CISC 636 Computational Biology & Bioinformatics
(Fall 2016)

Hidden Markov Models (III)

- Viterbi training
- Baum-Welch algorithm
- Maximum Likelihood
- Expectation Maximization

Model building

- Topology
  - Requires domain knowledge
- Parameters
  - When states are labeled for sequences of observables
    - Simple counting (Maximum Likelihood):

$$a_{kl} = A_{kl} / \Sigma_{l'} A_{kl'} \text{ and } e_k(b) = E_k(b) / \Sigma_{b'} E_k(b')$$

  - When states are not labeled

  Method 1 (Viterbi training)

  1. Assign random parameters
  2. Use Viterbi algorithm for labeling/decoding
  2. Do counting to collect new a_{kl} and $e_k(b)$;
  3. Repeat steps 2 and 3 until stopping criterion is met.

  Method 2 (Baum-Welch algorithm)

Baum-Welch algorithm (Expectation-Maximization)

- An iterative procedure similar to Viterbi training
- Probability that a_{kl} is used at position i in sequence j.

$$P(\pi_i = k, \pi_{i+1} = l \mid x, \theta) = f_k(i)\, a_{kl}\, e_l(x_{i+1})\, b_l(i+1) / P(x^j)$$

Calculate the expected number of times that is used by summing over all position and over all training sequences.

$$A_{kl} = \Sigma_j \{(1/P(x^j)\, [\Sigma_i f_k^j(i)\, a_{kl}\, e_l(x^j_{i+1})\, b_l^j(i+1)] \}$$

Similarly, calculate the expected number of times that symbol b is emitted in state k.

$$E_k(b) = \Sigma_j \{(1/P(x^j)\, [\Sigma_{\{i|x\_i\text{^}j = b\}} f_k^j(i)\, b_k^j(i)] \}$$

Maximum Likelihood

Define $L(\theta) = P(x|\theta)$

Estimate θ such that the distribution with the estimated θ best agrees with or support the data observed so far.

$$\theta^{ML} = \underset{\theta}{\operatorname{argmax}}\, L(\theta)$$

E.g. There are red and black balls in a box. What is the probability P of picking up a black ball?

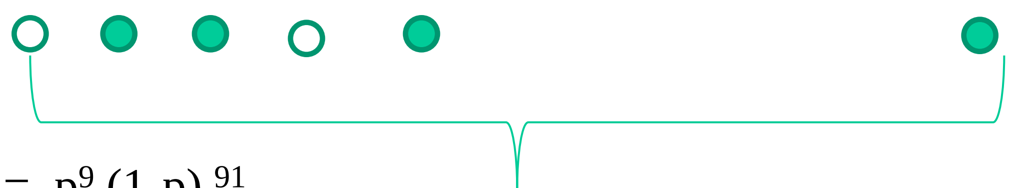
Do sampling (with replacement).

Maximum Likelihood

Define $L(\theta) = P(x|\ \theta)$
Estimate such that the distriibution with the estimated best agrees with or supports the data observed so far.

$\theta^{ML} = \text{argmax}_\theta\ L(\theta)$
When $L(\theta)$ is differentiable,
$$\frac{\partial L(\theta)}{\partial \theta}_{|\theta^{ML}} = 0$$

For example, want to know the ratio: # of blackball/# of whiteball, in other words, the probability P of picking up a black ball. Sampling (with replacement):



100 times

Counts: whiteball 91, blackball 9

Prob (iid) = $p^9\ (1\text{-}p)^{91}$
Likelihood $L(p) = p^9(1\text{-}p)^{91}$.

$$\frac{\partial L(p)}{\partial P} = 9p^8(1-p)^{91} - 91p^9(1-p)^{90} = 0$$

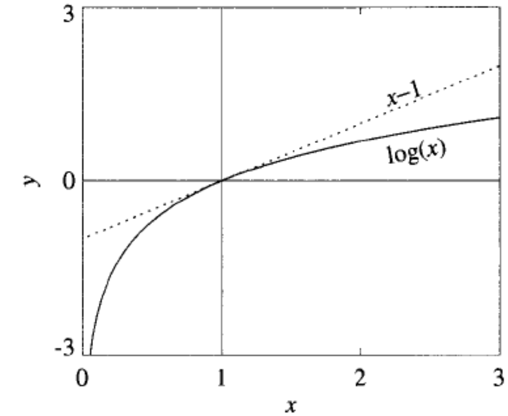
=> $P^{ML} = 9/100 = 9\%$.   The ML estimate of P is just the frequency.

