Algorithm Foundations of Data Science and Engineering

Lecture 13: EM Algorithm

MING GAO

DaSE @ ECNU
(for course related communications)
mgao@dase.ecnu.edu.cn

Jun. 11, 2019

Outline

A probabilistic game

There are 6 balls in a bag, 3 are red, 2 are yellow and 1 is blue. What is the probability of picking a yellow?

A probabilistic game

There are 6 balls in a bag, 3 are red, 2 are yellow and 1 is blue. What is the probability of picking a yellow?

- If we pick 20 times, we obtain 5 red, 12 yellow, and 3 blue;

A probabilistic game

There are 6 balls in a bag, 3 are red, 2 are yellow and 1 is blue. What is the probability of picking a yellow?

- If we pick 20 times, we obtain 5 red, 12 yellow, and 3 blue;
- Do you believe there are 3 are red, 2 are yellow and 1 is blue balls in the bag?

A probabilistic game

There are 6 balls in a bag, 3 are red, 2 are yellow and 1 is blue. What is the probability of picking a yellow?

- If we pick 20 times, we obtain 5 red, 12 yellow, and 3 blue;
- Do you believe there are 3 are red, 2 are yellow and 1 is blue balls in the bag?
- How about 2 are red, 3 are yellow and 1 is blue?

Intuition of likelihood

Intuition of likelihood

If we toss an unfair coin 10 times, the result is

$$T, T, T, H, T, T, T, H, T, T.$$

Would you guess the probability $P(H)$?

Intuition of likelihood

If we toss an unfair coin 10 times, the result is

$$T, T, T, H, T, T, T, H, T, T.$$

Would you guess the probability $P(H)$?

- We wish to fit the parameters of a model $p(x; \theta)$ to the data;

Intuition of likelihood

If we toss an unfair coin 10 times, the result is

$$T, T, T, H, T, T, T, H, T, T.$$

Would you guess the probability $P(H)$?

- We wish to fit the parameters of a model $p(x; \theta)$ to the data;
- We can infer the parameter via maximizing the joint probability that observes the above sample.

Definition of likelihood

Suppose that sample point $X_1, X_2, \cdots, X_n$ have a joint density or pmf $f(x_1, x_2, \cdots, x_n|\theta)$. Given observe value of $X_i = x_i$, the **likelihood** of θ as a function of $x_1, x_2, \cdots, x_n$ is defined as

Definition of likelihood

Suppose that sample point $X_1, X_2, \cdots, X_n$ have a joint density or pmf $f(x_1, x_2, \cdots, x_n|\theta)$. Given observe value of $X_i = x_i$, the **likelihood** of θ as a function of $x_1, x_2, \cdots, x_n$ is defined as

$$\begin{aligned} L(\theta|\mathbf{x}) &= f(x_1, x_2, \cdots, x_n|\theta) \\ &= \prod_{i=1}^{k} f(x_i|\theta). \end{aligned}$$

Definition of likelihood

Suppose that sample point $X_1, X_2, \cdots, X_n$ have a joint density or pmf $f(x_1, x_2, \cdots, x_n|\theta)$. Given observe value of $X_i = x_i$, the **likelihood** of θ as a function of $x_1, x_2, \cdots, x_n$ is defined as

$$L(\theta|\mathbf{x}) = f(x_1, x_2, \cdots, x_n|\theta)$$
$$= \prod_{i=1}^{k} f(x_i|\theta).$$

- The method of maximum likelihood can be applied to a great variety of other statistical problems, such as curve fitting, testing, and machine learning, etc.

Definition of likelihood

Suppose that sample point $X_1, X_2, \cdots, X_n$ have a joint density or pmf $f(x_1, x_2, \cdots, x_n|\theta)$. Given observe value of $X_i = x_i$, the **likelihood** of θ as a function of $x_1, x_2, \cdots, x_n$ is defined as

$$\begin{aligned} L(\theta|\mathbf{x}) &= f(x_1, x_2, \cdots, x_n|\theta) \\ &= \prod_{i=1}^{k} f(x_i|\theta). \end{aligned}$$

- The method of maximum likelihood can be applied to a great variety of other statistical problems, such as curve fitting, testing, and machine learning, etc.
- Maximum likelihood estimates have nice theoretical properties as well.

Maximum likelihood estimator

Definition

For each sample point $\mathbf{x}$, let $\widehat{\theta}(\mathbf{x})$ be a parameter value at which $L(\theta|\mathbf{x})$ attains its maximum as a function of θ, with $\mathbf{x}$ held fixed. That is,

$$\widehat{\theta}(\mathbf{x}) = argmax_{\theta} \;\; L(\theta|\mathbf{x}).$$

A **maximum likelihood estimator (MLE)** of the parameter θ based on a sample $\mathbf{X}$ is $\widehat{\theta}(\mathbf{x})$.

Maximum likelihood estimator

Definition

For each sample point $\mathbf{x}$, let $\widehat{\theta}(\mathbf{x})$ be a parameter value at which $L(\theta|\mathbf{x})$ attains its maximum as a function of θ, with $\mathbf{x}$ held fixed. That is,

$$\widehat{\theta}(\mathbf{x}) = argmax_{\theta} \;\; L(\theta|\mathbf{x}).$$

A **maximum likelihood estimator (MLE)** of the parameter θ based on a sample $\mathbf{X}$ is $\widehat{\theta}(\mathbf{x})$.

- The MLE is the parameter point for which the observed sample is most likely.

Maximum likelihood estimator

Definition

For each sample point $\mathbf{x}$, let $\widehat{\theta}(\mathbf{x})$ be a parameter value at which $L(\theta|\mathbf{x})$ attains its maximum as a function of θ, with $\mathbf{x}$ held fixed. That is,

$$\widehat{\theta}(\mathbf{x}) = argmax_{\theta} \;\; L(\theta|\mathbf{x}).$$

A **maximum likelihood estimator (MLE)** of the parameter θ based on a sample $\mathbf{X}$ is $\widehat{\theta}(\mathbf{x})$.

- The MLE is the parameter point for which the observed sample is most likely.
- There are two drawbacks:
  - It is actually to find the global maximum. In many cases, this problem reduces to a simple differential calculus exercise but, sometimes even for common densities, difficulties do arise.
  - It is numerical sensitivity.

How to obtain the maximum likelihood estimator

If the likelihood function is differentiable (in θ_i), possible candidates for the MLE are the values of $(\theta_1, \cdots, \theta_k)$ that solve

$$\frac{\partial}{\partial \theta_i} L(\theta|\mathbf{x}) = 0, i = 1, \cdots, k.$$

How to obtain the maximum likelihood estimator

If the likelihood function is differentiable (in θ_i), possible candidates for the MLE are the values of $(\theta_1, \cdots, \theta_k)$ that solve

$$\frac{\partial}{\partial \theta_i} L(\theta|\mathbf{x}) = 0, i = 1, \cdots, k.$$

Note that the solutions are only possible candidates for the MLE since the first derivative being 0 is only a necessary condition for a maximum, not a sufficient condition.

How to obtain the maximum likelihood estimator

If the likelihood function is differentiable (in θ_i), possible candidates for the MLE are the values of $(\theta_1, \cdots, \theta_k)$ that solve

$$\frac{\partial}{\partial \theta_i} L(\theta|\mathbf{x}) = 0, i = 1, \cdots, k.$$

Note that the solutions are only possible candidates for the MLE since the first derivative being 0 is only a necessary condition for a maximum, not a sufficient condition.

- The zeros of the first derivative locate only extreme points in the interior of the domain of a function.

How to obtain the maximum likelihood estimator

If the likelihood function is differentiable (in θ_i), possible candidates for the MLE are the values of $(\theta_1, \cdots, \theta_k)$ that solve

$$\frac{\partial}{\partial \theta_i} L(\theta|\mathbf{x}) = 0, i = 1, \cdots, k.$$

Note that the solutions are only possible candidates for the MLE since the first derivative being 0 is only a necessary condition for a maximum, not a sufficient condition.

- The zeros of the first derivative locate only extreme points in the interior of the domain of a function.
- If the extrema occur on the boundary the first derivative may not be 0.

