Lab 6 - EM

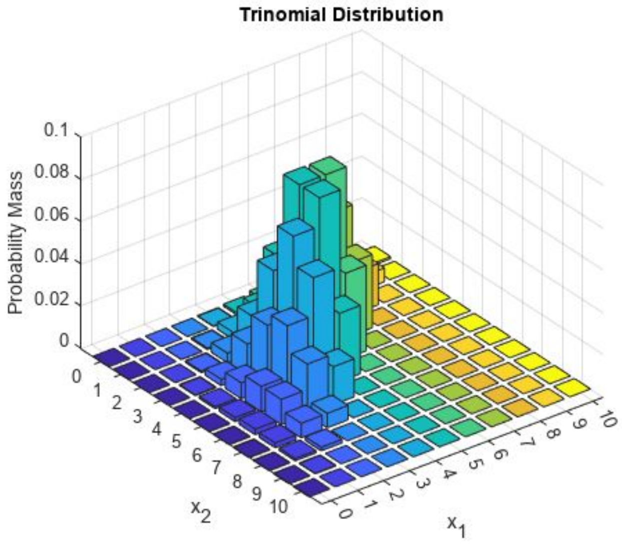
Expectation Maximization

Multinomial Distribution(n, π)

Example: Rolling a fair dice n=60 times.

Probabilities, π = (⅙, ⅙, ⅙, ⅙, ⅙, ⅙)

Observed data, X = (10, 9, 11, 10, 8, 12)



Multinomial Distribution(n, π)

Observed data, x = (x1,...,xm)

Constraint: x1 + … + xm = n

Probabilities, π = (θ1, …, θm)

Constraint: θ1 +...+ θm = 1

Multinomial PMF:

$$PMF = \frac{n!}{\Pi_{i=1}^{n} x_i!} \Pi_{i=1}^{n} (p(x_i))^{x_i}$$

Consider the following problem

Suppose X = (200, 34, 38, 98) is a sample from Mult(n=370, $\boldsymbol{\pi}$)

$$\pi_\theta \;=\; \left(\frac{1}{2}+\frac{1}{4}\theta,\; \frac{1}{4}(1-\theta),\; \frac{1}{4}(1-\theta),\; \frac{1}{4}\theta\right).$$

We want to solve for the value of theta.

Consider the following problem

Suppose X = (200, 34, 38, 98) is a sample from Mult(n=370, $\boldsymbol{\pi}$)

$$\pi_\theta = \left(\frac{1}{2} + \frac{1}{4}\theta,\ \frac{1}{4}(1-\theta),\ \frac{1}{4}(1-\theta),\ \frac{1}{4}\theta\right).$$

We want to solve for the value of theta.

Note: we can find the MLE in this case without using EM

MLE of multinomial

The likelihood, $L(\theta; \mathbf{x})$, is then given by

$$L(\theta; \mathbf{x}) = \frac{n!}{x_1!x_2!x_3!x_4!} \left(\frac{1}{2} + \frac{1}{4}\theta\right)^{x_1} \left(\frac{1}{4}(1-\theta)\right)^{x_2} \left(\frac{1}{4}(1-\theta)\right)^{x_3} \left(\frac{1}{4}\theta\right)^{x_4}$$

so that the log-likelihood $l(\theta; \mathbf{x})$ is

$$l(\theta; \mathbf{x}) \; = \; C \; + \; x_1 \ln\left(\frac{1}{2} + \frac{1}{4}\theta\right) \; + \; (x_2 + x_3)\ln(1-\theta) \; + \; x_4 \ln(\theta)$$

Final step is to solve $\frac{dl}{d\theta} = 0$

Remember that X = (200, 34, 38, 98)

Slight change to the problem

Original

Suppose X = (x1 = 200, x2 = 34, x3 =3
8, x4 = 98) is a sample from
Mult(370, $\boldsymbol{\pi}$)

$$\pi_\theta = \left(\frac{1}{2} + \frac{1}{4}\theta,\ \frac{1}{4}(1-\theta),\ \frac{1}{4}(1-\theta),\ \frac{1}{4}\theta\right).$$

We want to solve for the value of theta.

