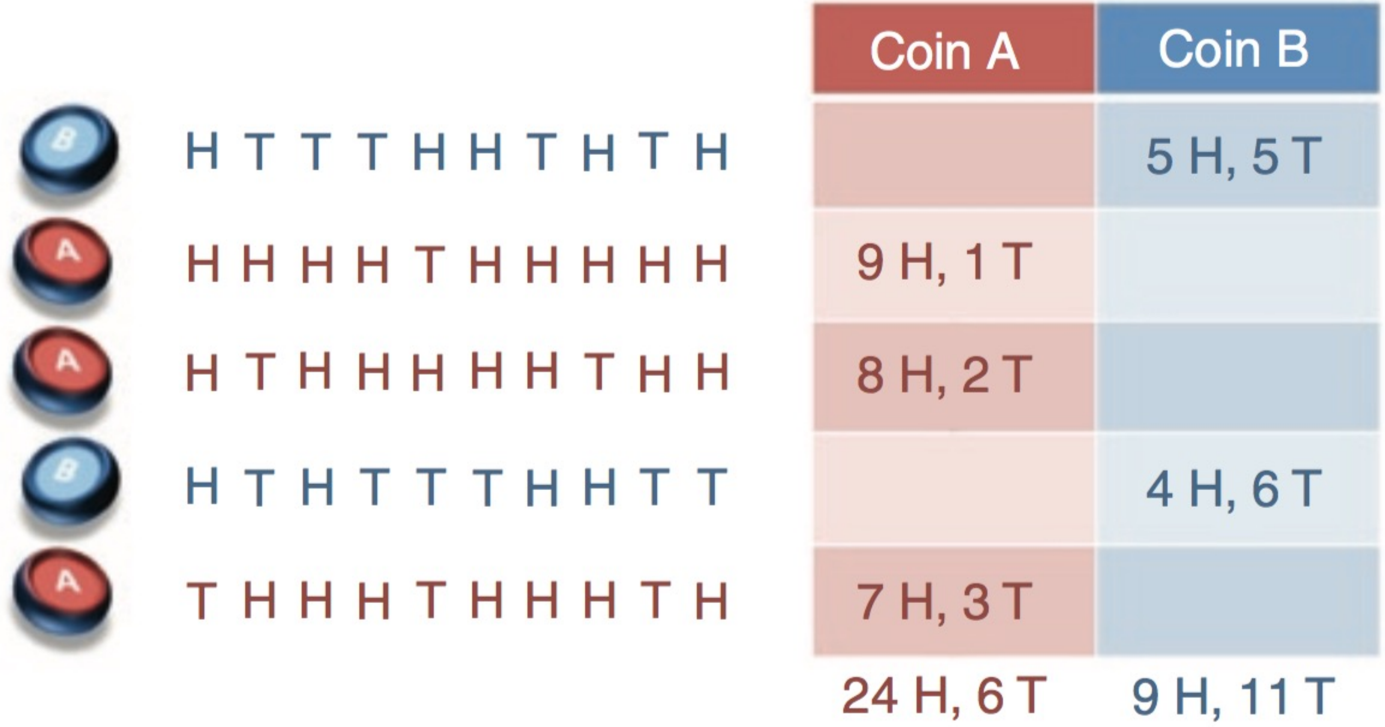
$\hat{\theta}_B = \frac{9}{9+11} = 0.45$

↓

model의 parameter

data의 확률을 maximize 할수 있는
모델 파라미터를 정할 수 있다

Maximum likelihood estimation (MLE)



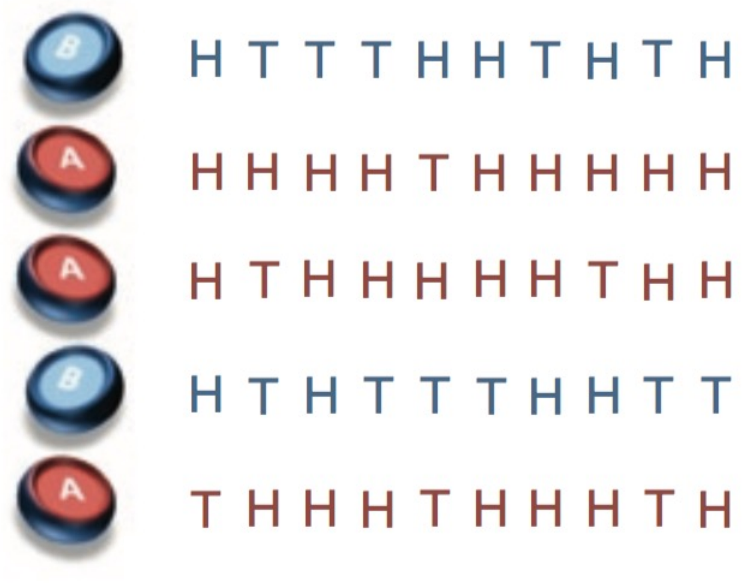
		Coin A	Coin B
B	H T T T H H T H T H		5 H, 5 T
A	H H H H T H H H H H	9 H, 1 T	
A	H T H H H H H T H H	8 H, 2 T	
B	H T H T T T H H T T		4 H, 6 T
A	T H H H T H H H T H	7 H, 3 T	
		24 H, 6 T	9 H, 11 T

