Data Mining and Machine Learning

Statistical Modelling of Sequences (2)

Peter Jančovič

UNIVERSITY OF BIRMINGHAM

Objectives

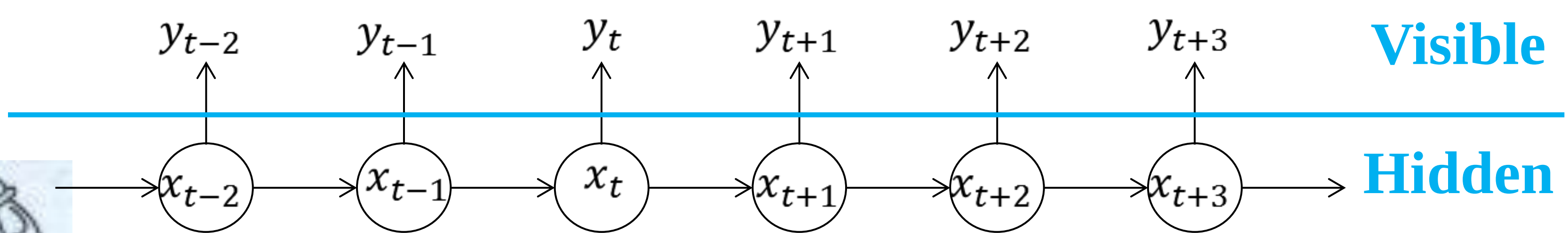
- So far, we introduced Markov models
- Hidden Markov models (HMMs)
- Calculating the probability of an observation sequence
- The Forward Probability calculation
- HMM training

Hidden Markov Models (HMMs)

- Let's go back to our original shopping example
- Suppose that when a shopper visits a shop he or she makes a single purchase from a set of M possible items $I_1, \ldots I_M$
- Suppose that, instead of observing the sequence of shops, we observe the sequence of purchases
- Because different shops may sell the same item it is in general not possible to know the shop sequence unambiguously from the purchase sequence
- This is an example of a <u>Hidden</u> Markov process

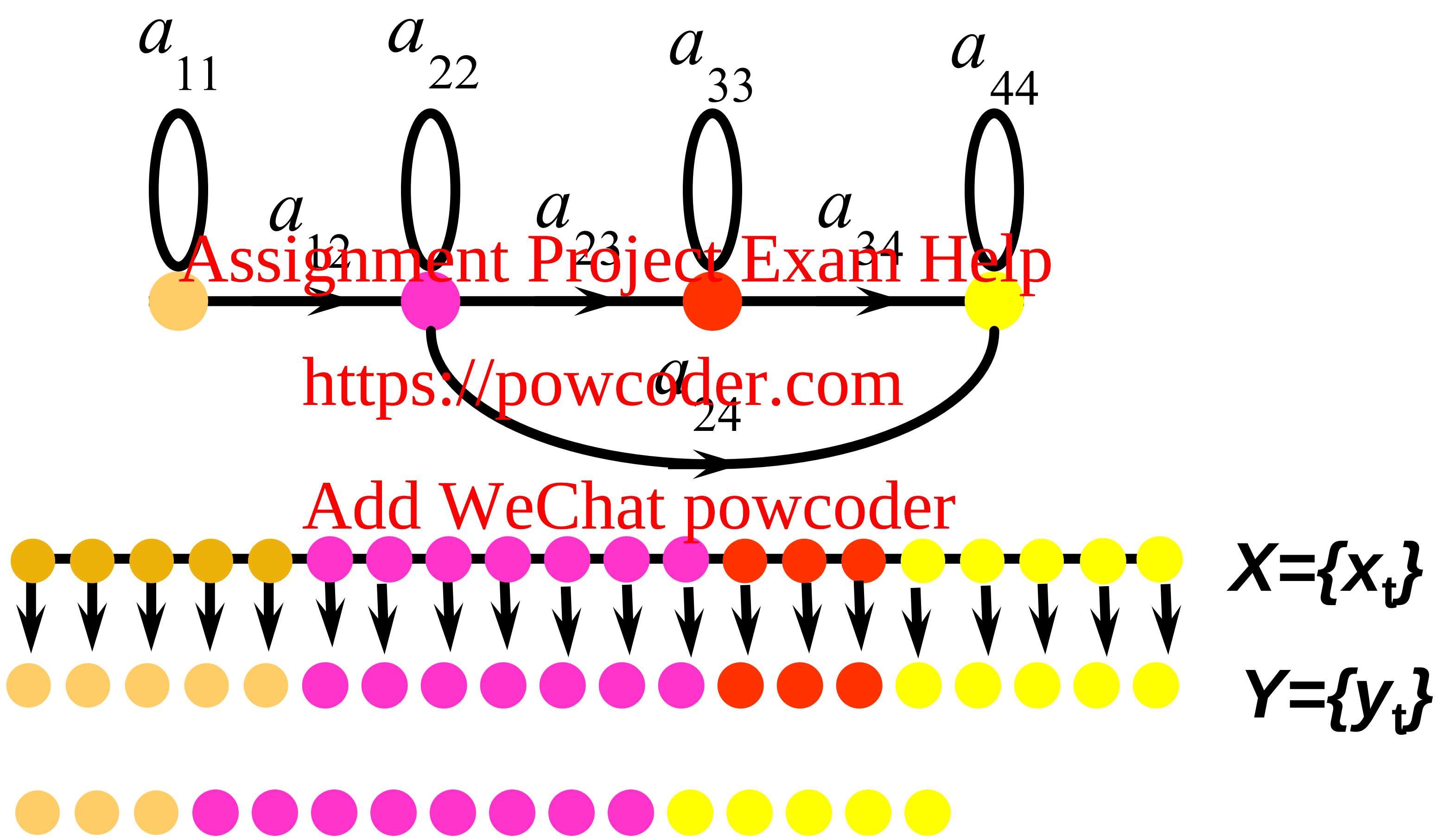
HMMs (continued)

