

## Expectation Maximization (EM)

- EM is used to estimate parameters in the presence of missing attributes.
- Converges to a local maximum of the maximum likelihood function.
- Some uses:
  - Train Bayesian Belief Networks
  - Unsupervised clustering
  - Learning Hidden Markov Models

1

## Three Coin Example

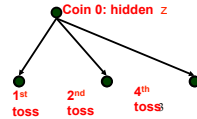
- We observe a series of coin tosses generated in the following way:
  - A person has three coins.
    - Coin 0: probability of Head is  $\lambda$
    - Coin 1: probability of Head  $p$
    - Coin 2: probability of Head  $q$
- Consider the following coin-tossing scenarios:

2

## Estimation Problems

- Scenario I: Toss one of the coins six times.  
Observing HHHHTT  
Which coin is more likely to produce this sequence? Suppose we know the probability of H for each coin.
- Scenario II: Toss coin 0. If Head – toss coin 1; o/w – toss coin 2  
Observing the sequence HHHHT, THHT, HHHHT, HHTTH  
produced by Coin 0, Coin 1 and Coin 2  
Estimate most likely values for  $p, q, \lambda$  (the probability of H in each coin)
- Scenario III: Toss coin 0. If Head – toss coin 1; o/w – toss coin 2  
Observing the sequence HHHHT, THHT, HHHHT, HHTTH  
produced by Coin 1 and/or Coin 2  
Estimate most likely values for  $\lambda, p, q$ .

The label of the first toss ( $z$ ) is hidden, we want to estimate the most likely hypothesis  $\theta = (\lambda, p, q)$  under hidden  $z$ .



3

## Key Intuition (1)

- If we knew which of the data points (HHHT), (HTHT), (HTTH) came from Coin 1 and which from Coin 2, there was no problem.
- Recall that the “simple” estimation is the ML estimation:
- Assume that you toss a  $(p, 1-p)$  coin  $m$  times and get  $k$  Heads  $m-k$  Tails.

$$\log[P(D|p)] = \log [p^k (1-p)^{m-k}] = k \log p + (m-k) \log (1-p)$$

- To maximize, set the derivative w.r.t.  $p$  equal to 0:

$$d/dp \{ \log P(D|p) \} = k/p - (m-k)/(1-p) = 0$$

- Solving this for  $p$ , gives:  $p = k/m$

4

## Key Intuition (2)

- Since we do not know which of the data points (HHHT), (HTHT), (HTTH) came from Coin 1 and which from Coin 2, we use an iterative approach for estimating  $(\lambda, p, q)$ .
- Let  $z$  be the hidden variable, which coin was tossed at each attempt.
- Compute  $P(z)$ , given the current hypothesis  $(\tilde{\lambda}, \tilde{p}, \tilde{q})$  and the observed data  $D$ . This will lead to a distribution.
- Compute the expected likelihood of the data  $D$  using the distribution over  $z$ .
  - Each value of  $z$  has a probability of occurrence: defines a possible world.
  - Compute the likelihood in each world.
  - Compute the expected likelihood.
- Maximize the likelihood of the data to recompute  $(\lambda, p, q)$ .
- This process can be iterated and can be shown to converge to a local maximum of the parameters  $(\lambda, p, q)$ .

5

## EM Algorithm (Coins) - I

- Suppose  $(\tilde{\lambda}, \tilde{p}, \tilde{q})$  is the current estimate of parameters.
- What is the probability  $P(z)$  given  $(\tilde{\lambda}, \tilde{p}, \tilde{q})$  and  $D$ ?
- Suppose there were  $m$  coin tosses and  $h$  heads in  $D^i$ . Given the current parameters,

$$P_i = P(z_i = 1 | D^i) = P(\text{Coin} | D^i) = \frac{P(D^i | \text{Coin} 1) P(\text{Coin} 1)}{P(D^i)}$$

$$= \frac{\tilde{\lambda} \tilde{p}^h (1 - \tilde{p})^{m-h}}{\tilde{\lambda} \tilde{p}^h (1 - \tilde{p})^{m-h} + (1 - \tilde{\lambda}) \tilde{q}^h (1 - \tilde{q})^{m-h}}$$

$$E[Y] = \sum_{y_i} y_i P(Y = y_i)$$

$$E[z_i] = 1 \times P(D_i \text{ was obtained from Coin 1}) + 0 \times P(D_i \text{ was obtained from Coin 2}) = P_i$$