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

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

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

Maximum likelihood estimation (MLE)



? H H H H T T H T H T $h = 6,\ t = 4$

$$P(y = A) = ((\hat{\theta}_A)^h (1 - \hat{\theta}_A)^t \quad \hat{\theta}_A = 0.8$$

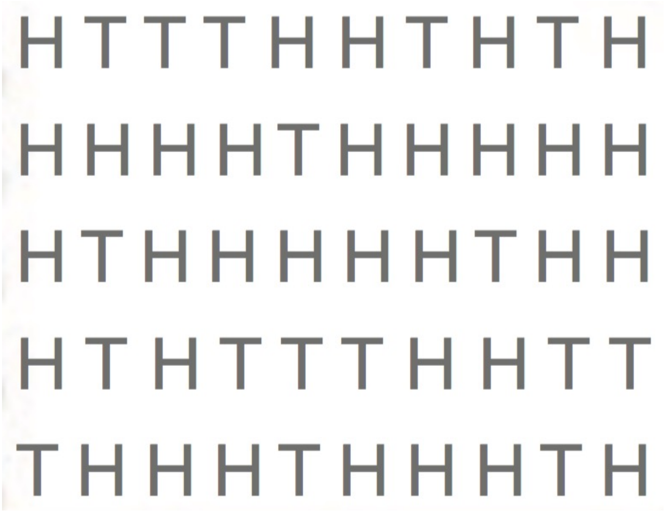
$$P(y = B) = ((\hat{\theta}_B)^h (1 - \hat{\theta}_B)^t \quad \hat{\theta}_B = 0.45$$

$$P(y = A) > P(y = B)$$

label 을 통해 model parameter를 구함

Expectation-Maximization (EM) vs. MLE

Unsupervised



? 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

label estimation

$\hat{\theta}_A =$?

$\hat{\theta}_B =$?

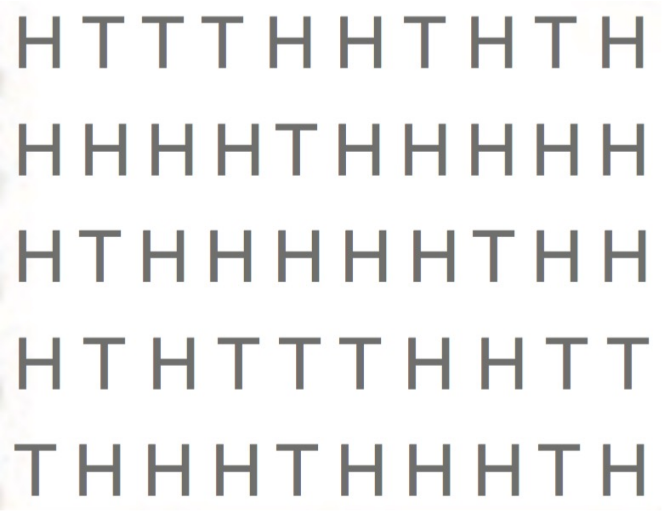
model estimation

? H H H H T T H T H T

label, model parameter 둘 다 모름 → ① Expectation ② Maximization

EM 알고리즘 사용

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)

Unknown variable,
model parameter) 둘을 모를때 EM 사용

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

⇒ model parameter set 가능

supervised 처럼

X: observed variable
앞면의 개수

Z: hidden variable
동전의 type (A/B)

θ : parameters for model
앞면이 나올 확률

D: data set

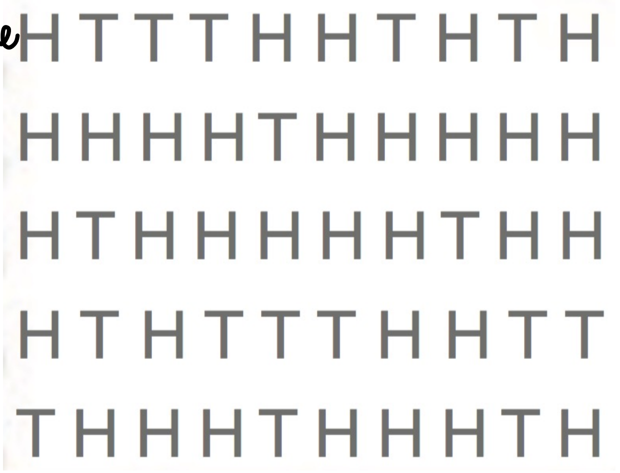
global optimal 보장 X

```
☆
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)   data의 probability를 최대화하는 쪽으로 model parameter를 정함

        $\theta = \text{argmax}\ P( D \mid \theta, Z_{\text{estimated}} )$
```

☆ 각 step이 무얼 의미하는지 시험 나옴

EM: coin example

Hidden Variable



d^1 ? H T T T H H T H T H

d^2 ? H H H H T H H H H H

d^3 ? H T H H H H H T H H

d^4 ? H T H T T T H H T T

d^5 ? T H H H T H H H T H

data point

$\hat{\theta}_A =$?

