Expectation-Maximization & Baum-Welch

The probabilistic setting

Input: data *x* coming from a probabilistic model with hidden information *y*

Goal: Learn the model's parameters so that the likelihood of the data is maximized.

Example: a mixture of two Gaussians

$$P(y_i = 1) = p_1 \; ; P(y_i = 2) = p_2 = 1 - p_1$$

$$P(x_i \mid y_i = j) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-(x_i - \mu_j)^2}{2\sigma^2} \right)$$

The likelihood function

$$P(y_i = 1) = p_1 \; ; P(y_i = 2) = p_2 = 1 - p_1$$

$$P(x_i \mid y_i = j) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-(x_i - \mu_j)^2}{2\sigma^2} \right)$$

$$L(\theta) = \prod_i P(x_i \mid \theta) = \prod_i \sum_j P(x_i, y_i = j \mid \theta)$$

$$\log L(\theta) = \sum_i \log\left(\sum_j \frac{p_j}{\sigma\sqrt{2\pi}} \exp\left(\frac{-(x_i - \mu_j)^2}{2\sigma^2} \right) \right)$$

Kullback-Leibler divergence is positive

log(x)≤x-1 for all x>0

$$\sum_{i\in +} P(x_i)\cdot \log\frac{Q(x_i)}{P(x_i)} \le \sum_{i\in +} P(x_i)\cdot\left(\frac{Q(x_i)}{P(x_i)}-1\right)$$

$$=\sum_{i\in +} Q(x_i)-1\le 0$$

CG

The EM algorithm

Goal: max $\log P(x|\theta)=\log (\Sigma\, P(x,y|\theta))$

Assume we have a model θ^t which we wish to improve.

Note: $P(x|\theta) = P(x,y|\theta) / P(y|x,\theta)$

$$P(y|x,\theta^t)\cdot\log P(x|\theta) = P(y|x,\theta^t)\cdot\log P(x,y|\theta) - P(y|x,\theta^t)\cdot\log P(y|x,\theta)$$

$$\sum_y P(y|x,\theta^t)\cdot\log P(x|\theta) = \sum_y P(y|x,\theta^t)\cdot\log P(x,y|\theta) - \sum_y P(y|x,\theta^t)\cdot\log P(y|x,\theta)$$

$$\log P(x|\theta) = \sum_y P(y|x,\theta^t)\cdot\log P(x,y|\theta) - \sum_y P(y|x,\theta^t)\cdot\log P(y|x,\theta)$$

$$\log P(x|\theta^t) = \sum_y P(y|x,\theta^t)\cdot\log P(x,y|\theta^t) - \sum_y P(y|x,\theta^t)\cdot\log P(y|x,\theta^t)$$

$$\Delta \quad = Q(\theta|\theta^t) - Q(\theta^t|\theta^t) + \sum_y P(y|x,\theta^t)\cdot\log \frac{P(y|x,\theta^t)}{P(y|x,\theta)}$$

Constant >=0

The EM algorithm (cont.)

Main component:

$$Q(\theta \mid \theta^t) = \sum_y P(y \mid x, \theta^t) \cdot \log P(x, y \mid \theta)$$

is the expectation of logP(x,y|θ) over the distribution of y given by the current parameters θ^t

The algorithm:

- E-step: Calculate the Q function
- M-step: Maximize $Q(\theta|\theta^t)$ with respect to θ
- Stopping criterion: improvement in log likelihood $\leq \varepsilon$

Application to the mixture model

$$Q(\theta \mid \theta^t) = \sum_y P(y \mid x, \theta^t) \cdot \log P(x, y \mid \theta)$$

$$P(x, y \mid \theta) = \prod_i P(x_i, y_i \mid \theta) = \prod_i \prod_j P(x_i, y_i = j \mid \theta)^{y_{ij}}$$

$$y_{ij} = \begin{cases} 1 & y_i = j \\ 0 & y_i \neq j \end{cases}$$

$$\log P(x, y \mid \theta) = \sum_i \sum_j y_{ij} \log P(x_i, y_i = j \mid \theta)$$

$$Q(\theta \mid \theta^t) = \sum_y P(y \mid x, \theta^t) \sum_i \sum_j y_{ij} \log P(x_i, y_i = j \mid \theta) =$$

$$\sum_i \sum_j \sum_y P(y \mid x, \theta^t) y_{ij} \log P(x_i, y_i = j \mid \theta)$$