- In a HMM we assume that the current item purchased depends only on the current state (shop) and not on items previously purchased or shops previously visited
- Suppose x_t is the state and y_t is the item purchased at time t
- The diagram indicates the dependencies:



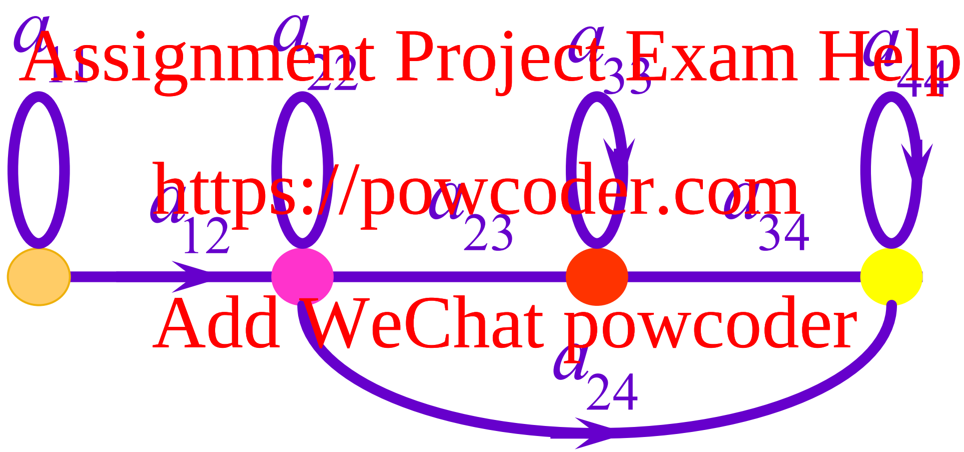
UNIVERSITY OF BIRMINGHAM

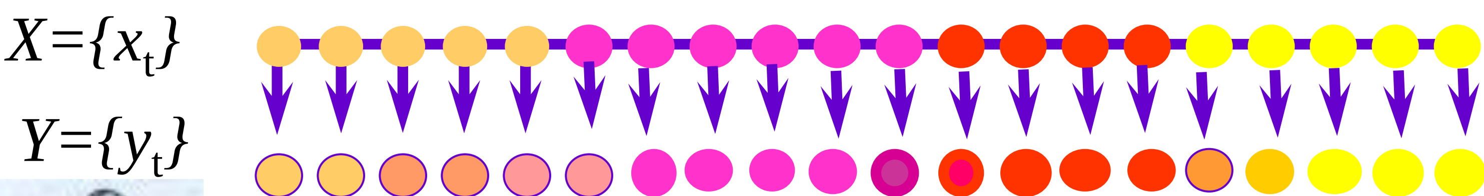
Markov Model



Hidden Markov Model

- In a **hidden** Markov model, the relationship between the observation sequence and the state sequence is ambiguous.