$\hat{\theta}_B =$?

observed data

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

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

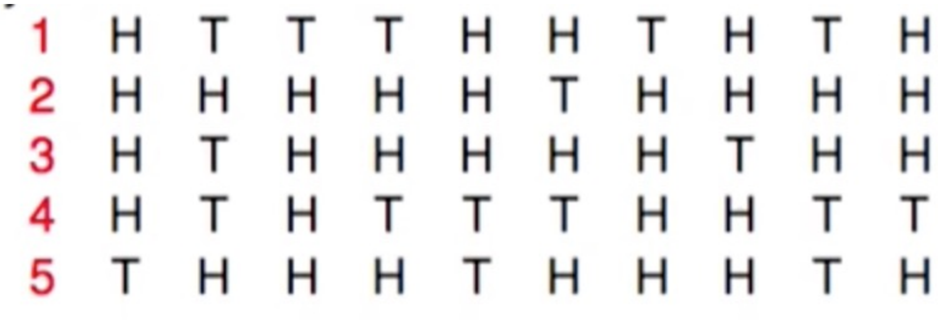
For example, $x^1 = 5$, $x^2 = 9$, $x^3 = 8$, $x^4 = 4$, $x^5 = 7$

$Z = \{z^1, z^2, z^3, z^4, z^5\}$ 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



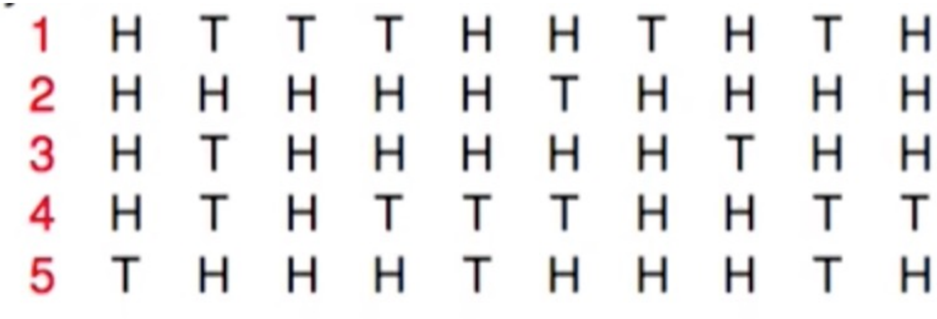
1	H	T	T	T	H	H	T	H	T	H
2	H	H	H	H	H	T	H	H	H	H
3	H	T	H	H	H	H	H	T	H	H
4	H	T	H	T	T	T	H	H	T	T
5	T	H	H	H	T	H	H	H	T	H

Is the first toss from A or B? z^1 =A or B when $x^1 = 5$?

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

→ $P(z^1 = A \mid d^1) > P(z^1 = B \mid d^1)$?

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

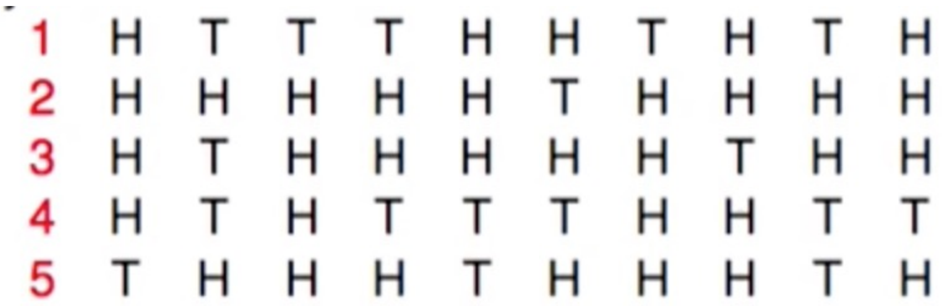
베이지안 //

$\theta_A = 0.6$, $\theta_B = 0.5$ (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 $P(d^i \mid \theta_A)$ and $P(d^i \mid \theta_B)$

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

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

$P(d^1 \mid \theta_A) = {}_{10}C_5 \ 0.6^5 \ 0.4^5$

$P(d^1 \mid \theta_B) = {}_{10}C_5 \ 0.5^5 \ 0.5^5$

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

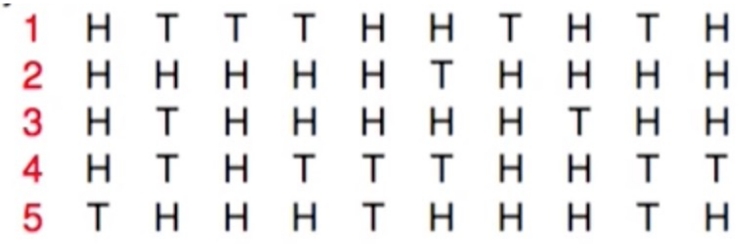
$P(z^1 = B \mid d^1) = 0.55$

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

k is the number of heads-up

θ is the probability of heads-up

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

$\theta_A^{(0)} = 0.6, \quad \theta_B^{(0)} = 0.5$ → 랜덤하게 model parameter assign

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

	X	PA	PB	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

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

observed data

앞면 나올 개수

	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

unobserved data estimation

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

observed data

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

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

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

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

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

Estimation

Set Unknown variable Z

data가 나올 확률이 더 높은 모델로 assign

[expectation]

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

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

$$P(z^I = A \mid d^I) = \frac{P(d_I \mid \theta_A^{(I)})}{P(d^I \mid \theta_A^{(I)}) + P(d^I \mid \theta_B^{(I)})} = 0.1$$