6

## EM Algorithm (Coins) - II

- We would like to compute the likelihood of the data, and find the parameters that maximize it.
- We want to maximize the log likelihood of the data:
  - $LL = \log P(D, z | \lambda, p, q)$
- But,  $z$  is hidden.
- Which value do we use for it in order to compute the likelihood of the data?
- We think of the likelihood  $\log P(D, z | \lambda, p, q)$  as a random variable that depends on  $z$ . Therefore, instead of maximizing the LL, we maximize the expectation of this random variable.

7

## EM Algorithm (Coins) - III

$$\begin{aligned}
 P(D^1 | \lambda, p, q) &= \lambda p^{h_1} (1-p)^{m-h_1} \\
 P(D^1 | \lambda, p, q) &= (1-\lambda) q^{h_1} (1-q)^{m-h_1} \quad z_i \text{ is an indicator variable} \\
 P(D^1 | z_i | \lambda, p, q) &= [\lambda p^{z_i} (1-p)^{m-z_i}]^{z_i} [(1-\lambda) q^{z_i} (1-q)^{m-z_i}]^{1-z_i} \\
 &= \lambda^{z_i} p^{z_i h_i} (1-p)^{z_i(m-h_i)} (1-\lambda)^{1-z_i} q^{(1-z_i)h_i} (1-q)^{(1-z_i)(m-h_i)} \\
 \log P(D^1 | z_i | \lambda, p, q) &= z_i \log \lambda + z_i h_i \log p + z_i(m-h_i) \log(1-p) + \\
 &\quad (1-z_i) \log(1-\lambda) + (1-z_i) h_i \log q + (1-z_i)(m-h_i) \log(1-q) \\
 P(D, z | \lambda, p, q) &= \prod_i P(D^1 | z_i | \lambda, p, q) \\
 \log P(D, z | \lambda, p, q) &= \sum_i \log P(D^1 | z_i | \lambda, p, q) \\
 E[\log P(D, z | \lambda, p, q)] &= E[\sum_i \log P(D^1 | z_i | \lambda, p, q)] = \sum_i E[\log P(D^1 | z_i | \lambda, p, q)] \\
 &= \sum_i E[z_i \log \lambda + z_i h_i \log p + z_i(m-h_i) \log(1-p) + (1-z_i) \log(1-\lambda) + (1-z_i) h_i \log q + (1-z_i)(m-h_i) \log(1-q)] \\
 &= \sum_i [p_i \log \lambda + p_i h_i \log p + p_i(m-h_i) \log(1-p) + (1-p_i) \log(1-\lambda) + (1-p_i) h_i \log q + (1-p_i)(m-h_i) \log(1-q)] \\
 E[X+Y] &= E[X] + E[Y] \\
 E[z_i] &= p_i
 \end{aligned}$$

We maximize this quantity wrt  $\lambda, p, q$

8

## EM Algorithm (Coins) - IV

When computing the derivatives, notice  $E[z_i]$  here is a constant; it is computed using the current parameters.

- In order to find the most likely parameters we maximize the derivatives with respect to  $\lambda, p, q$ :

$$\begin{aligned}
 \frac{dE}{d\lambda} &= \sum_i \left( \frac{p_i}{\lambda} - \frac{1-p_i}{1-\lambda} \right) = 0 \Rightarrow \lambda = \frac{\sum_i p_i}{n} \\
 \frac{dE}{dp} &= \sum_i p_i \left( \frac{h_i}{p} - \frac{m-h_i}{1-p} \right) = 0 \Rightarrow p = \frac{\sum_i p_i h_i}{\sum_i p_i m} \\
 \frac{dE}{dq} &= \sum_i (1-p_i) \left( \frac{h_i}{q} - \frac{m-h_i}{1-q} \right) = 0 \Rightarrow q = \frac{\sum_i (1-p_i) h_i}{\sum_i (1-p_i) m}
 \end{aligned}$$