HMMs Continued

- Let B_n be the probability distribution for items bought in shop S_n ($n=1,\ldots,N$)
- Then $B_n = [B_n(1) \ldots B_n(m) \ldots B_n(M)]$, where $B_n(m)$ is the probability of buying item I_m in shop S_n
- Or (better), $B_n(m) = P(y_t = I_m | x_t = S_n)$. Note that this is independent of t.
- We can write all of these probabilities as a $N \times M$ matrix B whose n^{th} row is B_n.

Formal definition of a HMM

- An N state HMM with observations $\{I_1,\ldots,I_M\}$ comprises:
- An underlying N state Markov model, defined by an initial state probability vector P_0 and $N\times N$ state transition probability matrix A, where:
  - $P_0(n) = P(x_1 = S_n)$
  - $A_{nm} = P(x_t = S_m \mid x_{t-1} = S_n)$
- An $N\times M$ state output probability matrix B where
  - $B_{nm} = B_n(m) = P(y_t = I_m \mid x_t = S_n)$

Example HMM Probability Calculation

- Let's start with our simple 3 state Markov model

$$P_0 = \begin{bmatrix} 0.5 \\ 0.2 \\ 0.3 \end{bmatrix}, A = \begin{bmatrix} 0.5 & 0.2 & 0.3 \\ 0.1 & 0.1 & 0.8 \\ 0.4 & 0 & 0.6 \end{bmatrix}$$

- In addition let's suppose that there are 4 possible items $I_1,\ldots,I_4$ that can be purchased. We need to specify the probabilities $B_n(m)$ for $n = 1, 2, 3$ and $m = 1, 2, 3, 4$. Suppose

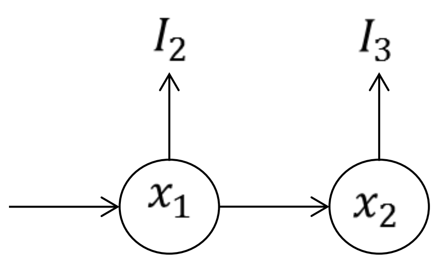
$$B = \begin{bmatrix} 0.6 & 0.1 & 0.1 & 0.2 \\ 0.1 & 0.1 & 0.1 & 0.7 \\ 0.2 & 0.2 & 0.3 & 0.3 \end{bmatrix}$$

Example (Continued)

- What is the probability of observing the sequence $I = I_2 I_3$?
- This sequence must correspond to an underlying state sequence $x = x_1 x_2$. Suppose $x_1 = S_1, x_2 = S_2$

- Then $P(I, x) = P(I|x)P(x)$

$$
\begin{aligned}
&= P(I_2|x_1) \times P(I_3|x_2) \times P_0(x_1) \times P(x_2|x_1) \\
&= B_1(2) \times B_2(3) \times P_0(1) \times a_{12} \\
&= 0.1 \times 0.1 \times 0.5 \times 0.2 = 0.001
\end{aligned}
$$



Example (Continued)

- So, $P(I, x) = 0.001$
- But $S_1 S_2$ is just one of the state sequences that could have generated I. It could also have arisen from $S_1 S_1$ or $S_1 S_3$ or $S_2 S_1$ or $S_2 S_2$ or $S_2 S_3$ or $S_3 S_1$ or $S_3 S_2$ or $S_3 S_3$!
- As always, when calculating the probability of an event I which may have arisen through a number of ways x, we have to sum the joint probability $P(I,x)$ over all possible values of x. In other words:

Example (Continued)

- So,
  - $P(I, S_1 S_1) = 0.0025$
  - $P(I, S_1 S_2) = 0.0010$
  - $P(I, S_1 S_3) = 0.0045$
  - $P(I, S_2 S_1) = 0.0002$
  - $P(I, S_2 S_2) = 0.0002$
  - $P(I, S_2 S_3) = 0.0048$
  - $P(I, S_3 S_1) = 0.0024$
  - $P(I, S_3 S_2) = 0$
  - $P(I, S_3 S_3) = 0.0108$