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

→

Hard assignment

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

다시
model parameter
estimation

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

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

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

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

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

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

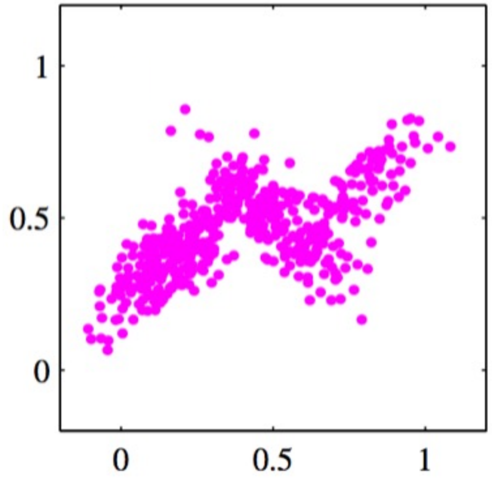
θ : parameters for model

```
assign arbitrary values for parameters $\theta$
iterate until convergence
    E step: estimate the values of hidden variable Z by using $\theta$ and X
            $Z = \text{argmax } P(Z \mid X, \theta)$
    M step: obtain more accurate  parameters $\theta$ using observed variable X and estimated Z
            (use MLE for parameters)
            $\theta = \text{argmax } P( D \mid \theta, Z_{\text{estimated}} )$
```

Types of assignments

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

학습속도는 좀 느려도 optimal



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	

	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

x is the number of heads

z is the type of coin

학습의 속도는 조금 느리지만
더 optimal하게 갈수 있음

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

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

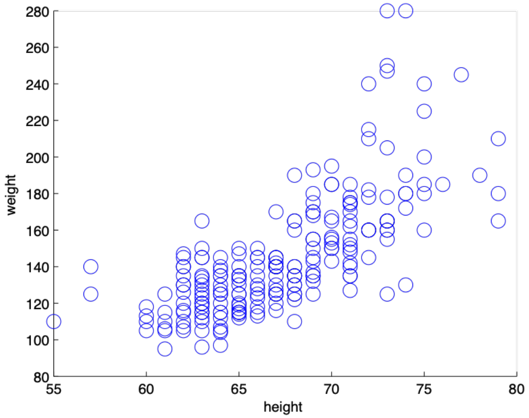
E-step: assign the expected values to the hidden variable based on the given model

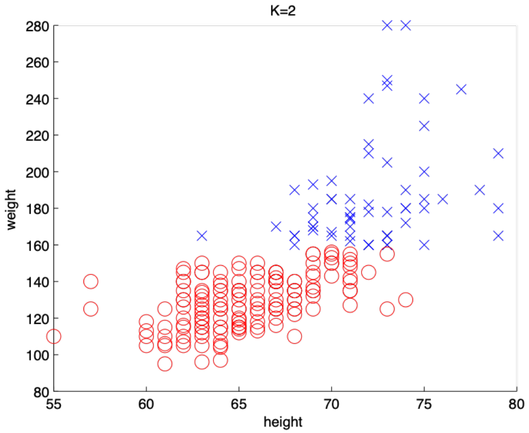
M-step: update the parameters that maximize the probability

Unsupervised learning

- Discovering clusters

Clustering





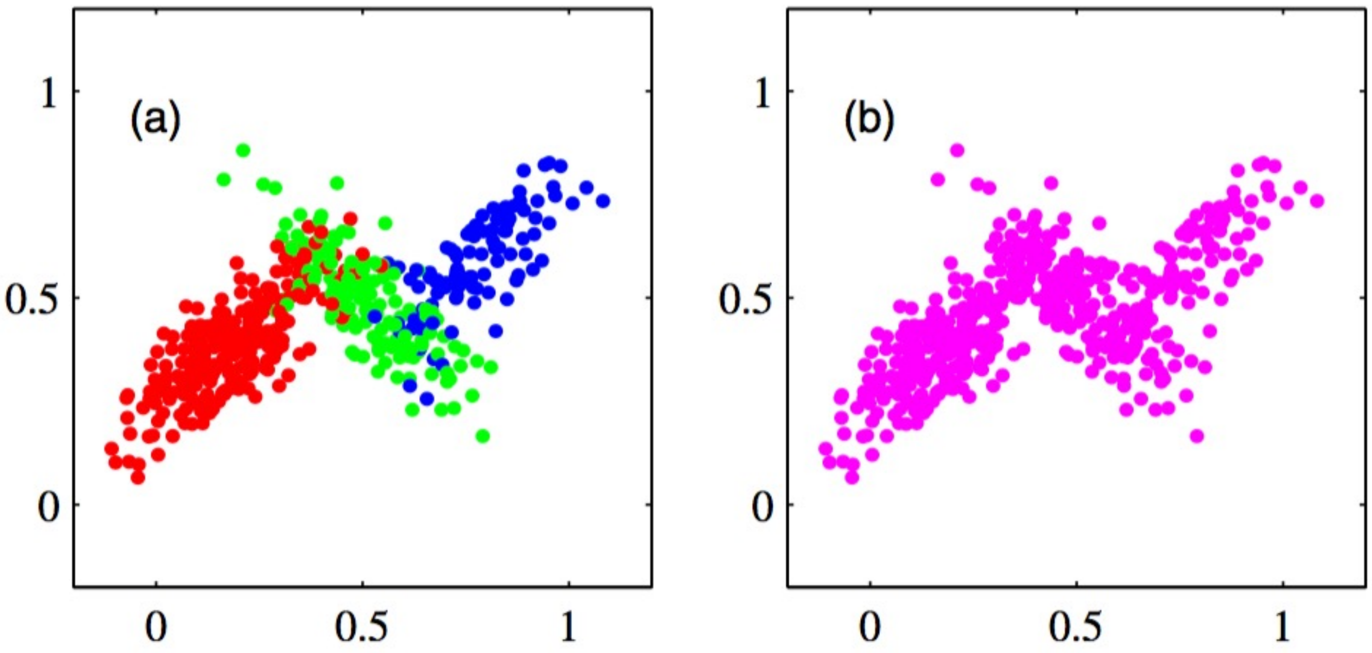
$$z_i^* = \operatorname{argmax}_k p(z_i = k|\mathbf{x}_i, \mathcal{D})$$