9

## Justification for EM algorithm

- Find parameter  $\theta$  that maximizes  $\ln P(x|\theta) = \ln \sum_z P(x, z|\theta)$  all summation over hidden  $z$
- $P(x, z|\theta) = P(z|x, \theta) P(x|\theta)$   
 $\ln P(x|\theta) = \ln P(x, z|\theta) - \ln P(z|x, \theta)$   $Q(\theta|\theta^t)$
- Suppose we have a current estimate  $\theta^t$ . Take expectation over  $P(z|x, \theta^t)$ :
- $\ln P(x|\theta) = \sum_z P(z|x, \theta^t) \ln P(x, z|\theta) - \sum_z P(z|x, \theta^t) \ln P(z|x, \theta)$
- $\ln P(x|\theta) = Q(\theta|\theta^t) - \sum_z P(z|x, \theta^t) \ln P(z|x, \theta)$
- Instantiate above equation for  $\ln P(x|\theta^t)$  and subtract:  $KL(P(z|x, \theta^t) || P(z|x, \theta)) \geq 0$
- $\ln P(x|\theta) - \ln P(x|\theta^t) = Q(\theta|\theta^t) - Q(\theta^t|\theta^t) + \sum_z P(z|x, \theta^t) \ln [P(z|x, \theta^t) / P(z|x, \theta)]$
- $\ln P(x|\theta) - \ln P(x|\theta^t) \geq Q(\theta|\theta^t) - Q(\theta^t|\theta^t)$
- Estimation (E) step:** Calculate  $Q(\theta|\theta^t)$  using the current hypothesis  $\theta^t$  and the observed data  $x$ .  
 $Q(\theta|\theta^t) := \sum_z P(z|x, \theta^t) \ln P(x, z|\theta) = E[\ln P(x, z|\theta) | \theta^t, x]$
- Maximization (M) step:** Replace  $\theta^t$  by the hypothesis  $\theta$  that maximizes  $Q$   
 $\theta^{t+1} := \underset{\theta}{\operatorname{argmax}} Q(\theta|\theta^t)$
- Converges to a local maxima

10

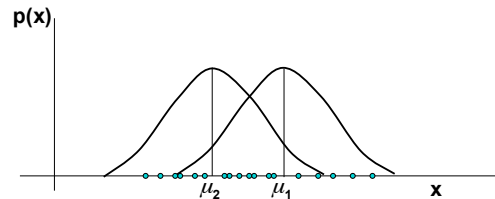
## EM Summary (so far)

- EM is a general procedure for learning in the presence of unobserved variables.**
- EM is an iterative algorithm that can be shown to converge to a local maximum of the likelihood function.**

11

## Example: K-Means Algorithm

**K-means is a clustering algorithm.**  
 We are given data points, known to be sampled independently from a mixture of  $k$  Normal distributions, with means  $\mu_i, i=1, \dots, k$  and the same standard variation  $\sigma$



12

## Example: K-Means Algorithm

First, notice that if we knew that all the data points are taken from a normal distribution with mean  $\mu$ , finding its most likely value is easy.

$$p(x | \mu) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}(x - \mu)^2\right]$$

We get many data points,  $D = \{x_1, \dots, x_m\}$

Maximizing the log-likelihood is equivalent to maximizing:

$$\ln L(D; \mu) = \ln P(D; \mu) = \sum_i \ln \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}(x_i - \mu)^2\right]$$

Calculate the derivative with respect to  $\mu$ , we get that the minimal point, that is, the most likely mean is

$$\mu_{ML} = \arg \min_{\mu} \sum_i (x_i - \mu)^2$$

$$\mu = \frac{1}{m} \sum_i x_i$$

13

## A Mixture of Distributions

As in the coin example, the problem is that data is sampled from a mixture of  $k$  different normal distributions, and we do not know, for a given data point  $x$ , where is it sampled from.