New Problem:

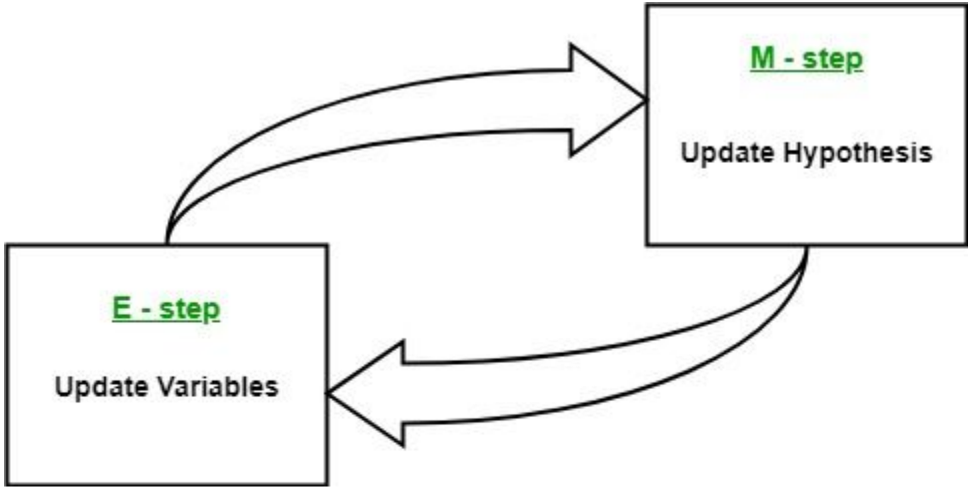
Suppose Y = (y1, y2, y3 = 34, y4 = 38, y5 = 98) is a sample from Mult(n=370, $\boldsymbol{\pi}$), but we only observe X (original problem)

(y1 + y2 = 200)

$$\pi_\theta^* = \left(\frac{1}{2},\ \frac{1}{4}\theta,\ \frac{1}{4}(1-\theta),\ \frac{1}{4}(1-\theta),\ \frac{1}{4}\theta\right).$$

Classical EM

EM algorithm is used for obtaining MLEs of parameters when some of the data is *missing* or *unobserved*.



E Step

Given the statistical model which generates a set X of observed data, a set of unobserved latent data Y, and a vector of unknown parameters θ, along with the log likelihood function $l(\theta; X, Y)$.

E-Step: The E-step of the EM algorithm computes the expected value of $l(\theta; \mathcal{X}, \mathcal{Y})$ given the observed data, $\mathcal{X}$, and the current parameter estimate, θ_{old} say. In particular, we define

$$\begin{aligned} Q(\theta; \theta_{old}) &:= \mathsf{E}\left[l(\theta; \mathcal{X}, \mathcal{Y}) \mid \mathcal{X}, \theta_{old}\right] \\ &= \int l(\theta; \mathcal{X}, y)\, p(y \mid \mathcal{X}, \theta_{old})\, dy \end{aligned} \tag{1}$$

where $p(\cdot \mid \mathcal{X}, \theta_{old})$ is the conditional density of $\mathcal{Y}$ given the observed data, $\mathcal{X}$, and assuming $\theta = \theta_{old}$.

Y - Problem Likelihood

$$\mathcal{L}(\theta; \mathcal{X}, \mathcal{Y}) = \frac{n!}{y_1!y_2!y_3!y_4!y_5!}(\frac{1}{2})^{y_1}(\frac{1}{4}\theta)^{y_2}(\frac{1}{4}(1-\theta))^{y_3}(\frac{1}{4}(1-\theta))^{y_4}(\frac{1}{4}\theta)^{y_5}$$

$$l(\theta; \mathcal{X}, \mathcal{Y}) \ = \ C \ + \ y_2 \ln(\theta) \ + \ (y_3 + y_4) \ln(1-\theta) \ + \ y_5 \ln(\theta)$$

E Step

$$l(\theta;\mathcal{X},\mathcal{Y}) \;=\; C \;+\; y_2 \ln(\theta) \;+\; (y_3+y_4)\ln(1-\theta) \;+\; y_5 \ln(\theta)$$

E-Step: Recalling that $Q(\theta;\theta_{old}) := \mathsf{E}\left[l(\theta;\mathcal{X},\mathcal{Y}) \mid \mathcal{X},\theta_{old}\right]$, we have

$$Q(\theta;\theta_{old}) \;:=\; C \;+\; \mathsf{E}\left[y_2 \ln(\theta) \mid \mathcal{X},\theta_{old}\right] \;+\; (y_3+y_4)\ln(1-\theta) \;+\; y_5 \ln(\theta)$$

E Step

$$l(\theta; \mathcal{X}, \mathcal{Y}) \; = \; C \; + \; y_2 \ln(\theta) \; + \; (y_3 + y_4) \ln(1-\theta) \; + \; y_5 \ln(\theta)$$

E-Step: Recalling that $Q(\theta; \theta_{old}) := \mathsf{E}\left[l(\theta; \mathcal{X}, \mathcal{Y}) \mid \mathcal{X}, \theta_{old}\right]$, we have

$$Q(\theta; \theta_{old}) \; := \; C \; + \; \mathsf{E}\left[y_2 \ln(\theta) \mid \mathcal{X}, \theta_{old}\right] \; + \; (y_3 + y_4) \ln(1-\theta) \; + \; y_5 \ln(\theta)$$

Y = (y1, y2, y3 = 34, y4 = 38, y5 = 98)

Why is the expectation of the term with y2?