How to obtain the maximum likelihood estimator

If the likelihood function is differentiable (in θ_i), possible candidates for the MLE are the values of $(\theta_1, \cdots, \theta_k)$ that solve

$$\frac{\partial}{\partial \theta_i} L(\theta|\mathbf{x}) = 0, i = 1, \cdots, k.$$

Note that the solutions are only possible candidates for the MLE since the first derivative being 0 is only a necessary condition for a maximum, not a sufficient condition.

- The zeros of the first derivative locate only extreme points in the interior of the domain of a function.
- If the extrema occur on the boundary the first derivative may not be 0.
- Thus, the boundary must be checked separately for extrema.

Poisson MLE

Let $X_1, X_2, \cdots, X_n$ be i.i.d. Poisson sample $\frac{\lambda^x}{x!}e^{-\lambda}$, and let $L(\lambda|\mathbf{x})$ denote the likelihood function. Then

$$\log L(\lambda|\mathbf{x}) = \sum_{i=1}^{n}(X_i \log\lambda - \lambda - \log X_i!) = \log\lambda \sum_{i=1}^{n} X_i - n\lambda - \sum_{i=1}^{n}\log(X_i!)$$

Poisson MLE

Let $X_1, X_2, \cdots, X_n$ be i.i.d. Poisson sample $\frac{\lambda^x}{x!}e^{-\lambda}$, and let $L(\lambda|\mathbf{x})$ denote the likelihood function. Then

$$\log L(\lambda|\mathbf{x}) = \sum_{i=1}^{n}(X_i \log\lambda - \lambda - \log X_i!) = \log\lambda \sum_{i=1}^{n} X_i - n\lambda - \sum_{i=1}^{n}\log(X_i!)$$

The equation $\frac{d}{d\lambda}\log L(\lambda|\mathbf{x}) = 0$ reduces to

$$\frac{1}{\lambda}\sum_{i=1}^{n} x_i - n = 0,$$

which has the solution $\hat{\lambda} = \overline{x}$.

Poisson MLE

Let $X_1, X_2, \cdots, X_n$ be i.i.d. Poisson sample $\frac{\lambda^x}{x!}e^{-\lambda}$, and let $L(\lambda|\mathbf{x})$ denote the likelihood function. Then

$$\log L(\lambda|\mathbf{x}) = \sum_{i=1}^{n}(X_i \log\lambda - \lambda - \log X_i!) = \log\lambda \sum_{i=1}^{n} X_i - n\lambda - \sum_{i=1}^{n}\log(X_i!)$$

The equation $\frac{d}{d\lambda}\log L(\lambda|\mathbf{x}) = 0$ reduces to

$$\frac{1}{\lambda}\sum_{i=1}^{n} x_i - n = 0,$$

which has the solution $\hat{\lambda} = \overline{x}$. The MLE agrees with the method of moments for this case and thus has the same sampling distribution.

Poisson MLE

Let $X_1, X_2, \cdots, X_n$ be i.i.d. Poisson sample $\frac{\lambda^x}{x!}e^{-\lambda}$, and let $L(\lambda|\mathbf{x})$ denote the likelihood function. Then

$$\log L(\lambda|\mathbf{x}) = \sum_{i=1}^{n}(X_i \log\lambda - \lambda - \log X_i!) = \log\lambda \sum_{i=1}^{n} X_i - n\lambda - \sum_{i=1}^{n}\log(X_i!)$$

The equation $\frac{d}{d\lambda}\log L(\lambda|\mathbf{x}) = 0$ reduces to

$$\frac{1}{\lambda}\sum_{i=1}^{n} x_i - n = 0,$$

which has the solution $\hat{\lambda} = \overline{x}$. The MLE agrees with the method of moments for this case and thus has the same sampling distribution.

- Note that $\widehat{\theta} = \overline{x}$ is the only solution to $\sum_{i=1}^{n}(x_i - \theta) = 0$.

Poisson MLE

Let $X_1, X_2, \cdots, X_n$ be i.i.d. Poisson sample $\frac{\lambda^x}{x!}e^{-\lambda}$, and let $L(\lambda|\mathbf{x})$ denote the likelihood function. Then

$$\log L(\lambda|\mathbf{x}) = \sum_{i=1}^{n}(X_i \log\lambda - \lambda - \log X_i!) = \log\lambda \sum_{i=1}^{n} X_i - n\lambda - \sum_{i=1}^{n}\log(X_i!)$$

The equation $\frac{d}{d\lambda}\log L(\lambda|\mathbf{x}) = 0$ reduces to

$$\frac{1}{\lambda}\sum_{i=1}^{n} x_i - n = 0,$$

which has the solution $\hat{\lambda} = \overline{x}$. The MLE agrees with the method of moments for this case and thus has the same sampling distribution.

- Note that $\widehat{\theta} = \overline{x}$ is the only solution to $\sum_{i=1}^{n}(x_i - \theta) = 0$.
- We can verify that $\frac{d^2}{d\theta^2}L(\theta|\mathbf{x})|\theta = \overline{x} < 0$. Thus $\overline{x}$ is the only extreme point in the interior and it is a maximum.

Poisson MLE

Let $X_1, X_2, \cdots, X_n$ be i.i.d. Poisson sample $\frac{\lambda^x}{x!}e^{-\lambda}$, and let $L(\lambda|\mathbf{x})$ denote the likelihood function. Then

$$\log L(\lambda|\mathbf{x}) = \sum_{i=1}^{n}(X_i \log\lambda - \lambda - \log X_i!) = \log\lambda \sum_{i=1}^{n} X_i - n\lambda - \sum_{i=1}^{n}\log(X_i!)$$

The equation $\frac{d}{d\lambda}\log L(\lambda|\mathbf{x}) = 0$ reduces to

$$\frac{1}{\lambda}\sum_{i=1}^{n} x_i - n = 0,$$

which has the solution $\hat{\lambda} = \overline{x}$. The MLE agrees with the method of moments for this case and thus has the same sampling distribution.

- Note that $\widehat{\theta} = \overline{x}$ is the only solution to $\sum_{i=1}^{n}(x_i - \theta) = 0$.
- We can verify that $\frac{d^2}{d\theta^2}L(\theta|\mathbf{x})|\theta = \overline{x} < 0$. Thus $\overline{x}$ is the only extreme point in the interior and it is a maximum.
- By taking limits it is easy to establish that the likelihood is 0 at $\pm\infty$.

Poisson MLE

Let $X_1, X_2, \cdots, X_n$ be i.i.d. Poisson sample $\frac{\lambda^x}{x!}e^{-\lambda}$, and let $L(\lambda|\mathbf{x})$ denote the likelihood function. Then

$$\log L(\lambda|\mathbf{x}) = \sum_{i=1}^{n}(X_i \log\lambda - \lambda - \log X_i!) = \log\lambda \sum_{i=1}^{n} X_i - n\lambda - \sum_{i=1}^{n}\log(X_i!)$$

The equation $\frac{d}{d\lambda}\log L(\lambda|\mathbf{x}) = 0$ reduces to

$$\frac{1}{\lambda}\sum_{i=1}^{n} x_i - n = 0,$$

which has the solution $\hat{\lambda} = \overline{x}$. The MLE agrees with the method of moments for this case and thus has the same sampling distribution.

- Note that $\widehat{\theta} = \overline{x}$ is the only solution to $\sum_{i=1}^{n}(x_i - \theta) = 0$.
- We can verify that $\frac{d^2}{d\theta^2}L(\theta|\mathbf{x})|\theta = \overline{x} < 0$. Thus $\overline{x}$ is the only extreme point in the interior and it is a maximum.
- By taking limits it is easy to establish that the likelihood is 0 at $\pm\infty$.

Therefore, $\widehat{\theta} = \overline{x}$ is a global maximum and $\overline{X}$ is the MLE.