Unknown variable

Latent variable

↪ cluster

Unsupervised learning

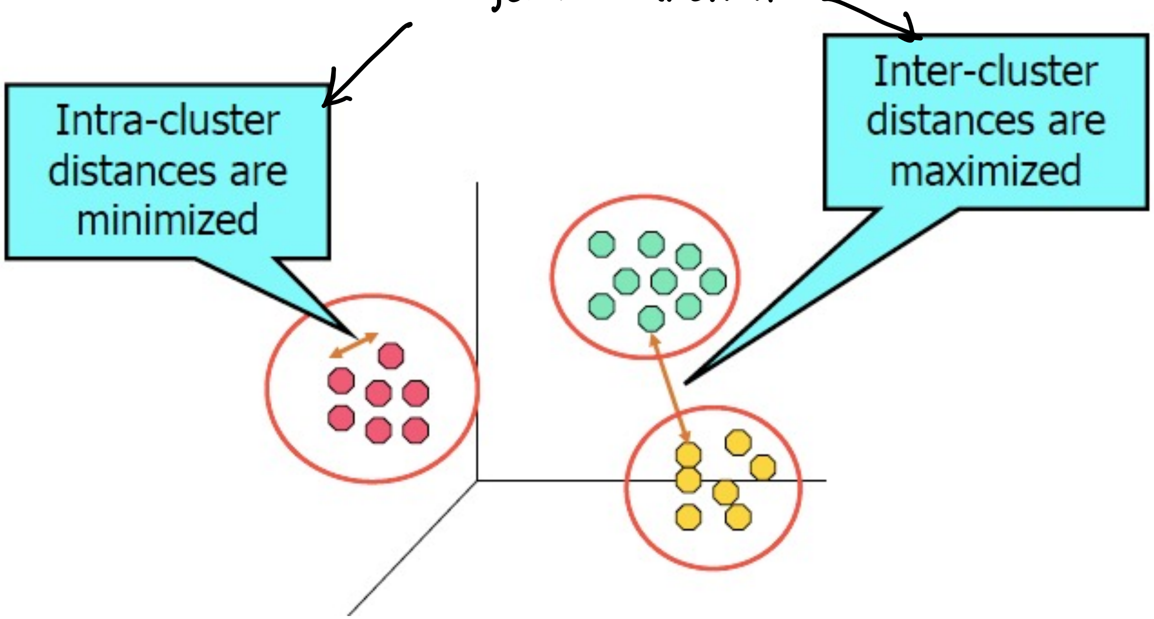


γ

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

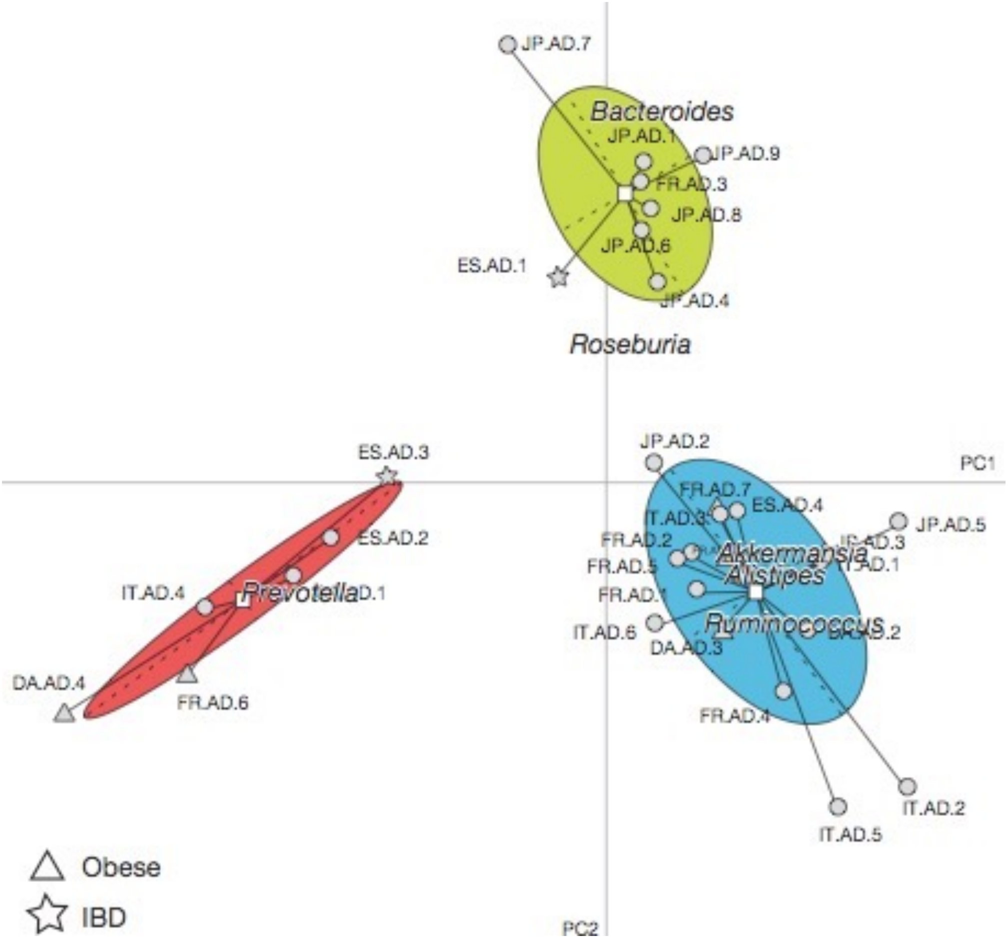
Clustering의 loss function, objective function

