EM as a Maximization-Maximization procedure (Gibbs sampler)

S.V.Harshith - EE19BTECH11018

September 8, 2020

1. Expectation – maximization algorithm
2. Kullback–Leibler divergence
3. EM as a Maximization-Maximization procedure
4. Applications

Expectation–maximization algorithm

In statistics, an expectation–maximization (EM) algorithm is an iterative method to find maximum likelihood or maximum a posteriori (MAP) estimates of parameters in statistical models, where the model depends on unobserved latent variables.
The EM algorithm seeks to find the maximum likelihood estimate(MLE) of the marginal likelihood by iteratively applying these two steps:

Expectation step (E step): Define $Q(\theta|\theta^{(t)})$ as the expected value of the log likelihood function of θ, with respect to the current conditional distribution of Z given X and the current estimates of the parameters $\theta^{(t)}$:

$$Q(\theta|\theta^{(t)}) = E_{Z|X,\theta^{(t)}}[logL(\theta; X, Z)]$$

Maximization step (M step): Find the parameters that maximize this quantity:

$$\theta^{(t+1)} = arg_\theta max Q\left(\theta|\theta^{(t)}\right)$$

Kullback–Leibler divergence

In mathematical statistics, the Kullback–Leibler divergence (also called relative entropy) is a measure of how one probability distribution is different from a second, reference probability distribution.
For discrete probability distributions P and Q defined on the same probability space,X, the Kullback–Leibler divergence of Q from P is defined to be

$$D_{KL}(P||Q) = \sum_{x \in X} P(x) log\left(\frac{P(x)}{Q(x)}\right)$$

EM as a Maximization-Maximization procedure

The EM algorithm can be viewed as two alternating maximization steps, that is, as an example of coordinate ascent.
Consider the function:

$$F(q, \theta) := E_q[logL(\theta; x, Z)] + H(q)$$

where q is an arbitrary probability distribution over the unobserved data z and H(q) is the entropy of the distribution q. This function can be written as

$$F(q, \theta) = -D_{KL}\left(q||p_{Z|X}(\cdot|x;\theta)\right) + logL(\theta; x)$$

where $p_{Z|X}(\cdot|x;\theta)$ is the conditional distribution of the unobserved data given the observed data x and D_{KL} is the Kullback–Leibler divergence.

Then the steps in the EM algorithm may be viewed as:

Expectation step: Choose q to maximize F:

$$q^{(t)} = \mathit{argmax}_{\theta} F\left(q, \theta^{(t)}\right)$$

Maximization step: Choose θ to maximize F:

$$\theta^{(t+1)} = \mathit{argmax}_{\theta} F\left(q^{(t)}, \theta\right)$$

Appliations

EM is frequently used for **data clustering** in machine learning and computer vision.
Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense) to each other than to those in other groups (clusters). It is a main task of exploratory data mining, and a common technique for statistical data analysis, used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, bioinformatics, data compression, and computer graphics.

Example

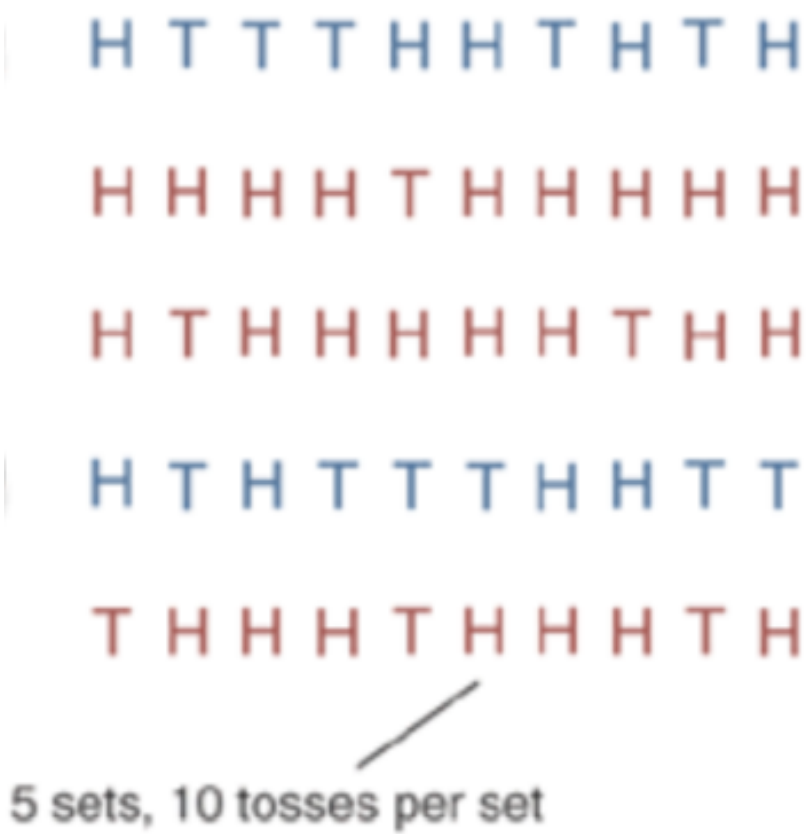
Assume we have 2 coins , C1 and C2.
Let the probability of getting heads with C1 and C2 are θ_1 and θ_2 respectively.
Suppose we choose a coin 5 times and toss it 10 times to get the distribution :