Normal MLE

Let $X_1, X_2, \cdots, X_n$ be i.i.d. $N(\mu, \sigma)$, and let $L(\theta|\mathbf{x})$ denote the likelihood function. Then

$$\log L(\mu, \sigma|\mathbf{x}) = -n \log \sigma - \frac{n}{2} \log 2\pi - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (X_i - \mu)^2$$

Normal MLE

Let $X_1, X_2, \cdots, X_n$ be i.i.d. $N(\mu, \sigma)$, and let $L(\theta|\mathbf{x})$ denote the likelihood function. Then

$$\log L(\mu, \sigma|\mathbf{x}) = -n \log \sigma - \frac{n}{2} \log 2\pi - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (X_i - \mu)^2$$

The equations $\frac{\partial}{\partial \mu} \log L(\mu, \sigma|\mathbf{x}) = 0$ and $\frac{\partial}{\partial \sigma} \log L(\mu, \sigma|\mathbf{x}) = 0$ reduce to

$$\frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \mu) = 0 \textit{ and } -\frac{n}{\sigma} + \sigma^{-3} \sum_{i=1}^{n} (x_i - \mu)^2 = 0$$

which has the solution

$$\widehat{\mu} = \overline{x} \textit{ and } \widehat{\sigma} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})^2}$$

Normal MLE Cont'd

Recall that for any number a,

$$\sum_{i=1}^{n}(x_i - a)^2 \geq \sum_{i=1}^{n}(x_i - \overline{x})^2$$

with equality if and only if $a = \overline{x}$.

Normal MLE Cont'd

Recall that for any number a,

$$\sum_{i=1}^{n}(x_i - a)^2 \geq \sum_{i=1}^{n}(x_i - \overline{x})^2$$

with equality if and only if $a = \overline{x}$. This implies that for any θ,

$$e^{-(1/2)\sum_{i=1}^{n}(x_i - \theta)^2} \leq e^{-(1/2)\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

with equality if and only if $\theta = \overline{x}$. Hence $\overline{X}$ is the MLE.

Normal MLE Cont'd

Recall that for any number a,

$$\sum_{i=1}^{n}(x_i - a)^2 \geq \sum_{i=1}^{n}(x_i - \overline{x})^2$$

with equality if and only if $a = \overline{x}$. This implies that for any θ,

$$e^{-(1/2)\sum_{i=1}^{n}(x_i-\theta)^2} \leq e^{-(1/2)\sum_{i=1}^{n}(x_i-\overline{x})^2}$$

with equality if and only if $\theta = \overline{x}$. Hence $\overline{X}$ is the MLE.
In most cases, especially when differentiation is to be used, it is easier to work with the natural logarithm of $L(\theta|\mathbf{x})$, $\log L(\theta|\mathbf{x})$ (known as the **log likelihood**), than it is to work with $L(\theta|\mathbf{x})$ directly.

Bernoulli MLE

Let $X_1, X_2, \cdots, X_n$ be i.i.d. *Bernoulli*(p). Then the log likelihood function is

$$l(p|\mathbf{x}) = y \log p + (n-y) \log (1-p).$$

where $y = \sum_{i=1}^{n} x_i$.

Bernoulli MLE

Let $X_1, X_2, \cdots, X_n$ be i.i.d. *Bernoulli*(p). Then the log likelihood function is

$$l(p|\mathbf{x}) = y \log p + (n-y) \log(1-p).$$

where $y = \sum_{i=1}^{n} x_i$. If $0 < y < n$, differentiating $\log L(p|\mathbf{x})$ and setting the result equal to 0 give the solution, $\widehat{p} = \frac{1}{n} \sum_{i=1}^{n} x_i$.

Bernoulli MLE

Let $X_1, X_2, \cdots, X_n$ be i.i.d. *Bernoulli*(p). Then the log likelihood function is

$$l(p|\mathbf{x}) = y \log p + (n-y) \log (1-p).$$

where $y = \sum_{i=1}^{n} x_i$. If $0 < y < n$, differentiating $\log L(p|\mathbf{x})$ and setting the result equal to 0 give the solution, $\widehat{p} = \frac{1}{n} \sum_{i=1}^{n} x_i$. If $\sum_{i=1}^{n} x_i = 0$ or $\sum_{i=1}^{n} x_i = n$, then

$$\log L(p|\mathbf{x}) = \begin{cases} n \log (1-p), & \text{if } y = 0 \\ n \log p, & \text{if } y = n \end{cases}$$

In either case $l(p|\mathbf{x})$ is a monotone function of p, and it is again straightforward to verify that $\widehat{p} = \frac{1}{n} \sum_{i=1}^{n} x_i$ in each case.

Bernoulli MLE

Let $X_1, X_2, \cdots, X_n$ be i.i.d. *Bernoulli*(p). Then the log likelihood function is

$$l(p|\mathbf{x}) = y \log p + (n-y) \log (1-p).$$

where $y = \sum_{i=1}^n x_i$. If $0 < y < n$, differentiating $\log L(p|\mathbf{x})$ and setting the result equal to 0 give the solution, $\widehat{p} = \frac{1}{n}\sum_{i=1}^n x_i$. If $\sum_{i=1}^n x_i = 0$ or $\sum_{i=1}^n x_i = n$, then

$$\log L(p|\mathbf{x}) = \begin{cases} n \log (1-p), & \text{if } y = 0 \\ n \log p, & \text{if } y = n \end{cases}$$

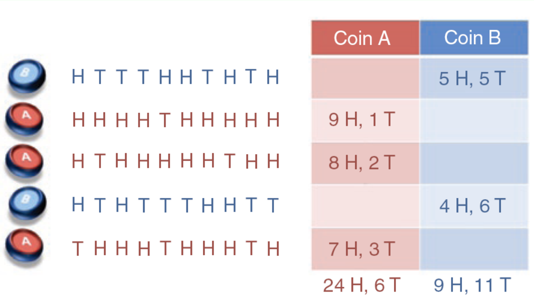
In either case $l(p|\mathbf{x})$ is a monotone function of p, and it is again straightforward to verify that $\widehat{p} = \frac{1}{n}\sum_{i=1}^n x_i$ in each case. Thus, we have shown that $\frac{1}{n}\sum_{i=1}^n x_i$ is the MLE of p.

Outline

EM algorithm: the intuition

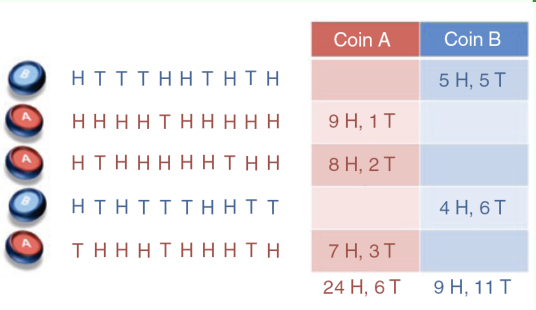
		Coin A	Coin B
B	H T T T H H T H T H		5 H, 5 T
A	H H H H T H H H H H	9 H, 1 T	
A	H T H H H H H T H H	8 H, 2 T	
B	H T H T T T H H T T		4 H, 6 T
A	T H H H T H H H T H	7 H, 3 T	
		24 H, 6 T	9 H, 11 T

EM algorithm: the intuition



Assume that we have two coins, A and B

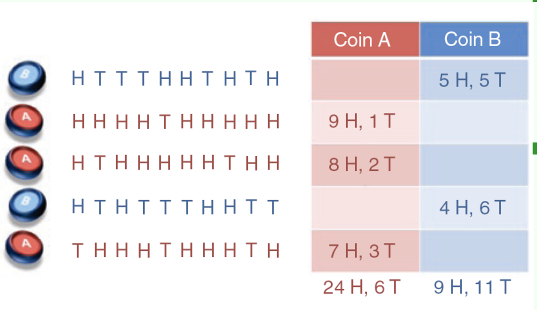
EM algorithm: the intuition



Assume that we have two coins, A and B

- Assume the bias of A is θ_1 (i.e., probability of getting heads with A);