A proof that the observed frequency -> ML estimate of probabilities for polynomial distribution

Let Counts n_i for outcome i

The observed frequencies $\theta_i = n_i /N$, where $N = \sum_i n_i$

If $\theta_i^{ML} = n_i /N$, then $P(n|\theta^{ML}) > p(n|\theta)$ for any $\theta \neq \theta^{ML}$



Proof:

$$\log\frac{P(n\,|\,\theta^{ML})}{P(n\,|\,\theta)} = \log\frac{\prod_i (\theta_i^{ML})^{n_i}}{\prod_i (\theta_i)^{n_i}} = \log\prod_i \left(\frac{\theta_i^{ML}}{\theta_i}\right)^{n_i}$$

$$= \sum_i n_i \log\left(\frac{\theta_i^{ML}}{\theta_i}\right) = \mathrm{N}\sum_i \frac{n_i}{N}\log\left(\frac{\theta_i^{ML}}{\theta_i}\right) = \sum_i \theta_i^{ML}\log\left(\frac{\theta_i^{ML}}{\theta_i}\right)$$

$$= \mathrm{H}(\theta^{ML}\,||\,\theta) \geq 0$$

Maximum Likelihood: pros and cons

- Consistent, i.e., in the limit of a large amount of data, ML estimate converges to the true parameters by which the data are created.
- Simple
- Poor estimate when data are insufficient.

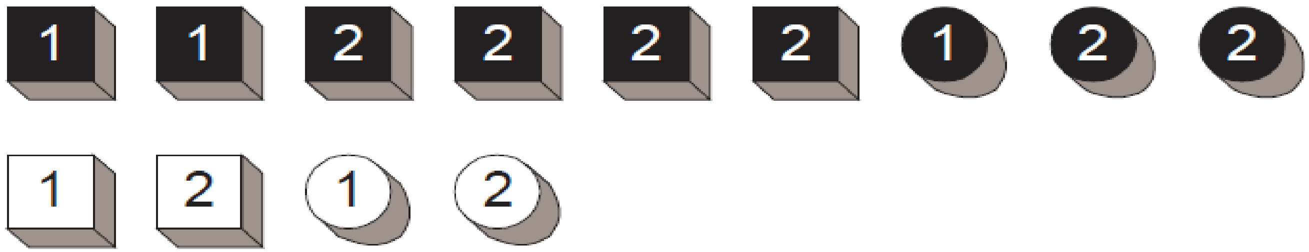
  e.g., if you roll a die for less than 6 times, the ML estimate for some numbers would be zero.

Pseudo counts:

$$\theta_i = \frac{n_i + \alpha_i}{N + A},$$

where $A = \sum_i \alpha_i$

Conditional Probability and Join Probability



P(one) = 5/13

P(square) = 8/13

P(one, square) = 3/13

P(one | square) = 3/8 = P(one, square) / P(square)

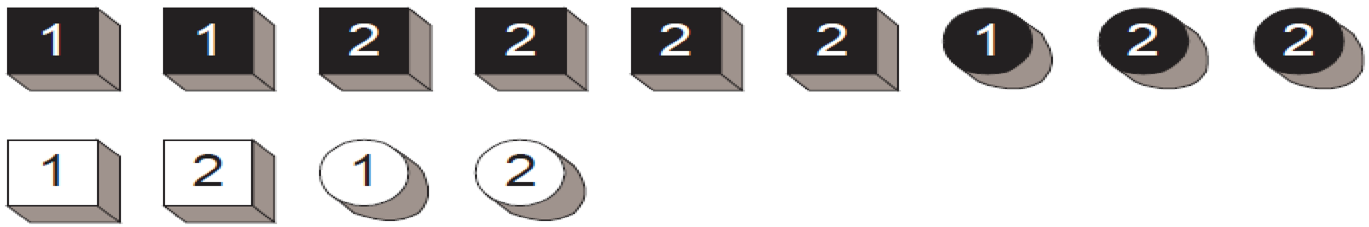
In general, P(D,M) = P(D|M)P(M) = P(M|D)P(D)

