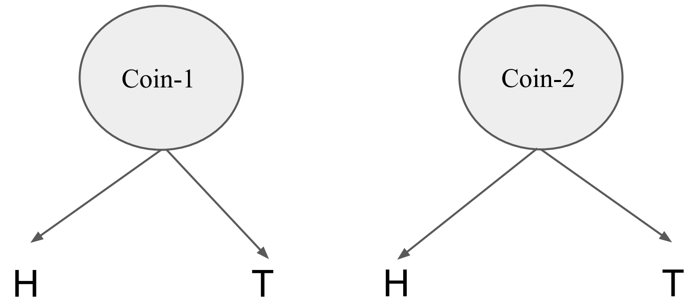
Expectation Maximization (EM)

Course - CS 571

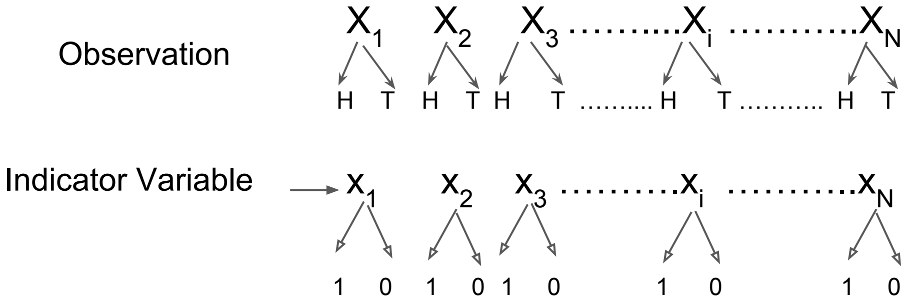
Prof. Pushpak Bhattacharyya
Department of Computer Science & Engineering
Indian Institute of Technology Patna

Expectation Maximization (EM): Introduction

- Choose a coin and toss.
- Repeat it for N number of times.
- **p** = probability of choosing **Coin-1**.
- $\mathbf{p_1}$ = probability of Head from **Coin-1**.
- $\mathbf{p_2}$ = probability of Head from **Coin-2**.
- **Task:** Estimate parameters **p**, $\mathbf{p_1}$ and $\mathbf{p_2}$



- **Sample Observation:**
  N=10
  H H T H T T H T H T
- $x_i \rightarrow$ 1 1 0 1 0 0 1 0 1 0
- M = number of heads = 5



Expectation Maximization

- Estimate parameters p, p_1 and p_2 by Expectation Maximization (EM) Algorithm.

$$\left\{ p=\frac{\sum_{i=1}^{N} E(z_i)}{N} \qquad p_1=\frac{\sum_{i=1}^{N} x_i E(z_i)}{\sum_{i=1}^{N} E(z_i)} \qquad p_2=\frac{M-\sum_{i=1}^{N} x_i E(z_i)}{N-\sum_{i=1}^{N} E(z_i)} \right\}$$

M-Step

- z_i is Hidden variable.
- z_i = 1, if coin-1 is chosen.
- z_i = 0, elsewhere

$$\left\{ E(z_i)=\frac{pp_1^{x_i}(1-p_1)^{1-x_i}}{pp_1^{x_i}(1-p_1)^{1-x_i}+(1-p)p_2^{x_i}(1-p_2)^{1-x_i}} \right\}$$

Expectation Step

Maximum Likelihood Estimation (MLE)

- **Represented by:** f: Data (D) $\longrightarrow$ Hypothesis (H)
- **Aim:** Maximize the likelihood (probability) of the observation.
- **Answers the question:** What sequence is most likely to appear?
- **Example:** Tossing a coin N number of times.
- M = observed number of heads in the sequence.
- Probability of Head $p_H = \frac{M}{N}$
- Maximum Likelihood Estimate: $p_H^* = \underset{p_H}{argmax} \quad p(X|p_H)$ where X=observation
- If probability of Head is less then less probability of Head in the sequence.

Maximum Likelihood Estimation (MLE)

- In general, Maximize $p(X|\theta)$, where X: observation, θ: parameters.
- $\theta^* = \underset{\theta}{argmax} \quad p(X|\theta)$
- **Proof:** Toss a coin N times where probability of head $p_H = \frac{M}{N}$
- Let, p(H) = p then Likelihood L =$p(X|\theta) = \prod_{i=1}^{N} p^{x_i} (1-p)^{1-x_i}$ (Bernoulli trial)
- LL = log likelihood = $\sum_{i=1}^{N} x_i \log p + \sum_{i=1}^{N} (1-x_i) \log(1-p)$

Maximum Likelihood Estimation (MLE)