Hence $\underline{P(I) = 0.0264}$

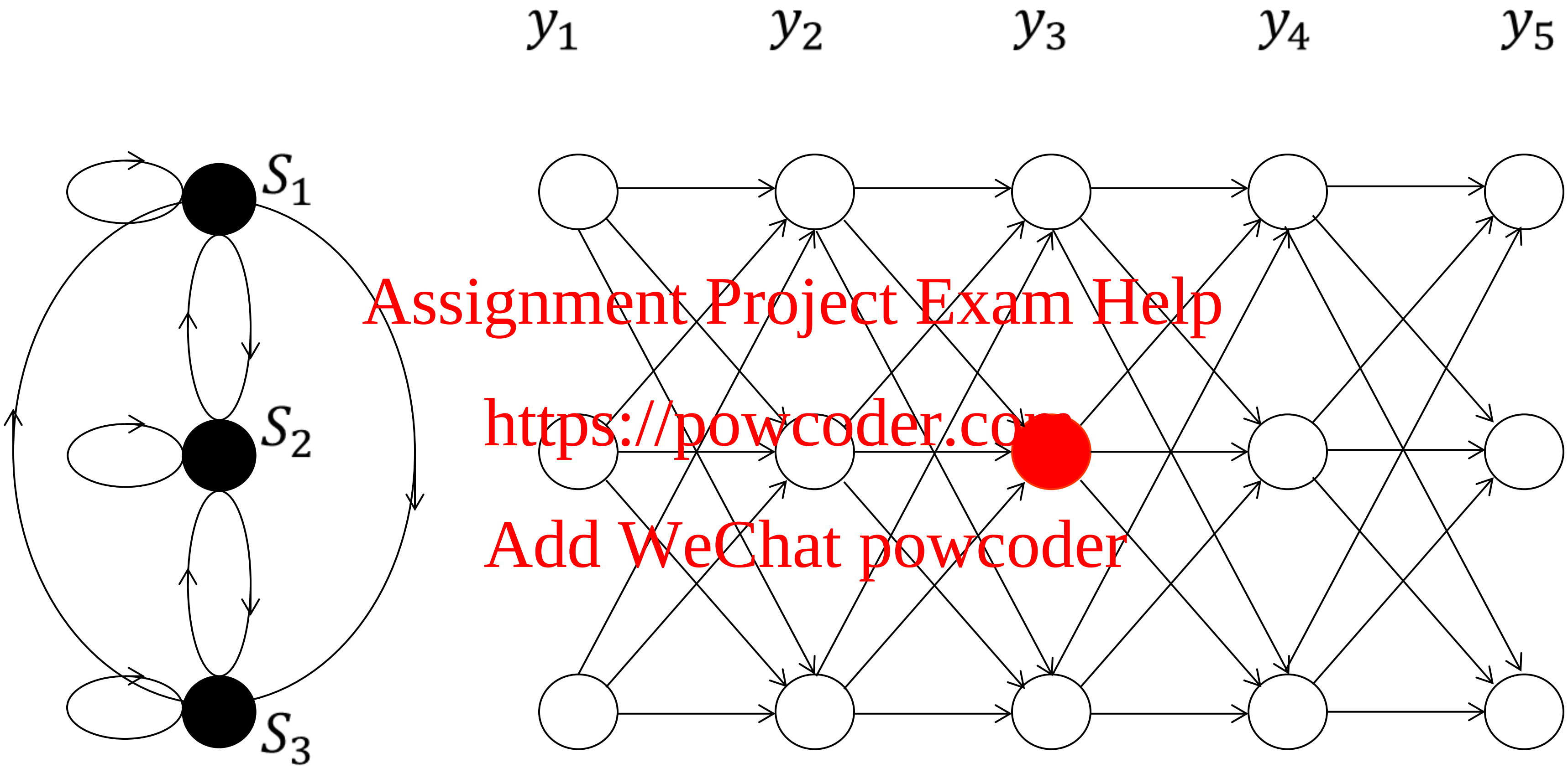
Calculating the probability of an observed sequence

- Even in our simple example with 3 states and 2 observations there are 9 terms in the summation
- In general, if the Markov model is fully connected and has N states, and we have T observations, then the number of state sequences (and therefore the number of terms in the summation) is N^T. This makes direct calculation of $P(I)$ computationally impractical.
- However, there is an efficient solution….

The Forward Probability calculation

- This is very similar to Dynamic Programming
- Given a sequence of observations $y_1, y_2, \ldots, y_T$, for each t and i define $\alpha_t(i) = P(y_1, y_2, \ldots, y_t, x_t = S_i)$
- In words, $\alpha_t(i)$ is the probability that the sub-sequence $y_1, y_2, \ldots, y_t$ is observed <u>and</u> the Markov process is in state S_i at time t.
- This is easier to understand with a picture…