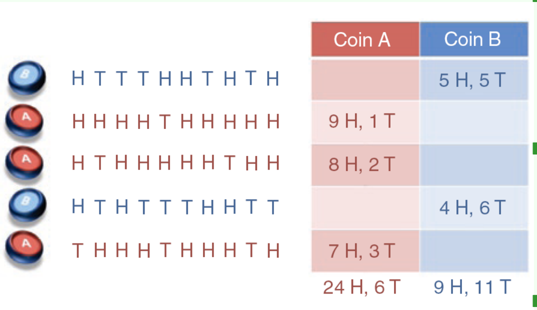
EM algorithm: the intuition



Assume that we have two coins, A and B

- Assume the bias of A is θ_1 (i.e., probability of getting heads with A);
- Assume the bias of B is θ_2 (i.e., probability of getting heads with B);

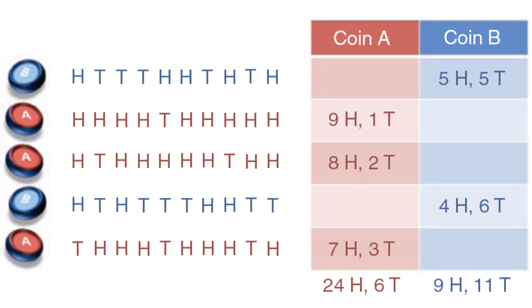
EM algorithm: the intuition



Assume that we have two coins, A and B

- Assume the bias of A is θ_1 (i.e., probability of getting heads with A);
- Assume the bias of B is θ_2 (i.e., probability of getting heads with B);
- We want to find the values of θ_1 and θ_2.

EM algorithm: the intuition

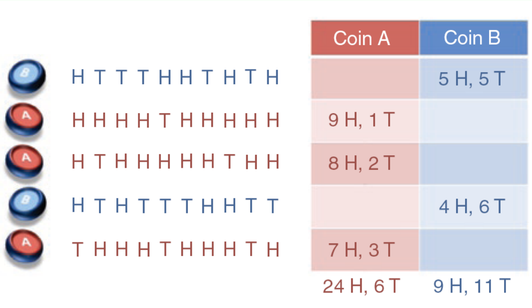


Assume that we have two coins, A and B

Assume a more challenging problem:

- Assume the bias of A is θ_1 (i.e., probability of getting heads with A);
- Assume the bias of B is θ_2 (i.e., probability of getting heads with B);
- We want to find the values of θ_1 and θ_2.

EM algorithm: the intuition



Assume that we have two coins, A and B

- Assume the bias of A is θ_1 (i.e., probability of getting heads with A);
- Assume the bias of B is θ_2 (i.e., probability of getting heads with B);
- We want to find the values of θ_1 and θ_2.

Assume a more challenging problem: we do not know the identities of the coins used for each set of tosses (we treat them as hidden variables).

Outline

Bernoulli mixture models (BMMs)

First experiment

- We choose 5 times one of the coins;

Bernoulli mixture models (BMMs)

First experiment

- We choose 5 times one of the coins;
- We toss the chosen coin 10 times.

Bernoulli mixture models (BMMs)

First experiment

- We choose 5 times one of the coins;
- We toss the chosen coin 10 times.

Bernoulli mixture models (BMMs)

First experiment

- We choose 5 times one of the coins;
- We toss the chosen coin 10 times.

$$\widehat{\theta}_1 = \frac{\#\ \text{heads using } A}{\#\ \text{flips using } A} \quad \textbf{?}$$

$$\widehat{\theta}_2 = \frac{\#\ \text{heads using } B}{\#\ \text{flips using } B} \quad \textbf{?}$$

Bernoulli mixture models (BMMs)

First experiment

- We choose 5 times one of the coins;
- We toss the chosen coin 10 times.

$$\widehat{\theta}_1 = \frac{\#\text{ heads using } A}{\#\text{ flips using } A} \;\; \textbf{?}$$

$$\widehat{\theta}_2 = \frac{\#\text{ heads using } B}{\#\text{ flips using } B} \;\; \textbf{?}$$

The parameters of the model cannot be estimated. Thus, it becomes a more challenging problem.

Unobserved Variables

A variable can be unobserved (latent) because:

Unobserved Variables

A variable can be unobserved (latent) because:

- It is an imaginary quantity meant to provide some simplified and abstractive view of the data generation process, e.g., speech recognition models, mixture models;

Unobserved Variables

A variable can be unobserved (latent) because:

- It is an imaginary quantity meant to provide some simplified and abstractive view of the data generation process, e.g., speech recognition models, mixture models;
- It is a real-world object and/or phenomena, but difficult or impossible to measure, e.g., the temperature of a star, causes of a disease, evolutionary ancestors;

Unobserved Variables

A variable can be unobserved (latent) because:

- It is an imaginary quantity meant to provide some simplified and abstractive view of the data generation process, e.g., speech recognition models, mixture models;
- It is a real-world object and/or phenomena, but difficult or impossible to measure, e.g., the temperature of a star, causes of a disease, evolutionary ancestors;
- It is a real-world object and/or phenomena, but sometimes was not measured, because of faulty sensors; or was measure with a noisy channel, etc., e.g., traffic radio, aircraft signal on a radar screen.

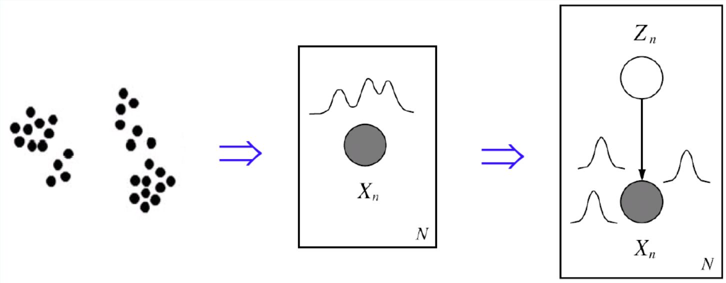
Unobserved Variables

A variable can be unobserved (latent) because:

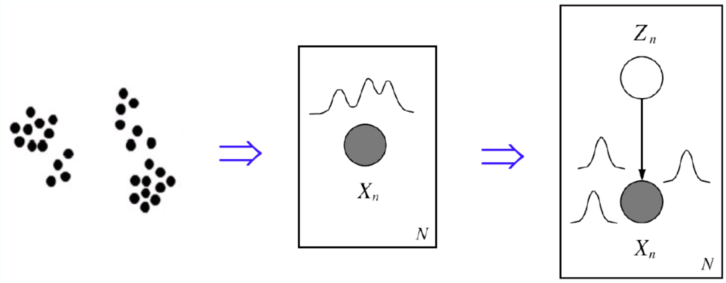
- It is an imaginary quantity meant to provide some simplified and abstractive view of the data generation process, e.g., speech recognition models, mixture models;
- It is a real-world object and/or phenomena, but difficult or impossible to measure, e.g., the temperature of a star, causes of a disease, evolutionary ancestors;
- It is a real-world object and/or phenomena, but sometimes was not measured, because of faulty sensors; or was measure with a noisy channel, etc., e.g., traffic radio, aircraft signal on a radar screen.

Discrete latent variables can be used to partition/cluster data into sub-groups (mixture models).

Mixed models

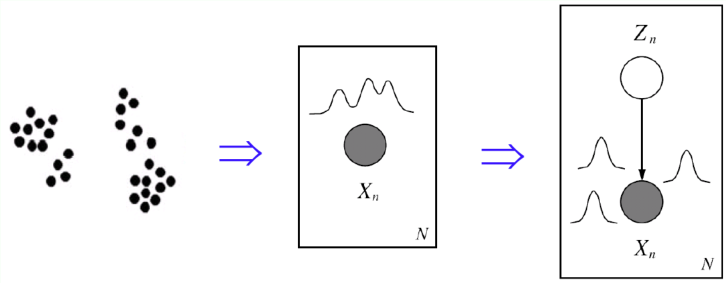


Mixed models



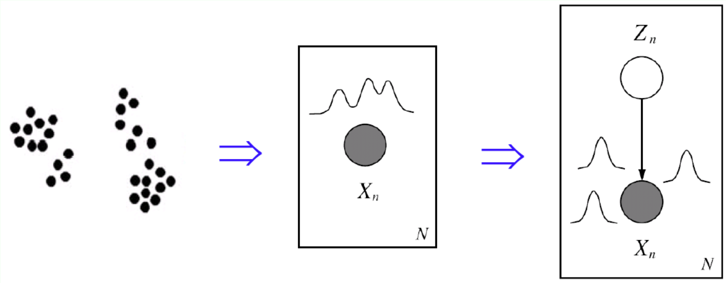
- A density model $p(x)$ may be multi-modal;

Mixed models



- A density model $p(x)$ may be multi-modal;
- We may be able to model it as a mixture of uni-modal distributions (e.g., Bernoulli or Gaussians, etc).