- LL = $\sum_{i=1}^{N} x_i \log p + \sum_{i=1}^{N} (1-x_i) \log(1-p)$
- To maximize LL, set $\frac{dLL}{dp} = 0$
- $\sum_{i=1}^{N} x_i \frac{1}{p} + \sum_{i=1}^{N} (1-x_i) \frac{1}{(1-p)} \times (-1) = 0$
- $\frac{M}{p} - \frac{(N-M)}{(1-p)} = 0$

  $p = \frac{M}{N}$

Maximum Entropy Principle (ME)

- Alternate to MLE.
- Used when there is no information about the observation.
- There is an uncertainty involved.
- **Example:** Tossing an unbiased coin.
- Let *p* = probability of Head, *1-p* = probability of Tail.
- Entropy $E = -p \log p - (1-p)\log(1-p)$

Maximum Entropy Principle (ME)

- Entropy $E = -p \log p - (1-p) \log(1-p)$
- To Maximize E, set $\frac{dE}{dp} = 0$
- $\frac{dE}{dp} = -[p.\frac{1}{p} + \log p] - [(1-p).\frac{1}{(1-p)}.(0-1) + \log(1-p).(0-1)] = 0$
- $-1 - \log p + 1 + \log(1-p) = 0$
- $\frac{1-p}{p} = 1$
- $p = \frac{1}{2}$

Review

- Parameter θ, Observation X, likelihood $p(X|\theta)$, log likelihood $\log(p(X|\theta))$
- Entropy $E = -\sum p_i \log p_i$
- Maximum Entropy Principle
- Two parameter estimation techniques
  - MLE, when observation given
  - ME, when no observation given
- Parameter estimation is also called **statistical inferences**.

EM: Introducing hidden variable

- X: X_1, X_2, X_3, ……………, X_i, …………………… X_{N-1}, X_N
- x: x_1, x_2, x_3, ……………., x_i, …………………… x_{N-1}, x_N
- z: z_1, z_2, z_3, ……………., z_i, …………………. z_{N-1}, z_N
- **Example:** One coin toss, p = probability of Head = θ
- Suppose X : H T H T H H T T T H

$$\begin{array}{cccccccccc} H & T & H & T & H & H & T & T & T & H \\ \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow \\ \theta & (1-\theta) & \theta & (1-\theta) & \theta & \theta & (1-\theta) & (1-\theta) & (1-\theta) & \theta \end{array}$$

- $L(\theta) = p(X|\theta) = \theta.(1-\theta).\theta.(1-\theta).\theta.\theta.(1-\theta).(1-\theta).(1-\theta).\theta$ (independent events)
- If there are M heads then
- $L(\theta) = p(X|\theta) = \theta^M(1-\theta)^{N-M}$

EM: Introducing hidden variable …...contd

- $L(\theta) = p(X|\theta) = \theta^M(1-\theta)^{N-M}$
- $\frac{dLL}{d\theta} = 0$ gives θ = M/N.
- Whatever be the sequence, $L(\theta) = \theta^M(1-\theta)^{N-M}$
- If we introduce indicator variable then,
- $$L(\theta) = \prod_{i=1}^{N} \theta^{x_i} (1-\theta)^{1-x_i}$$

EM: Two Coin Toss Example

- **Example:** Two coin toss, $\theta = \langle p, p_1, p_2 \rangle$
- p = probability of choosing coin-1/coin-2.
- p_1 = probability of Head from coin-1.
- p_2 = probability of Head from coin-2.
- Suppose X : H T H T H H T T T H

$$\underbrace{\text{H}}_{\{pp_1 + (1-p)p_2\}} \qquad \underbrace{\text{T}}_{\{p(1-p_1) + (1-p)(1-p_2)\}}$$

- If there are M heads then
- Let W = $\{pp_1 + (1-p)p_2\}^M . \{p(1-p_1) + (1-p)(1-p_2)\}^{N-M}$

EM: Two Coin Toss Example

- W = $\{pp_1 + (1-p)p_2\}^M . \{p(1-p_1) + (1-p)(1-p_2)\}^{N-M}$
- A = Log likelihood of W = $M.\log Q_1 + (N-M).\log Q_2$, *where*
  - $Q_1 = pp_1 + (1-p)p_2$
  - $Q_2 = p(1-p_1) + (1-p)(1-p_2)$
- To maximize *A*, make $\frac{\delta A}{\delta \theta} = 0$
- As θ = <p, p_1, p_2>, we need to do partial derivatives.