Assume that we observe data point  $x$ . What is the probability that was sampled from the distribution  $\mu_i$ ?

$$P_{ij} = P(\mu_j | x_i) = \frac{P(x_i | \mu_j) P(\mu_j)}{P(x_i)} = \frac{\frac{1}{k} P(x = x_i | \mu = \mu_j)}{\sum_{n=1}^k \frac{1}{k} P(x = x_i | \mu = \mu_n)} = \frac{\exp\left[-\frac{1}{2\sigma^2}(x_i - \mu_j)^2\right]}{\sum_{n=1}^k \exp\left[-\frac{1}{2\sigma^2}(x_i - \mu_n)^2\right]}$$

14

## A Mixture of Distributions

For a data point  $x_i$ , define  $k$  binary hidden variables,  $Z_{i1}, Z_{i2}, \dots, Z_{ik}$ , s.t.  $Z_{ij} = 1$  iff  $x_i$  is sampled from the  $j$ -th distribution.

$$E[Z_{ij}] = 1 \times P(x_i \text{ was sampled from } \mu_j) + 0 \times P(x_i \text{ was not sampled from } \mu_j) = P_{ij}$$

15

## The EM Algorithm

### Algorithm:

- Guess initial values for the hypothesis  $h = \mu_1, \mu_2, \dots, \mu_k$
- **Expectation:** Calculate  $Q(h' | h) = E(\ln P(Y|h') | h, X)$  using the current hypothesis  $h$  and the observed data  $X$ .  $Y =$  observed data  $X +$  unobserved data  $Z$ .
- **Maximization:** Replace the current hypothesis  $h$  by  $h'$ , that maximizes the  $Q$  function (the likelihood function):  
Set  $h = h'$ , such that  $Q(h' | h)$  is maximal
- **Repeat:** Estimate the Expectation again.

16

## Example: k-means Algorithm

### Expectation:

$$p(y_i | h) = p(x_i, z_{i1}, \dots, z_{ik} | h) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2} \sum_j z_{ij} (x_i - \mu_j)^2\right]$$

Computing the likelihood given the observed data  $D = \{x_1, \dots, x_m\}$  and the hypothesis  $h$  (w/o the constant coefficient)

$$\ln P(Y|h) = \sum_{i=1}^m \ln \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2} \sum_j z_{ij} (x_i - \mu_j)^2\right]$$

$$E[\ln P(Y|h)] = E\left[\sum_{i=1}^m \ln \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2} \sum_j z_{ij} (x_i - \mu_j)^2\right]\right]$$

$$= \sum_{i=1}^m \ln \frac{1}{\sqrt{2\pi\sigma^2}} \sum_j E[z_{ij}] (x_i - \mu_j)^2$$

17

## Example: k-means Algorithm

### Maximization: Maximizing

$$Q(h|h') = \sum_{i=1}^m \ln \frac{1}{\sqrt{2\pi\sigma^2}} \sum_j E[z_{ij}] (x_i - \mu_j)^2$$

with respect to  $\mu_j$  we get:

$$\frac{dQ}{d\mu_j} = C \sum_{i=1}^m E[z_{ij}] (x_i - \mu_j) \mu_j = 0$$

Which yields:

$$\mu_j = \frac{\sum_{i=1}^m E[z_{ij}] x_i}{\sum_{i=1}^m E[z_{ij}]}$$

18

### Summary: k-means Algorithm

Given a set  $D = \{x_1, \dots, x_n\}$  of data points,  
 guess initial parameters  $\mu_1, \mu_2, \dots, \mu_k$

Compute (for all  $i, j$ )  $p_{ij} = E[z_{ij}] = \frac{\exp[-\frac{1}{2\sigma^2}(x_i - \mu_j)^2]}{\sum_{n=1}^k \exp[-\frac{1}{2\sigma^2}(x_i - \mu_n)^2]}$

and a new set of means:  $\mu_j = \frac{\sum_{i=1}^n E[z_{ij}]x_i}{\sum_{i=1}^n E[z_{ij}]}$

repeat to convergence

19

### Summary: EM

- EM is a general procedure for learning in the presence of unobservable variables.
- We have shown how to use it in order to estimate the most likely density function for a mixture of probability distributions.
- EM is an iterative algorithm that can be shown to converge to a local maximum of the likelihood function.
- It has been shown to be quite useful in practice.