Application (cont.)

$$Q(\theta \mid \theta^t) = \sum_i \sum_j P(y_{ij} = 1 \mid x_i, \theta^t) \log P(x_i, y_i = j \mid \theta)$$

$$w_{ij}^t := P(y_{ij} = 1 \mid x_i, \theta^t) = \frac{P(x_i, y_i = j \mid \theta^t)}{\sum_j P(x_i, y_i = j \mid \theta^t)}$$

$$Q(\theta \mid \theta^t) = \sum_i \sum_j w_{ij}^t \left(\log \frac{1}{\sqrt{2\pi}} - \log \sigma + \log p_j - \frac{(x_i - \mu_j)^2}{2\sigma^2} \right)$$

Recap & another example

Input: data x coming from a probabilistic model with hidden information y

Example 2: a mixture of two coins (Bernoulis)

$$P(y_i = 1) = p_1 \; ; P(y_i = 2) = p_2 = 1 - p_1$$

$$P(x_i \mid y_i = j) = h_j$$

$$L(\theta) = \prod_i P(x_i \mid \theta) = \prod_i \sum_j P(x_i, y_i = j \mid \theta) =$$

$$\prod_i \sum_j p_j h_j^{x_i} (1 - h_j)^{1 - x_i}$$

Complete liklihood

Two ways to write it: one based on indicator variables and one based on counts.

$$\log P(x, y \mid \theta) = \sum_i \sum_j y_{ij} \log P(x_i, y_i = j \mid \theta) =$$

$$\sum_i \sum_j y_{ij} [\log p_j + x_i \log h_j + (1 - x_i) \log(1 - h_j)]$$

Define $n_{j,H}(y)$ to be the number of heads obtained when coin *j* was used, and similarly define $n_{j,T}(y)$ then:

$$\log P(x, y \mid \theta) = \sum_j [(n_{j,H}(y) + n_{j,T}(y)) \log p_j + n_{j,H}(y) \log h_j + n_{j,T}(y) \log(1 - h_j)]$$

What is the relation between the two, i.e. between the *y* variables and the *n* variables?

Baum-Welch: EM for HMM

y=π, i.e. the log likelihood is

$$\log P(x \mid \theta) = \log \sum_{\pi} P(x, \pi \mid \theta)$$

And the Q function is:

$$Q(\theta \mid \theta^t) = \sum_{\pi} P(\pi \mid x, \theta^t) \cdot \log P(x, \pi \mid \theta)$$

Baum-Welch (cont.)

$$P(x,\pi \mid \theta) = \prod_{k=1}^{M} \prod_{b} [e_k(b)]^{E_k(b,\pi)} \cdot \prod_{k=1}^{M} \prod_{l=1}^{M} a_{kl}^{A_{kl}(\pi)}$$

Emission probability, state k character b

Transition probability, state k to state l

Number of times we saw b from k at π

Baum-Welch (cont.)

$$Q(\theta \mid \theta^t) = \sum_{\pi} P(\pi \mid x, \theta^t) \cdot \left[\sum_{k=1}^{M} \sum_{b} E_k(b,\pi) \cdot \log(e_k(b)) + \sum_{k=1}^{M} \sum_{l=1}^{M} A_{kl}(\pi) \cdot \log a_{kl} \right] =$$

$$= \sum_{k=1}^{M} \sum_{b} \sum_{\pi} P(\pi \mid x, \theta^t) \cdot E_k(b,\pi) \cdot \log(e_k(b)) + \sum_{k=1}^{M} \sum_{l=1}^{M} \sum_{\pi} P(\pi \mid x, \theta^t) \cdot A_{kl}(\pi) \cdot \log a_{kl}$$

$$\sum_{\pi} P(\pi \mid x, \theta^t) \cdot E_k(b,\pi) = E_k(b)$$

probability — value — expectation

$$\sum_{\pi} P(\pi \mid x, \theta^t) \cdot A_{kl}(\pi) = A_{kl}$$

probability — value — expectation

Baum-Welch (cont.)