Mixed models



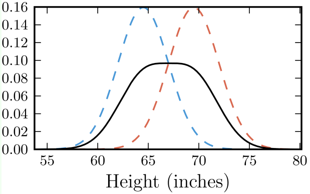
- A density model $p(x)$ may be multi-modal;
- We may be able to model it as a mixture of uni-modal distributions (e.g., Bernoulli or Gaussians, etc).
- Each mode may correspond to a different sub-population (e.g., male and female).

Gaussian mixture models (GMMs)

For example, the height of a randomly chosen man is normally distributed with a mean around $5'9.5''$ and standard deviation around $2.5''$. Similarly, the height of a randomly chosen woman is normally distributed with a mean around $5'4.5''$ and standard deviation around $2.5''$. Is the height of a randomly chosen person normally distributed?

Gaussian mixture models (GMMs)

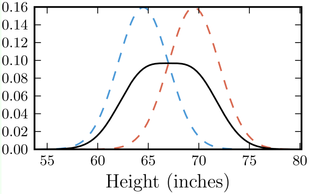
For example, the height of a randomly chosen man is normally distributed with a mean around $5'9.5''$ and standard deviation around $2.5''$. Similarly, the height of a randomly chosen woman is normally distributed with a mean around $5'4.5''$ and standard deviation around $2.5''$. Is the height of a randomly chosen person normally distributed?



Gaussian mixture models (GMMs)

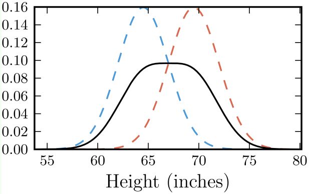
For example, the height of a randomly chosen man is normally distributed with a mean around $5'9.5''$ and standard deviation around $2.5''$. Similarly, the height of a randomly chosen woman is normally distributed with a mean around $5'4.5''$ and standard deviation around $2.5''$. Is the height of a randomly chosen person normally distributed?

The answer is no.



Gaussian mixture models (GMMs)

For example, the height of a randomly chosen man is normally distributed with a mean around $5'9.5''$ and standard deviation around $2.5''$. Similarly, the height of a randomly chosen woman is normally distributed with a mean around $5'4.5''$ and standard deviation around $2.5''$. Is the height of a randomly chosen person normally distributed?



The answer is no. This one is a little more deceptive: because there's so much overlap between the height distributions for men and for women, the overall distribution is in fact highest at the center. But it's still not normally distributed: it's too wide and at in the center (we'll formalize this idea in just a moment).

GMM

Formally, suppose we have people numbered $i = 1, \cdots, n$. We observe r.v. $Y_i \in \mathcal{R}$ for each person's height, and assume there is an unobserved label $C_i \in \{M, F\}$ for each person representing that person's gender, where c stands for "class". Assume that the two groups have the same known variance σ^2, but different unknown means μ_M and μ_F.

GMM

Formally, suppose we have people numbered $i = 1, \cdots, n$. We observe r.v. $Y_i \in \mathcal{R}$ for each person's height, and assume there is an unobserved label $C_i \in \{M, F\}$ for each person representing that person's gender, where c stands for "class". Assume that the two groups have the same known variance σ^2, but different unknown means μ_M and μ_F. The distribution for the class labels is Bernoulli:

$$P(c_i) = q^{I_{c_i=M}}(1-q)^{I_{c_i=F}}.$$

GMM

Formally, suppose we have people numbered $i = 1, \cdots, n$. We observe r.v. $Y_i \in \mathcal{R}$ for each person's height, and assume there is an unobserved label $C_i \in \{M, F\}$ for each person representing that person's gender, where c stands for "class". Assume that the two groups have the same known variance σ^2, but different unknown means μ_M and μ_F. The distribution for the class labels is Bernoulli:

$$P(c_i) = q^{I_{c_i=M}}(1-q)^{I_{c_i=F}}.$$

Let q be known. To simplify notation later, let $\pi_M = q$ and $\pi_F = 1-q$, then

$$P(c_i) = \prod_{c \in \{M,F\}} \pi_c^{I_{c_i=c}}.$$

GMM

Formally, suppose we have people numbered $i = 1, \cdots, n$. We observe r.v. $Y_i \in \mathcal{R}$ for each person's height, and assume there is an unobserved label $C_i \in \{M, F\}$ for each person representing that person's gender, where c stands for "class". Assume that the two groups have the same known variance σ^2, but different unknown means μ_M and μ_F. The distribution for the class labels is Bernoulli:

$$P(c_i) = q^{I_{c_i=M}}(1-q)^{I_{c_i=F}}.$$

Let q be known. To simplify notation later, let $\pi_M = q$ and $\pi_F = 1-q$, then

$$P(c_i) = \prod_{c \in \{M,F\}} \pi_c^{I_{c_i=c}}.$$

The conditional distribution within each class is Gaussian:

$$P(y_i|c_i) = \prod_c N(y_i|\mu_c, \sigma^2)^{I_{c_i=c}}.$$

Solution for GMM: First attempt

Suppose we observe i.i.d. heights $\mathbf{y} = (y_1, \cdots, y_n)$, and we attempt to find the maximum likelihood estimates for the parameters μ_M and μ_F.

Solution for GMM: First attempt

Suppose we observe i.i.d. heights $\mathbf{y} = (y_1, \cdots, y_n)$, and we attempt to find the maximum likelihood estimates for the parameters μ_M and μ_F. We start with the density for a single data point $Y_i = y_i$:

$$P(y_i) = \sum_{c_i} P(y_i|c_i)P(c_i) = \sum_{c_i} \prod_{c \in \{M,F\}} (\pi_c N(y_i|\mu_c, \sigma^2))^{I_{c_i=c}}$$

Solution for GMM: First attempt

Suppose we observe i.i.d. heights $\mathbf{y} = (y_1, \cdots, y_n)$, and we attempt to find the maximum likelihood estimates for the parameters μ_M and μ_F. We start with the density for a single data point $Y_i = y_i$:

$$P(y_i) = \sum_{c_i} P(y_i|c_i)P(c_i) = \sum_{c_i} \prod_{c \in \{M,F\}} (\pi_c N(y_i|\mu_c, \sigma^2))^{I_{c_i=c}}$$

The log-likelihood function is

$$\log \prod_{i=1}^{n} P(y_i) = \sum_{i=1}^{n} \log \sum_{c_i} \prod_{c \in \{M,F\}} (\pi_c N(y_i|\mu_c, \sigma^2))^{I_{c_i=c}}$$

Solution for GMM: First attempt

Suppose we observe i.i.d. heights $\mathbf{y} = (y_1, \cdots, y_n)$, and we attempt to find the maximum likelihood estimates for the parameters μ_M and μ_F. We start with the density for a single data point $Y_i = y_i$:

$$P(y_i) = \sum_{c_i} P(y_i|c_i)P(c_i) = \sum_{c_i} \prod_{c \in \{M,F\}} (\pi_c N(y_i|\mu_c, \sigma^2))^{I_{c_i=c}}$$

The log-likelihood function is

$$\log \prod_{i=1}^{n} P(y_i) = \sum_{i=1}^{n} \log \sum_{c_i} \prod_{c \in \{M,F\}} (\pi_c N(y_i|\mu_c, \sigma^2))^{I_{c_i=c}}$$

Differentiating w.r.t. to μ_M, we obtain

$$\sum_{i=1}^{n} \frac{1}{\sum_{c_i} \prod_{c \in \{M,F\}} (\pi_c N(y_i|\mu_c, \sigma^2))^{I_{c_i=c}}} \pi_M N(y_i|\mu_M, \sigma^2) \frac{y_i - \mu_M}{\sigma^2} = 0$$