20

### Motif finding

- motif finding
  - $\theta$  = distributions of amino acids in motif and background
  - $x$  = observed sequences
  - $z_{ij} = 1$  if motif begins at  $j$ th symbol in string  $i$ .
  - What would you do if  $z$  were visible?

### Motif finding

- $X = \{X_1, \dots, X_N\}$  = set of given sequences
- Define motif model:
  - $M = (M_1, \dots, M_k)$
  - $M_i = (M_{iA}, \dots, M_{iT})$  (assume  $\{A, C, G, T\}$ )
  - where  $M_{ij} = \text{Prob}[\text{motif position } i \text{ is letter } j]$
- Define background model:
  - $B = B_A, \dots, B_T$
  - $B_i = \text{Prob}[\text{letter } i \text{ in background sequence}]$
- Parameter space  $\theta = (M, B)$
- Hidden variables
  - $Z_{ij} = 1$  if motif begins at position  $j$  in sequence  $i$
- [Lawrence and Reilly 1990]

### Computing $P[Z_{ij} | X, \theta^t]$

- $X_i = \text{GCTGAG}$
- $k = 3$
- $\theta^t =$ 

	0	1	2	3
A	0.25	0.1	0.5	0.2
C	0.25	0.4	0.2	0.1
G	0.25	0.3	0.1	0.6
T	0.25	0.2	0.2	0.1
- Define
  - $V_{i1} = 0.3 * 0.2 * 0.1 * (0.25)^3$
  - $V_{i2} = \dots$
  - $V_{i3} = \dots$
  - $V_{i4} = \dots$
- $P[Z_{i1} | X, \theta^t] = V_{i1} / \sum V_{ij}, \dots$

### Computing $\theta^{t+1}$

- $X_1 = \text{A C A G C A}$
  - $X_2 = \text{A G G C A G}$
  - $X_3 = \text{T C A G T C}$
  - $P[Z_1 | X, \theta^t] = \begin{matrix} 0.1 & 0.7 & 0.1 & 0.1 \end{matrix}$
  - $P[Z_2 | X, \theta^t] = \begin{matrix} 0.4 & 0.1 & 0.1 & 0.4 \end{matrix}$
  - $P[Z_3 | X, \theta^t] = \begin{matrix} 0.2 & 0.6 & 0.1 & 0.1 \end{matrix}$
- Assume these are the values
- $Q(\theta^t) = E[\ln P(X, Z | \theta) | \theta^t, X] = \sum E[\ln P(X_i, Z_i | \theta) | \theta^t, X]$
  - $P(X_1, Z_1 | \theta) = (\theta_{1,A} \times \theta_{2,C} \times \theta_{3,A} \times \theta_{0,G} \times \theta_{0,C} \times \theta_{0,A}, \dots)$
  - $P(X_2, Z_2 | \theta) = (\dots)$
  - $P(X_3, Z_3 | \theta) = (\dots)$
  - $Q(\theta^t) = \sum (\ln(\theta_{1,A} \times \theta_{2,C} \times \theta_{3,A} \times \theta_{0,G} \times \theta_{0,C} \times \theta_{0,A}) \times 0.1 + \dots)$
  - Solve for  $\theta$  that maximizes  $Q(\theta, \theta^t)$ .
- 4 terms

## Convergence of EM

- Usually converges in a small number of iterations
- Initial parameters matter
- MEME: Bailey and Elkan, ISMB 1994.
  - Improvement to basic EM
  - One of three motif models:
    - ♦ OOPS: One expected occurrence per sequence
    - ♦ ZOOPS: Zero or one expected occurrence per sequence
    - ♦ TCM: Any number of occurrences of the motif