=> **Baye's Rule:**

$$P(\mathrm{M} \mid \mathrm{D}) = \frac{P(D \mid M)P(M)}{P(D)}$$

$$\begin{aligned} P(\mathrm{One}|\mathrm{Black}) &= \frac{P(\mathrm{Black}|\mathrm{One})P(\mathrm{One})}{P(\mathrm{Black}|\mathrm{One})P(\mathrm{One}) + P(\mathrm{Black}|\mathrm{Two})P(\mathrm{Two})} \\ &= \frac{(\frac{3}{5})(\frac{5}{13})}{(\frac{3}{5})(\frac{5}{13}) + (\frac{6}{8})(\frac{8}{13})} = \frac{1}{3}, \end{aligned}$$

Conditional Probability and Conditional Independence



$$
\begin{aligned}
P(\text{One}) &= \frac{5}{13} \\
P(\text{One}|\text{Square}) &= \frac{3}{8}
\end{aligned}
$$

$$
\begin{aligned}
P(\text{One}|\text{Black}) &= \frac{3}{9} = \frac{1}{3} \\
P(\text{One}|\text{Square} \cap \text{Black}) &= \frac{2}{6} = \frac{1}{3}
\end{aligned}
$$

$$
\begin{aligned}
P(\text{One}|\text{White}) &= \frac{2}{4} = \frac{1}{2} \\
P(\text{One}|\text{Square} \cap \text{White}) &= \frac{1}{2}.
\end{aligned}
$$

So One *and* Square *are not independent, but they are conditionally independent given* Black *and given* White.

Baye's Rule:

$$P(\mathrm{M} \mid \mathrm{D}) = \frac{P(D \mid M)P(M)}{P(D)}$$

Example: disease diagnosis/inference

P(Leukemia | Fever) = ?

P(Fever | Leukemia) = 0.85

P(Fever) = 0.9

P(Leukemia) = 0.005

P(Leukemia | Fever) = P(F|L)P(L)/P(F) = 0.85*0.01/0.9 = 0.0047

Bayesian Inference
Maximum a posterior estimate

$$\theta^{MAP} = \arg\max_{\theta} P(\theta \mid \mathrm{x})$$

Expectation Maximization

$$\mathrm{P}(x, y \mid \theta) = P(y \mid x, \theta) P(\mathrm{x} \mid \theta)$$

$$P(x \mid \theta) = P(x, y \mid \theta) / P(y \mid x, \theta)$$

$$\log P(x \mid \theta) = \log P(x, y \mid \theta) - \log P(\mathrm{y} \mid \mathrm{x}, \theta)$$

$$\sum_y P(y \mid x, \theta^t) \quad (\qquad\qquad\qquad\qquad)$$

Expectation

$$\log P(x \mid \theta) = \sum_y P(y \mid x, \theta^t) \log P(x, y \mid \theta) - \sum_y P(y \mid x, \theta^t) \log P(y \mid \mathrm{x}, \theta)$$

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

$$\log P(x \mid \theta) - \log P(x \mid \theta^t)$$

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

$$\geq Q(\theta \mid \theta^t) - Q(\theta^t \mid \theta^t)$$

$$\theta^{t+1} = \arg\max_\theta Q(\theta \mid \theta^t)$$

Maximization

EM explanation of the Baum-Welch algorithm

We like to maximize by choosing θ

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

But state path π is hidden variable. Thus, EM.

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

$$P(x, \pi \mid \theta) = \prod_{k=1}^{M} \prod_{b} [e_k(b)]^{E_k(b,\pi)} \prod_{k=0}^{M} \prod_{l=1}^{M} a_{kl}^{A_{kl}(\pi)},$$

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

$$\left[\sum_{k=1}^{M} \sum_{b} E_k(b, \pi) \log e_k(b) + \sum_{k=0}^{M} \sum_{l=1}^{M} A_{kl}(\pi) \log a_{kl} \right].$$

EM Explanation of the Baum-Welch algorithm

$$E_k(b) = \sum_{\pi} P(\pi|x,\theta^t) E_k(b,\pi) \quad \text{and} \quad A_{kl} = \sum_{\pi} P(\pi|x,\theta^t) A_{kl}(\pi).$$

$$Q(\theta|\theta^t) = \sum_{k=1}^{M} \sum_{b} E_k(b) \log e_k(b) + \sum_{k=0}^{M} \sum_{l=1}^{M} A_{kl} \log a_{kl}.$$

E-term A-term

A-term is maximized if $a_{kl}^{EM} = \dfrac{A_{kl}}{\sum_{l'} A_{kl'}}$

E-term is maximized if $e_k^{EM}(b) = \dfrac{E_k(b)}{\sum_{b'} E_k(b')}$