Solution for GMM: First attempt

Suppose we observe i.i.d. heights $\mathbf{y} = (y_1, \cdots, y_n)$, and we attempt to find the maximum likelihood estimates for the parameters μ_M and μ_F. We start with the density for a single data point $Y_i = y_i$:

$$P(y_i) = \sum_{c_i} P(y_i|c_i)P(c_i) = \sum_{c_i} \prod_{c \in \{M,F\}} (\pi_c N(y_i|\mu_c, \sigma^2))^{I_{c_i=c}}$$

The log-likelihood function is

$$\log \prod_{i=1}^{n} P(y_i) = \sum_{i=1}^{n} \log \sum_{c_i} \prod_{c \in \{M,F\}} (\pi_c N(y_i|\mu_c, \sigma^2))^{I_{c_i=c}}$$

Differentiating w.r.t. to μ_M, we obtain

$$\sum_{i=1}^{n} \frac{1}{\sum_{c_i} \prod_{c \in \{M,F\}} (\pi_c N(y_i|\mu_c, \sigma^2))^{I_{c_i=c}}} \pi_M N(y_i|\mu_M, \sigma^2) \frac{y_i - \mu_M}{\sigma^2} = 0$$

There is no way we can solve this in closed form.

Outline

Solution of BMMs

Now consider a more challenging variant of the parameter estimation problem in which we are given the recorded head counts x but not the identities z of the coins used for each set of tosses.

Solution of BMMs

Now consider a more challenging variant of the parameter estimation problem in which we are given the recorded head counts x but not the identities z of the coins used for each set of tosses. This time, computing proportions of heads for each coin is no longer possible, because we do not know the coin used for each set of tosses.

Solution of BMMs

Now consider a more challenging variant of the parameter estimation problem in which we are given the recorded head counts x but not the identities z of the coins used for each set of tosses. This time, computing proportions of heads for each coin is no longer possible, because we do not know the coin used for each set of tosses.

- We refer to z as hidden variables or latent factors.

Solution of BMMs

Now consider a more challenging variant of the parameter estimation problem in which we are given the recorded head counts x but not the identities z of the coins used for each set of tosses. This time, computing proportions of heads for each coin is no longer possible, because we do not know the coin used for each set of tosses.

- We refer to z as hidden variables or latent factors.
- Parameter estimation in this new setting is known as the incomplete data case.

Solution of BMMs

Now consider a more challenging variant of the parameter estimation problem in which we are given the recorded head counts x but not the identities z of the coins used for each set of tosses. This time, computing proportions of heads for each coin is no longer possible, because we do not know the coin used for each set of tosses.

- We refer to z as hidden variables or latent factors.
- Parameter estimation in this new setting is known as the incomplete data case.

However, if we had some way of completing the data (in our case, guessing correctly which coin was used in each of the five sets), then we could reduce parameter estimation for this problem with incomplete data to maximum likelihood estimation with complete data.

Iterative scheme design

We start from some initial parameters $\widehat{\theta}^{(t)} = (\widehat{\theta}_1^{(t)}, \widehat{\theta}_2^{(t)})$:

Iterative scheme design

We start from some initial parameters $\widehat{\theta}^{(t)} = (\widehat{\theta}_1^{(t)}, \widehat{\theta}_2^{(t)})$:

- Determine for each of the five sets whether coin A or coin B was more likely to have generated the observed flips (using the current parameter estimates);

Iterative scheme design

We start from some initial parameters $\widehat{\theta}^{(t)} = (\widehat{\theta}_1^{(t)}, \widehat{\theta}_2^{(t)})$:

- Determine for each of the five sets whether coin A or coin B was more likely to have generated the observed flips (using the current parameter estimates);
- Then, assume these completions (that is, guessed coin assignments) to be correct, and apply the regular maximum likelihood estimation procedure to get $\widehat{\theta}^{(t+1)}$;

Iterative scheme design

We start from some initial parameters $\widehat{\theta}^{(t)} = (\widehat{\theta}_1^{(t)}, \widehat{\theta}_2^{(t)})$:

- Determine for each of the five sets whether coin A or coin B was more likely to have generated the observed flips (using the current parameter estimates);
- Then, assume these completions (that is, guessed coin assignments) to be correct, and apply the regular maximum likelihood estimation procedure to get $\widehat{\theta}^{(t+1)}$;
- Finally, repeat these two steps until convergence. As the estimated model improves, so too will the quality of the resulting completions.

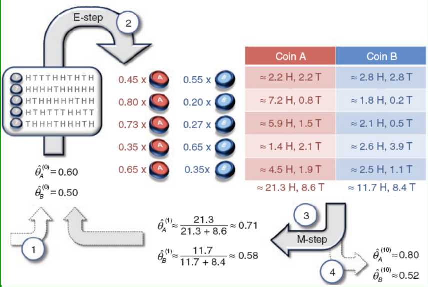
Iterative scheme design

We start from some initial parameters $\widehat{\theta}^{(t)} = (\widehat{\theta}_1^{(t)}, \widehat{\theta}_2^{(t)})$:

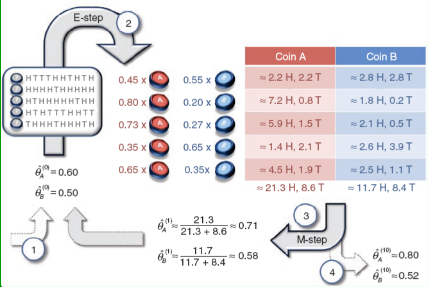
- Determine for each of the five sets whether coin A or coin B was more likely to have generated the observed flips (using the current parameter estimates);
- Then, assume these completions (that is, guessed coin assignments) to be correct, and apply the regular maximum likelihood estimation procedure to get $\widehat{\theta}^{(t+1)}$;
- Finally, repeat these two steps until convergence. As the estimated model improves, so too will the quality of the resulting completions.

The expectation maximization algorithm (EM algorithm) is a refinement on this basic idea.

Refinement


E-step
2
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
0.45 x A
0.55 x B
0.80 x A
0.20 x B
0.73 x A
0.27 x B
0.35 x A
0.65 x B
0.65 x A
0.35x B
Coin A
Coin B
≈ 2.2 H, 2.2 T
≈ 2.8 H, 2.8 T
≈ 7.2 H, 0.8 T
≈ 1.8 H, 0.2 T
≈ 5.9 H, 1.5 T
≈ 2.1 H, 0.5 T
≈ 1.4 H, 2.1 T
≈ 2.6 H, 3.9 T
≈ 4.5 H, 1.9 T
≈ 2.5 H, 1.1 T
≈ 21.3 H, 8.6 T
≈ 11.7 H, 8.4 T
$\hat{\theta}_A^{(0)} = 0.60$
$\hat{\theta}_B^{(0)} = 0.50$
1
$\hat{\theta}_A^{(1)} \approx \frac{21.3}{21.3 + 8.6} \approx 0.71$
$\hat{\theta}_B^{(1)} \approx \frac{11.7}{11.7 + 8.4} \approx 0.58$
3
M-step
4
$\hat{\theta}_A^{(10)} \approx 0.80$
$\hat{\theta}_B^{(10)} \approx 0.52$

Refinement



Let Y be each set of coin tossing.

$$P(A|Y)$$

$$= \frac{P(Y|A)P(A)}{P(Y|A)P(A) + P(Y|B)P(B)}$$

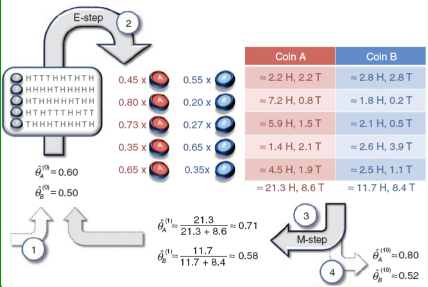
$$= \frac{0.24^5 \cdot P(A)}{0.24^5 \cdot P(A) + 0.25^5 \cdot P(B)}$$

$$= \frac{7962624}{7962624 + 9765625} \approx 0.45$$

$$P(B|Y) = 1 - P(A|Y) = 0.55$$

$$P(A) = \frac{\sum_i P(A|Y_i)}{N}$$

Refinement



Let Y be each set of coin tossing.