- So we want to find a set of parameters θ^{t+1} that maximizes:

$$\sum_{k=1}^{M} \sum_{b} E_k(b) \cdot \log(e_k(b)) + \sum_{k=1}^{M} \sum_{l=1}^{M} A_{kl} \cdot \log a_{kl}$$

- $E_k(b)$, A_{kl} can be computed using forward/backward:

$P(\pi_i=k, \pi_{i+1}=l \mid x, \Theta^t) = [1/P(x)] \cdot f_k(i) \cdot a_{kl} \cdot e_l(x_{i+1}) \cdot b_l(i+1)$

$A_{kl} = [1/P(x)] \cdot \sum_i f_k(i) \cdot a_{kl} \cdot e_l(x_{i+1}) \cdot b_l(i+1)$

similarly, $E_k(b) = [1/P(x)] \cdot \sum_{\{i|x_i=b\}} f_k(i) \cdot b_k(i)$

- For maximization, select:

$$a_{ij} = \frac{A_{ij}}{\sum_k A_{ik}} \quad , \quad e_k(b) = \frac{E_k(b)}{\sum_{b'} E_k(b')}$$

Baum-Welch: EM for HMM

Maximize:
$$\sum_{k=1}^{M}\sum_{b} E_k(b)\cdot\log(e_k(b)) + \sum_{k=1}^{M}\sum_{l=1}^{M} A_{kl}\cdot\log a_{kl}$$

Multiply and divide by same factor

$$a_{ij} = \frac{A_{ij}}{\sum_k A_{ik}} \quad \text{(denote as } a_{ij}^{\text{chosen}}\text{)}, \quad e_k(b) = \frac{E_k(b)}{\sum_{b'} E_k(b')}$$

Difference between chosen set and some other:
$$\sum_{k=1}^{M}\sum_{l=1}^{M} A_{kl}\cdot\log\left(\frac{a_{kl}^{chosen}}{a_{kl}^{other}}\right) = \sum_{k=1}^{M}\sum_{k'} A_{ik'}\sum_{l=1}^{M}\frac{A_{kl}}{\sum_{k'} A_{ik'}}\cdot\log\left(\frac{a_{kl}^{chosen}}{a_{kl}^{other}}\right) =$$

$$= \sum_{k=1}^{M}\sum_{k'} A_{ik'}\sum_{l=1}^{M} a_{kl}^{chosen}\cdot\log\left(\frac{a_{kl}^{chosen}}{a_{kl}^{other}}\right)$$

➔ always positive

Parameter Estimation in HMM

Case 2: -Estimation When States are Unknown

Input: $X^1, \ldots, X^n$ indep training sequences

Baum-Welch alg. (1972):

✶ **Expectation**:

- compute expected no. of $k \to l$ state transitions:

  $P(\pi_i=k,\ \pi_{i+1}=l \mid X,\ \Theta) = [1/P(x)] \cdot f_k(i) \cdot a_{kl} \cdot e_l(x_{i+1}) \cdot b_l(i+1)$

  ⇨ $A_{kl} = \sum_j [1/P(X^j)] \cdot \sum_i f_k^j(i) \cdot a_{kl} \cdot e_l(x^j_{i+1}) \cdot b_l^j(i+1)$

- compute expected no. of symbol b appearances in state k

  $E_k(b) = \sum_j [1/P(X^j)] \cdot \sum_{\{i|x^j_i=b\}} f_k^j(i) \cdot b_k^j(i)$ *(ex.)*

✶ **Maximization**:

- re-compute new parameters from A, E using max. likelihood.

repeat (1)+(2) until improvement $\leq \varepsilon$