EM: Two Coin Toss Example … Computing *p*

- $A = M.\log(pp_1 + (1-p)p_2) + (N-M)\log(p(1-p_1) + (1-p)(1-p_2))$
- Make $\frac{\delta A}{\delta p} = 0$
- $\frac{M.(p_1-p_2)}{pp_1+(1-p)p_2} + \frac{(N-M)(1-p_1-1+p_2)}{p(1-p_1)+(1-p)(1-p_2)} = 0$
- $\frac{M.(p_1-p_2)}{pp_1+p_2-pp_2} - \frac{(N-M)(p_1-p_2)}{p-pp_1+1-p_2-p+pp_2} = 0$
- Solving this equation you will get $p = \frac{M-Np_2}{N(p_1-p_2)}$
- So, $p = f_1(M, N, p_1, p_2)$

EM: Two Coin Toss Example … Computing p_1

- $A =$ $M.\log\left(pp_1 + (1-p)p_2\right) + (N-M)\log\left(p(1-p_1) + (1-p)(1-p_2)\right)$
- Make $\frac{\delta A}{\delta p_1} = 0$
- $\frac{M.p}{pp_1+(1-p)p_2} + \frac{(N-M).(-p)}{p(1-p_1)+(1-p)(1-p_2)} = 0$
- $\frac{M}{pp_1+p_2-pp_2} - \frac{N-M}{p-pp_1+1-p_2-p+pp_2} = 0$
- Solving this equation you will get $p_1 = \frac{M-Np_2+Npp_2}{Np}$
- So, $p_1 = f_2(M, N, p, p_2)$

EM: Two Coin Toss Example … Computing p_2

- $A = M.\log\left(pp_1 + (1-p)p_2\right) + (N-M)\log\left(p(1-p_1) + (1-p)(1-p_2)\right)$
- Make $\frac{\delta A}{\delta p_2} = 0$
- $\frac{M(1-p)}{pp_1+(1-p)p_2} + \frac{(N-M).(-1+p)}{p(1-p_1)+(1-p)(1-p_2)} = 0$
- $\frac{M}{pp_1+p_2-pp_2} - \frac{N-M}{p-pp_1+1-p_2-p+pp_2} = 0$
- Solving this equation you will get $p_2 = \frac{M-Npp_1}{N(1-p)}$
- So, $p_2 = f_3(M, N, p, p_1)$

EM: Why Hidden Variable?

- The calculation was very complex.
- **So, we introduce hidden variable**.
- We actually computed Joint probability of Coin choice (*Z*) and observation $(X|\theta)$ which is $p(Z, X|\theta)$.
- But, the relation between $p(Z, X|\theta)$ and $p(X|\theta)$ is
- $p(X|\theta) = \sum_Z p(Z, X|\theta)$ (Law of Marginalization)

Two Coin Toss Problem: Introducing CONCAVITY

- $\theta^* = \underset{\theta}{argmax} \quad p(X|\theta)$
- $\theta^* = \underset{\theta}{argmax} \quad \sum_Z p(X, Z|\theta)$
- Log likelihood = $LL(\theta)$ = $\log \sum_Z p(X, Z|\theta)$
- Log has an interesting property called **CONCAVITY** which says
- $\log\left(\sum_i \lambda_i x_i\right) \geq \sum_i \lambda_i \log\left(x_i\right)$ where $\lambda_i \geq 0, \forall i$ and $\sum_i \lambda_i = 1$
- Using this rule we get $\log \sum_Z p(X, Z|\theta) \geq \sum_Z \lambda_Z \log\left[\frac{p(X,Z|\theta)}{\lambda_Z}\right]$

Two Coin Toss Problem: Introducing Entropy

- So, $\log \sum_Z p(X, Z|\theta) \geq \sum_Z \lambda_Z \log \left[\frac{p(X,Z|\theta)}{\lambda_Z} \right]$ …..(1)
- Let's take $\lambda_Z = p(Z|X, \theta)$
- Putting the value of λ_Z in (1), we get

Entropy

- $\log \sum_Z p(X, Z|\theta) \geq \sum_Z p(Z|X, \theta) \log(p(X, Z|\theta)) - \sum_Z p(Z|X, \theta) \log(p(Z|X, \theta))$ (2)
- Now let us evaluate the first term in the r.h.s i.e. $\sum_Z p(Z|X, \theta) \log(p(X, Z|\theta))$
- By definition, $\sum_Z p(Z|X, \theta) \log(p(X, Z|\theta)) = \underset{Z|X,\theta}{E}[\log(p(X, Z|\theta))]$ …….(3)

Two Coin Toss Problem: Introducing Bernoulli trial

- We have $\sum_Z p(Z|X,\theta)\log(p(X,Z|\theta)) = \underset{Z|X,\theta}{E}[\log(p(X,Z|\theta))]$ …… (3)
- Now, $p(X,Z|\theta)$ $= \prod_{i=1}^{N}\left[pp_1^{x_i}(1-p_1)^{1-x_i}\right]^{z_i}.\left[(1-p)p_2{}^{x_i}(1-p_2)^{1-x_i}\right]^{1-z_i}$
  - where, z_i=1, if coin-1 chosen, z_i=0, if coin-2 chosen.
  - Here, both choosing coin and coin tosses follow **Bernoulli trial.**
- So, $\log(p(X,Z|\theta))$ = $\sum_{i=1}^{N} z_iQ_1 + (1-z_i)Q_2$
  - where, $Q_1 = \log p + x_i logp_1 + (1-x_i)log(1-p_1)$
  - And $Q_2 = \log(1-p) + x_i logp_2 + (1-x_i)log(1-p_2)$