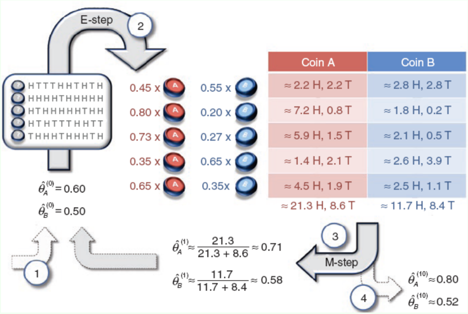
$$
\begin{aligned}
P(A|Y) &= \frac{P(Y|A)P(A)}{P(Y|A)P(A) + P(Y|B)P(B)} \\
&= \frac{0.24^5 \cdot P(A)}{0.24^5 \cdot P(A) + 0.25^5 \cdot P(B)} \\
&= \frac{7962624}{7962624 + 9765625} \approx 0.45
\end{aligned}
$$

$$P(B|Y) = 1 - P(A|Y) = 0.55$$

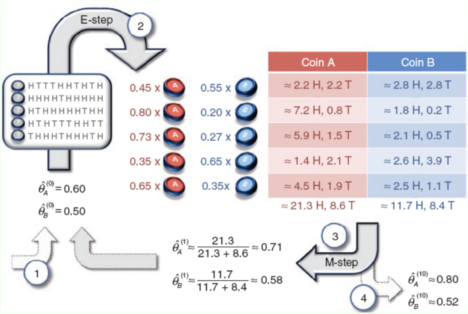
$$P(A) = \frac{\sum_i P(A|Y_i)}{N}$$

Rather than picking the single most likely completion of the missing coin assignments, EM algorithm refines the completion.

Refinement Cont'd

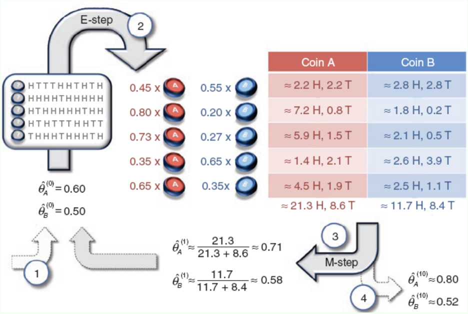


Refinement Cont'd



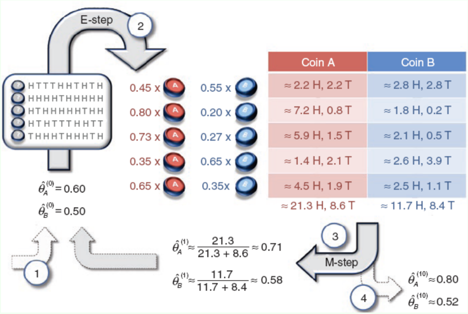
- EM algorithm computes probabilities for each possible completion of the missing data, using $\widehat{\theta}^{(t)}$;

Refinement Cont'd



- EM algorithm computes probabilities for each possible completion of the missing data, using $\widehat{\theta}^{(t)}$;
- These probabilities are used to create a weighted training set;

Refinement Cont'd



- EM algorithm computes probabilities for each possible completion of the missing data, using $\widehat{\theta}^{(t)}$;
- These probabilities are used to create a weighted training set;
- Finally, EM algorithm deals with weighted training examples provides new parameter estimates, $\widehat{\theta}^{(t+1)}$.

EM algorithm

In using the EM algorithm we consider two different likelihood problems

- "Incomplete-data" problem;
- Latent variable problem.

EM algorithm

In using the EM algorithm we consider two different likelihood problems

- "Incomplete-data" problem;
- Latent variable problem.

Filtering and smoothing EM algorithms arise by repeating this two-step procedure:

- **E-step**: Determine the conditional expectation

$$E_{Z|\theta_n}(l(\theta|\mathbf{X}))$$

- **M-step**: Maximize this expression w.r.t. θ.

$$\theta_{n+1} = argmax_\theta E_{Z|\theta_n}(l(\theta|\mathbf{X})).$$

Solution for GMM: EM algorithm

If we know the latent labels C_i exactly, then it would be easy to obtain the MLEs.

Solution for GMM: EM algorithm

If we know the latent labels C_i exactly, then it would be easy to obtain the MLEs.

$$P(y_i) = (P(y_i|M)P(M))^{I_{C_i=M}}(P(y_i|F)P(F))^{I_{C_i=F}}$$

Solution for GMM: EM algorithm

If we know the latent labels C_i exactly, then it would be easy to obtain the MLEs.

$$P(y_i) = (P(y_i|M)P(M))^{I_{C_i=M}}(P(y_i|F)P(F))^{I_{C_i=F}}$$

The log-likelihood function is

$$\log L(\mu_M, \mu_F|y_i) = \sum_{i=1}^{n} \left[I_{C_i=M} \log P(y_i|M)P(M) + I_{C_i=F} \log P(y_i|F)P(F) \right]$$

Solution for GMM: EM algorithm

If we know the latent labels C_i exactly, then it would be easy to obtain the MLEs.

$$P(y_i) = (P(y_i|M)P(M))^{I_{C_i=M}}(P(y_i|F)P(F))^{I_{C_i=F}}$$

The log-likelihood function is

$$\log L(\mu_M, \mu_F|y_i) = \sum_{i=1}^{n} \left[I_{C_i=M} \log P(y_i|M)P(M) + I_{C_i=F} \log P(y_i|F)P(F) \right]$$

Note that $I_{C_i=M}$ is a r.v.

Solution for GMM: EM algorithm

If we know the latent labels C_i exactly, then it would be easy to obtain the MLEs.

$$P(y_i) = (P(y_i|M)P(M))^{I_{C_i=M}}(P(y_i|F)P(F))^{I_{C_i=F}}$$

The log-likelihood function is

$$\log L(\mu_M, \mu_F|y_i) = \sum_{i=1}^{n} \left[I_{C_i=M} \log P(y_i|M)P(M) + I_{C_i=F} \log P(y_i|F)P(F) \right]$$

Note that $I_{C_i=M}$ is a r.v. Let's look at the $P(C_i = M|y_i)$

$$\begin{aligned} P(C_i = M|y_i) &= \frac{P(y_i|C_i = M)P(C_i = M)}{P(y_i|C_i = M)P(C_i = M) + P(y_i|C_i = F)P(C_i = F)} \\ &= \frac{\pi_M N(y_i|\mu_M, \sigma^2)}{\pi_M N(y_i|\mu_M, \sigma^2) + \pi_F N(y_i|\mu_F, \sigma^2)} \doteq q(M) \end{aligned}$$

Solution for GMM: EM algorithm Cont'd

- Initialize the values of $\mu_M = 0$ and $\mu_F = 0$;

Solution for GMM: EM algorithm Cont'd

- Initialize the values of $\mu_M = 0$ and $\mu_F = 0$;
- E-step:

$$\begin{aligned} q(M)^{(k)} &= P(C_i = M | y_i, \mu^{(k-1)}) \\ &= \frac{\pi_M N(y_i | \mu_M^{(k-1)}, \sigma^2)}{\pi_M N(y_i | \mu_M^{(k-1)}, \sigma^2) + \pi_F N(y_i | \mu_F^{(k-1)}, \sigma^2)} \end{aligned}$$

Solution for GMM: EM algorithm Cont'd

- Initialize the values of $\mu_M = 0$ and $\mu_F = 0$;
- E-step:

$$\begin{aligned} q(M)^{(k)} &= P(C_i = M | y_i, \mu^{(k-1)}) \\ &= \frac{\pi_M N(y_i | \mu_M^{(k-1)}, \sigma^2)}{\pi_M N(y_i | \mu_M^{(k-1)}, \sigma^2) + \pi_F N(y_i | \mu_F^{(k-1)}, \sigma^2)} \end{aligned}$$