$$\theta_1 = \frac{number\ of\ heads\ using\ C1}{total\ number\ of\ flips\ using\ C1}$$

$$\theta_2 = \frac{number\ of\ heads\ using\ C2}{total\ number\ of\ flips\ using\ C2}$$

From the above formulae, $\theta_1 = 0.8$, $\theta_2 = 0.45$

Suppose that we do not know the identities of the coins used for each set of tosses (we treat them as hidden variables).
The Distribution is :



• Initialize $\theta_A = 0.6$, $\theta_B = 0.5$ (Assume)
• Compute a probability distribution of possible completions of the data using current parameters.

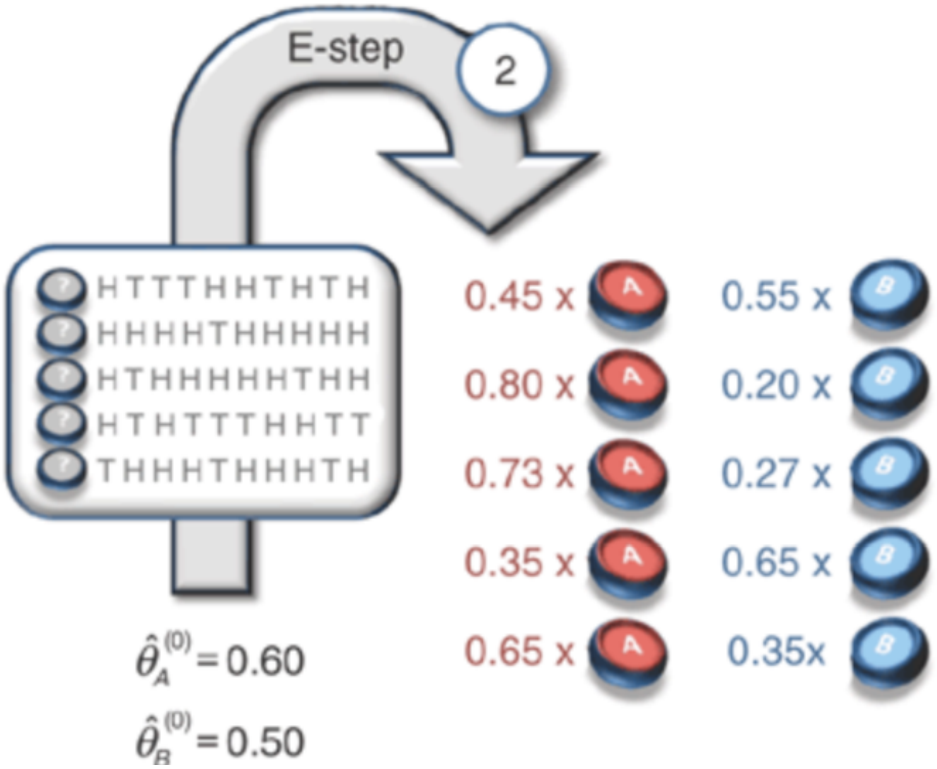
SET-1 : H T T T H H T H T H

• What is the probability that we observe 5 heads and 5 tails in coin A and B given the initializing parameters $\theta_A = 0.6$, $\theta_B = 0.5$?
• Compute likelihood of set 1 coming from coin A or B using the binomial distribution with mean probability on n trials with k successes

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

After Calculating,
Likelihood of A = 0.00079
Likelihood of B = 0.00097
Normalizing A by using (A/A+B), we get A = 0.45
Similarly after normalizing B we get B = 0.55

E - Step



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

M-Step

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

$$\theta_A{}^1 = \frac{21.3}{21.3 + 8.6} = 0.71$$

$$\theta_B{}^1 = \frac{11.7}{11.7 + 8.4} = 0.58$$

Now take the new values of θ_A and θ_B and repeat the process