Two Coin Toss Problem

- $\underset{Z|X,\theta}{E}[\log(p(X,Z|\theta))] = \underset{Z|X,\theta}{E}[\sum_{i=1}^{N}(z_i Q_1 + (1-z_i)Q_2)]$

How can we replace Zi to which expectations of E[Zi]

- $= \sum_{i=1}^{N}(\underset{Z|X,\theta}{E}[z_i]Q_1 + (1-\underset{Z|X,\theta}{E}[z_i])Q_2)$
- $= \sum_{i=1}^{N}\underset{Z|X,\theta}{E}[z_i]Q_1 + \sum_{i=1}^{N}(1-\underset{Z|X,\theta}{E}[z_i])Q_2$

Entropy

Study the EM tutorial the p(zlx, Q) that we have taken is suppsed to be for teh nth iteration of theta hence entopry becomes constant for that

- Our original equation was from (2)
- $\log\sum_Z p(X,Z|\theta) \geq \sum_Z p(Z|X,\theta)\log(p(X,Z|\theta)) - \sum_Z p(Z|X,\theta)\log(p(Z|X,\theta))$
- $= \sum_{i=1}^{N}\underset{Z|X,\theta}{E}[z_i]Q_1 + \sum_{i=1}^{N}(1-\underset{Z|X,\theta}{E}[z_i])Q_2$ + Entropy
- **If Entropy is constant** then we are safe in maximizing the r.h.s.

Two Coin Toss Problem: Final Equation

- **Our Final equation is:**
- $\log \sum_Z p(X, Z|\theta) \geq \sum_{i=1}^{N} \underset{Z|X,\theta}{E}[z_i] Q_1 + \sum_{i=1}^{N} (1 - \underset{Z|X,\theta}{E}[z_i]) Q_2$ + C, where

  C=Entropy

Why

- Now if $f(x) \geq g(x)$ then maximizing $g(x)$ will be enough to maximize $f(x)$.
- So, let A = $g(x)$ = $\sum_{i=1}^{N} \underset{Z|X,\theta}{E}[z_i] Q_1 + \sum_{i=1}^{N} (1 - \underset{Z|X,\theta}{E}[z_i]) Q_2$ + C
- To maximize A, make $\frac{\delta A}{\delta \theta} = 0$
- As θ = <p, p_1, p_2>, we need to do partial derivatives.

Two Coin Toss Problem: Evaluating parameters

- A = $\sum_{i=1}^{N} \underset{Z|X,\theta}{E}[z_i] Q_1 + \sum_{i=1}^{N} (1 - \underset{Z|X,\theta}{E}[z_i]) Q_2$ + C
- Make $\frac{\delta A}{\delta p} = 0$
- Now, $Q_1 = \log p + x_i logp_1 + (1 - x_i) log(1 - p_1)$and
- $Q_2 = \log(1-p) + x_i logp_2 + (1 - x_i) log(1 - p_2)$
- So, $\frac{\delta A}{\delta p} = \frac{\sum_{i=1}^{N} E(z_i)}{p} - \frac{\sum_{i=1}^{N} (1 - E(z_i))}{1-p} = 0$
- $\frac{\sum_{i=1}^{N} E(z_i)}{p} - \frac{N - \sum_{i=1}^{N} E(z_i)}{1-p} = 0$
- Solving this we get $p = \frac{\sum_{i=1}^{N} E(z_i)}{N}$

Two Coin Toss Problem: Evaluating parameters

- A = $\sum_{i=1}^{N} \underset{Z|X,\theta}{E}[z_i]Q_1 + \sum_{i=1}^{N}(1 - \underset{Z|X,\theta}{E}[z_i])Q_2$ + C
- Make $\frac{\delta A}{\delta p_1} = 0$
- Solving this we get $p_1 = \frac{\sum_{i=1}^{N} x_i E(z_i)}{\sum_{i=1}^{N} E(z_i)}$
- Make $\frac{\delta A}{\delta p_2} = 0$
- Solving this we get $p_2 = \frac{M - \sum_{i=1}^{N} x_i E(z_i)}{N - \sum_{i=1}^{N} E(z_i)}$

Questions to be answered

1. Why can Entropy be treated as Constant?

How is entropy being treated as constant it will chnange with teh balue of p

2. Why does maximizing $g(x)$ lead to maximizing $f(x)$?

Essentially g(x) is a linear combination of teh function of f(X) if we maximze each compoenent of g we are essetially maximzing the each compoenent of f only