- M-step:

$$\frac{\partial}{\partial \mu_c} \log L(\mu_M, \mu_F | y_i) = \sum_{i=1}^{n} q(c)^{(k)} \frac{y_i - \mu_c}{\sigma^2} = 0$$

Solution for GMM: EM algorithm Cont'd

- Initialize the values of $\mu_M = 0$ and $\mu_F = 0$;
- E-step:

$$\begin{aligned} q(M)^{(k)} &= P(C_i = M | y_i, \mu^{(k-1)}) \\ &= \frac{\pi_M N(y_i|\mu_M^{(k-1)}, \sigma^2)}{\pi_M N(y_i|\mu_M^{(k-1)}, \sigma^2) + \pi_F N(y_i|\mu_F^{(k-1)}, \sigma^2)} \end{aligned}$$

- M-step:

$$\frac{\partial}{\partial \mu_c} \log L(\mu_M, \mu_F | y_i) = \sum_{i=1}^{n} q(c)^{(k)} \frac{y_i - \mu_c}{\sigma^2} = 0$$

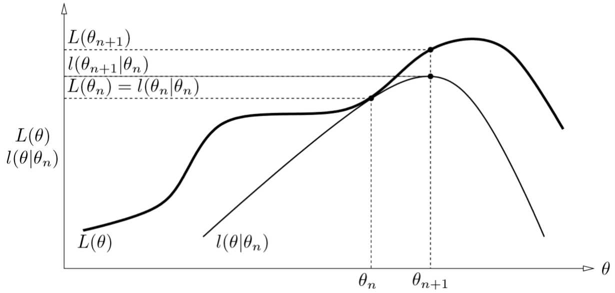
That is,

$$\mu_M^{(k)} = \frac{\sum_{i=1}^n q(M)^{(k)} y_i}{\sum_{i=1}^n q(M)^{(k)}}, \mu_F^{(k)} = \frac{\sum_{i=1}^n (1 - q(F)^{(k)}) y_i}{\sum_{i=1}^n (1 - q(F)^{(k)})}$$

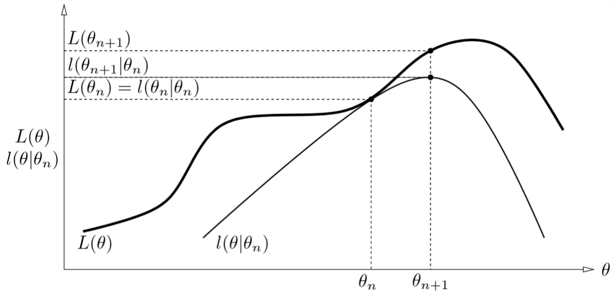
Convergence of EM algorithm

$$
\begin{aligned}
l(\theta|\mathbf{x}) &= \sum_{i=1}^{n} \log p(x_i;\theta) = \sum_{i=1}^{n} \log \sum_z p(x_i, z;\theta) \\
&= \sum_{i=1}^{n} \log \sum_z p(z;\theta_n) \frac{p(x_i, z;\theta)}{p(z;\theta_n)} \\
&\geq \sum_{i=1}^{n} \sum_z p(z;\theta_n) \log \frac{p(x_i, z;\theta)}{p(z;\theta_n)} \\
&= \sum_{i=1}^{n} \sum_z p(z;\theta_n) \log p(x_i, z;\theta) - n \sum_z p(z;\theta_n) \log p(z;\theta_n) \\
&= Q(\theta|\theta_n) + H(P(z;\theta_n)) \equiv F(p(z;\theta_n), \theta) \\
& Q(\theta_{n+1}|\theta_n) = Q(\theta_n|\theta_n)
\end{aligned}
$$

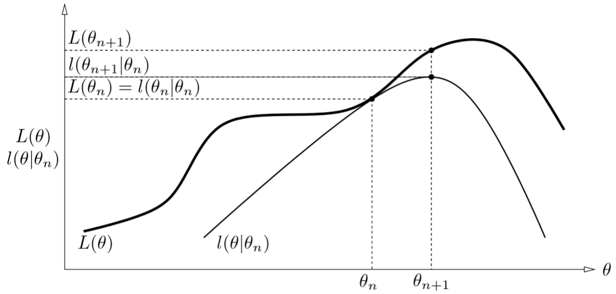
Convergence of EM algorithm Cont'd



Convergence of EM algorithm Cont'd

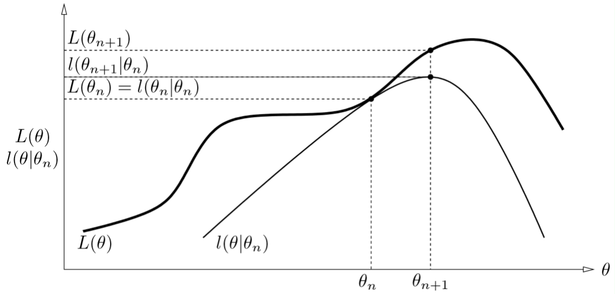


Convergence of EM algorithm Cont'd



- The EM Algorithm always improves a parameter's estimation through this multi-step process;

Convergence of EM algorithm Cont'd



- The EM Algorithm always improves a parameter's estimation through this multi-step process;
- However, it sometimes needs a few random starts to find the best model because the algorithm can hone in on a local maxima that is not that close to the (optimal) global maxima.

MLE VS. EM algorithm

Although Maximum Likelihood Estimation (MLE) and EM can both find "best-fit" parameters, how they find the models are very different.

MLE VS. EM algorithm

Although Maximum Likelihood Estimation (MLE) and EM can both find "best-fit" parameters, how they find the models are very different.

- MLE accumulates all of the data first and then uses that data to construct the most likely model;

MLE VS. EM algorithm

Although Maximum Likelihood Estimation (MLE) and EM can both find "best-fit" parameters, how they find the models are very different.

- MLE accumulates all of the data first and then uses that data to construct the most likely model;
- EM takes a guess at the parameters first accounting for the missing data then tweaks the model to fit the guesses and the observed data.

MLE VS. EM algorithm

Although Maximum Likelihood Estimation (MLE) and EM can both find "best-fit" parameters, how they find the models are very different.

- MLE accumulates all of the data first and then uses that data to construct the most likely model;
- EM takes a guess at the parameters first accounting for the missing data then tweaks the model to fit the guesses and the observed data.
  - An initial guess is made for the model's parameters and a probability distribution is created (E-Step);

MLE VS. EM algorithm

Although Maximum Likelihood Estimation (MLE) and EM can both find "best-fit" parameters, how they find the models are very different.

- MLE accumulates all of the data first and then uses that data to construct the most likely model;
- EM takes a guess at the parameters first accounting for the missing data then tweaks the model to fit the guesses and the observed data.
    - An initial guess is made for the model's parameters and a probability distribution is created (E-Step);
    - Newly observed data is fed into the model;

MLE VS. EM algorithm

Although Maximum Likelihood Estimation (MLE) and EM can both find "best-fit" parameters, how they find the models are very different.

- MLE accumulates all of the data first and then uses that data to construct the most likely model;
- EM takes a guess at the parameters first accounting for the missing data then tweaks the model to fit the guesses and the observed data.
  - An initial guess is made for the model's parameters and a probability distribution is created (E-Step);
  - Newly observed data is fed into the model;
  - The probability distribution from the E-step is tweaked to include the new data (M-step);

MLE VS. EM algorithm

Although Maximum Likelihood Estimation (MLE) and EM can both find "best-fit" parameters, how they find the models are very different.

- MLE accumulates all of the data first and then uses that data to construct the most likely model;
- EM takes a guess at the parameters first accounting for the missing data then tweaks the model to fit the guesses and the observed data.
  - An initial guess is made for the model's parameters and a probability distribution is created (E-Step);
  - Newly observed data is fed into the model;
  - The probability distribution from the E-step is tweaked to include the new data (M-step);
  - Steps 2 through 4 are repeated until stability

https://www.cs.utah.edu/~piyush/teaching/EM_
algori

Take-home messages

- Motivation
- MLE
- EM Algorithm