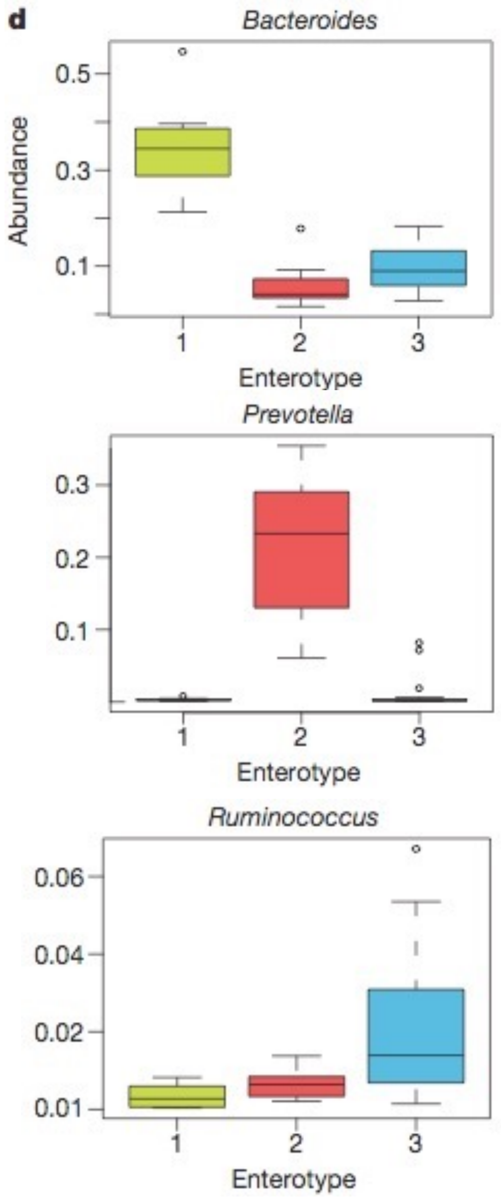


Intra-cluster distances are minimized

Inter-cluster distances are maximized

Clustering in biomedical data





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

K-means clustering

cluster label을 set하는 과정 : clustering

EM ☆

k-means

input : observed data x

- 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

output , unknown variable

estimate

→ 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

model parameter 역할

- μ_k: prototype associated with the kth 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 ≠ k

data point

$$
\begin{array}{cc}
 & \begin{array}{ccc} \boxed{1} & \boxed{2} & \boxed{3} \end{array} \\
(r_{nk}) = & \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}
\end{array}
$$

cluster

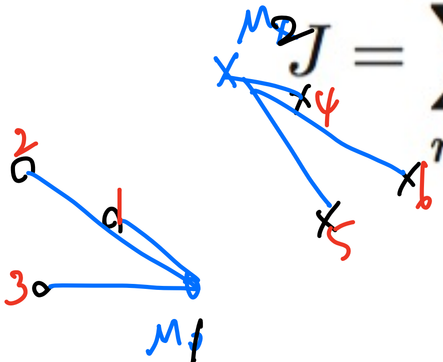
one hot encoding

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

objective function → 각 data point터 가장 가까운 벡터라고 지정된 것의 거리^{2}의 합을 minimize하는 function

$$
\begin{array}{c|cc}
1 & 1 & 0 \\
2 & 1 & 0 \\
3 & 1 & 0 \\
4 & 0 & 1 \\
5 & 0 & 1 \\
6 & 0 & 1
\end{array}
$$

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



1에 해당하는 center 값만 더함

⇒ 자기가 assign된 k에 해당하는 그 가운데 있는 센터값과 자기의 거리를 전부 더함

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;
      =center

      → determine r_{nk}

labeling

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

data x와 center값의 거리가 가장 최소인 J를 K에 assign

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

  → determine μ_k with the new r_{nk}

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

distance를 → 최고로 낮게 (MLE)