Expected value of y2 term

Given n=370 and Y = (y1, y2, y3 = 34, y4 = 38, y5 = 98), y1+y2 = 200

$$\pi_\theta^* = \left(\frac{1}{2}, \frac{1}{4}\theta, \frac{1}{4}(1-\theta), \frac{1}{4}(1-\theta), \frac{1}{4}\theta\right).$$

We can think of choosing between y2 and y1 as a binomial distribution. (Y = y2 in the formula below)

$$f(\mathcal{Y} \mid \mathcal{X}, \theta) = \text{Bin}\left(y_1 + y_2, \frac{\theta/4}{1/2 + \theta/4}\right).$$

E Step Complete

E-Step: Recalling that $Q(\theta;\theta_{old}) := \mathsf{E}\left[l(\theta;\mathcal{X},\mathcal{Y}) \mid \mathcal{X},\theta_{old}\right]$, we have

$$\begin{aligned} Q(\theta;\theta_{old}) &:= C + \mathsf{E}\left[y_2 \ln(\theta) \mid \mathcal{X},\theta_{old}\right] + (y_3+y_4)\ln(1-\theta) + y_5 \ln(\theta) \\ &= C + (y_1+y_2)p_{old}\ln(\theta) + (y_3+y_4)\ln(1-\theta) + y_5\ln(\theta) \end{aligned}$$

where

$$p_{old} := \frac{\theta_{old}/4}{1/2+\theta_{old}/4}.$$

Notice the difference between theta_old and theta

M Step

M-Step: The M-step consists of maximizing over θ the expectation computed in (1). That is, we set

$$\theta_{new} \ := \ \max_{\theta} \ Q(\theta; \theta_{old}).$$

We then set $\theta_{old} = \theta_{new}$.

M Step

$$\begin{aligned} Q(\theta;\theta_{old}) &:= C + \mathsf{E}\left[y_2 \ln(\theta) \mid \mathcal{X}, \theta_{old}\right] + (y_3+y_4)\ln(1-\theta) + y_5 \ln(\theta) \\ &= C + (y_1+y_2)p_{old}\ln(\theta) + (y_3+y_4)\ln(1-\theta) + y_5\ln(\theta) \end{aligned}$$

M-Step: We now maximize $Q(\theta;\theta_{old})$ to find θ_{new}. Taking the derivative we obtain

$$p_{old} := \frac{\theta_{old}/4}{1/2+\theta_{old}/4}.$$

$$\frac{dQ}{d\theta} = \frac{(y_1+y_2)}{\theta}p_{old} - \frac{(y_3+y_4)}{1-\theta} + \frac{y_5}{\theta}$$

which is zero when we take $\theta = \theta_{new}$ where

$$\theta_{new} := \frac{y_5 + p_{old}(y_1+y_2)}{y_3+y_4+y_5+p_{old}(y_1+y_2)}$$

Writing Code to find Theta

1. Start with initial guess of theta_old

2. From E Step: $$p_{old} := \frac{\theta_{old}/4}{1/2 + \theta_{old}/4}.$$

3. From M Step: solve for theta_new

$$\theta_{new} \ := \ \frac{y_5 + p_{old}(y_1 + y_2)}{y_3 + y_4 + y_5 + p_{old}(y_1 + y_2)}$$

Repeat 2-3 for N iterations

Notice the following

From M-step (update rule)

$$\theta_{new} := \frac{y_5 + p_{old}(y_1 + y_2)}{y_3 + y_4 + y_5 + p_{old}(y_1 + y_2)}$$

Our update rule is based on our data and p_old from the E step

$$p_{old} := \frac{\theta_{old}/4}{1/2 + \theta_{old}/4}.$$

If I wanted to code a solution, I only need to code these two functions.

(Note: nowhere in my coding solution do I need to compute the log likelihood probability)

General Takeaways/Tips

1. Start with log-likelihood function for the problem
2. Find Q

$$Q(\theta; \theta_{old}) \quad := \quad \mathsf{E}\left[l(\theta; \mathcal{X}, \mathcal{Y}) \mid \mathcal{X}, \theta_{old}\right]$$

1. Derive update rule from Q for desired parameter(s)