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

For each k,

set the derivative of J to 0 with respect to μ_k

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

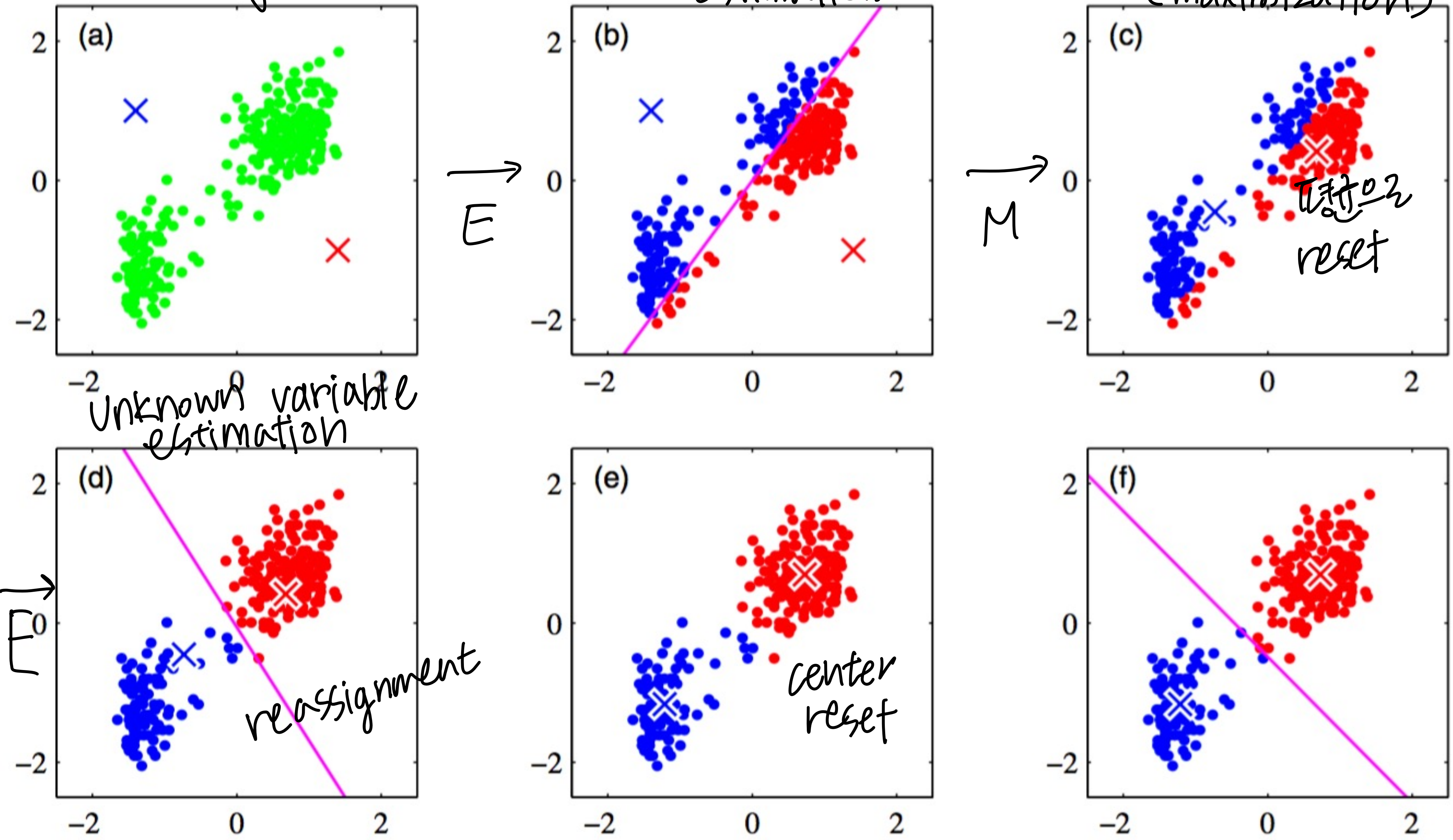
모든 x 값의 평균으로 모델 parameter set

K-means clustering (EM에 기반한 clustering 방법)

아무렇게 μ값 assign
Random guess

Unknown variable estimation

model parameter set (maximization)



E

M

평균으로 reset

Unknown variable estimation

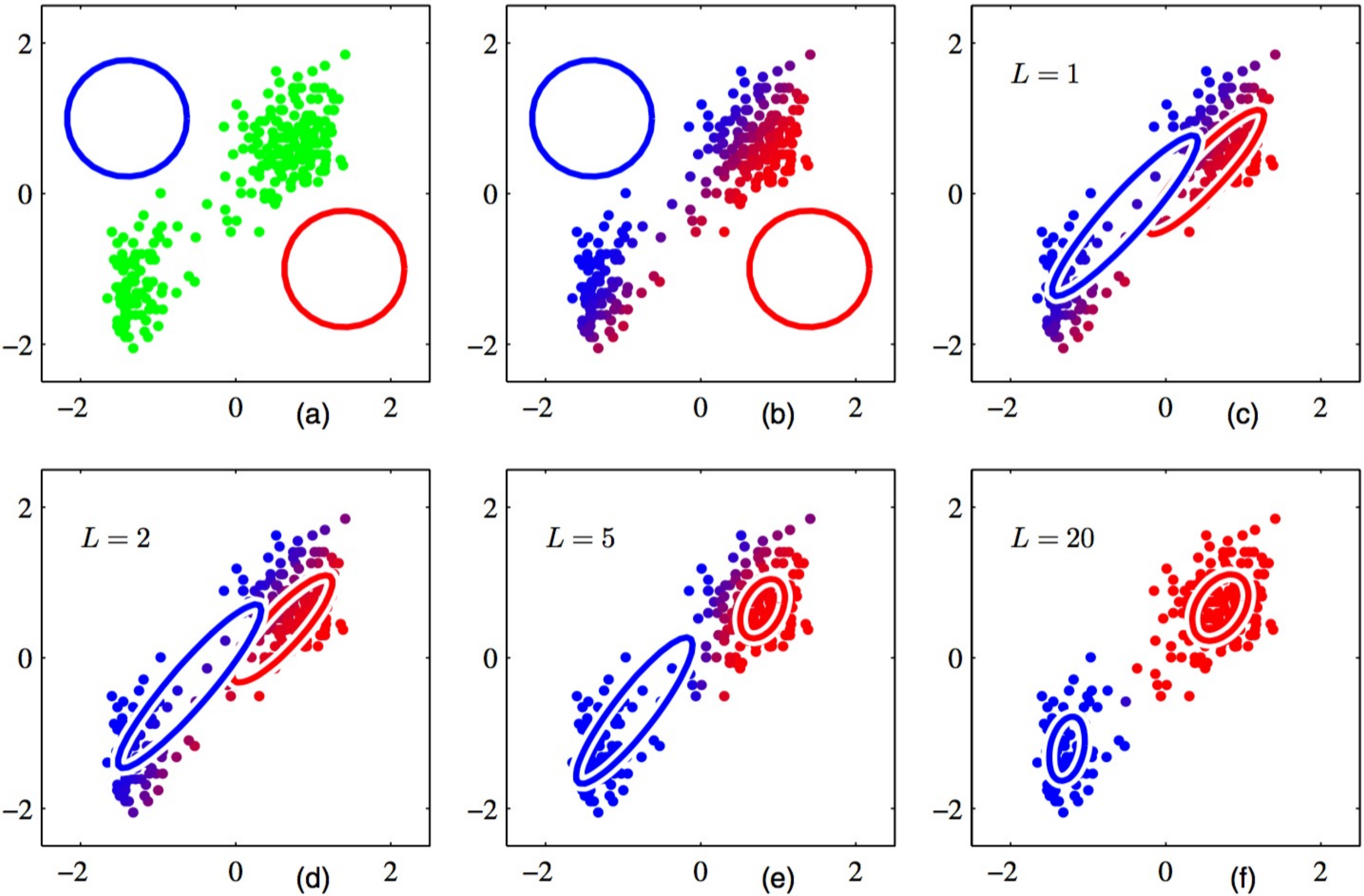
E

reassignment

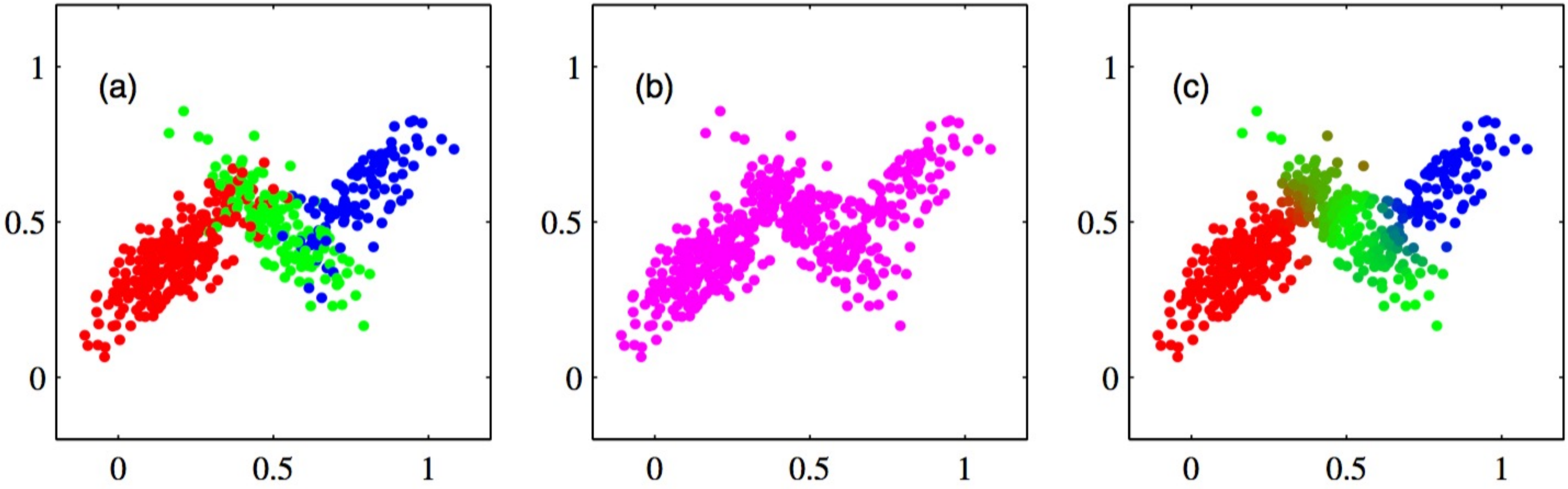
center reset

반복

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