Graphical interpretation of $\alpha_t(i)$



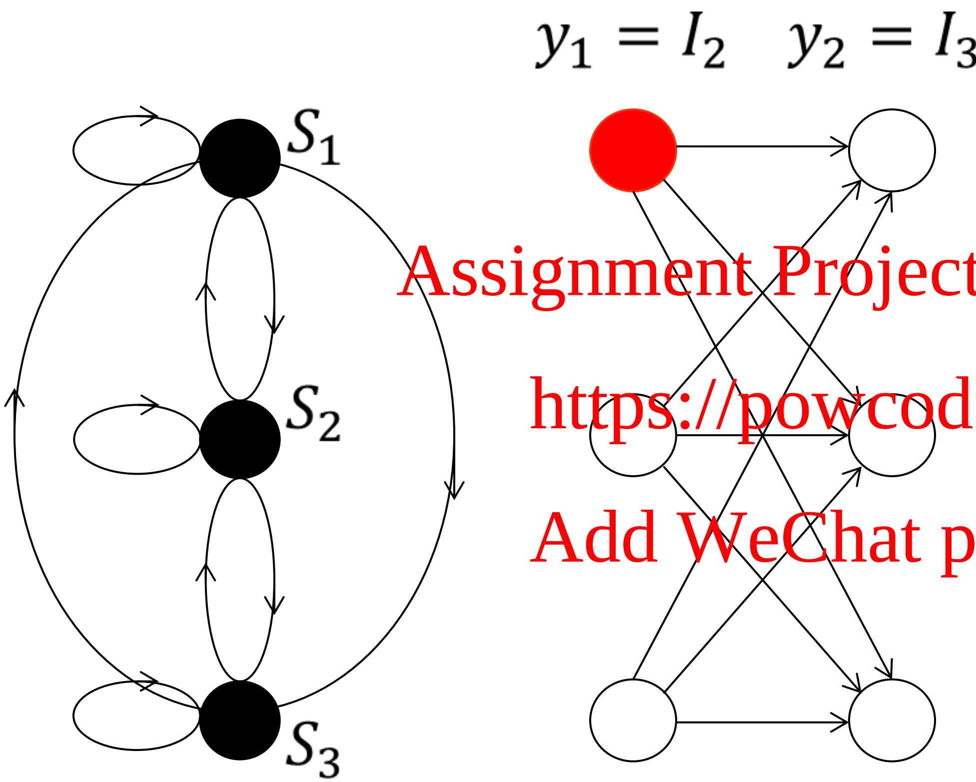
● Corresponds to $\alpha_3(2)$

Recursive equation for $\alpha_t(i)$

- From the diagram,

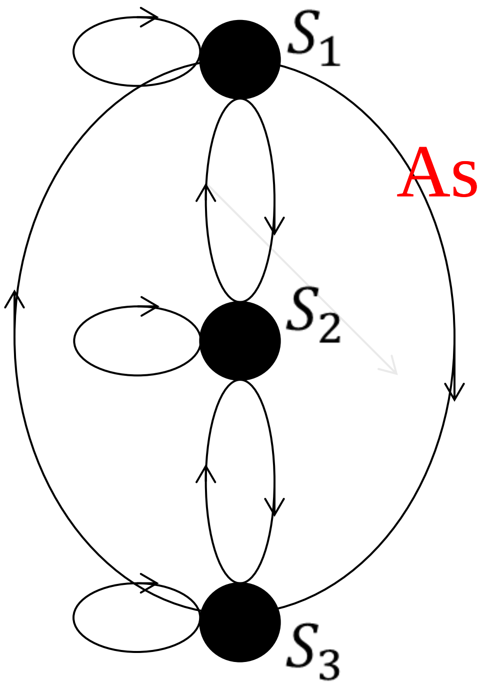
$$\alpha_t(i) = \sum_{j=1}^{N} \alpha_{t-1}(j)\, a_{ji} B_i(y_t)$$

Example

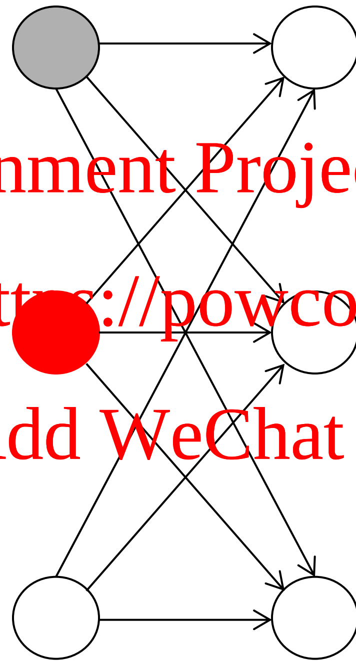


$\alpha_1(1) = P_0(1) \times B_1(2) = 0.05$

Example



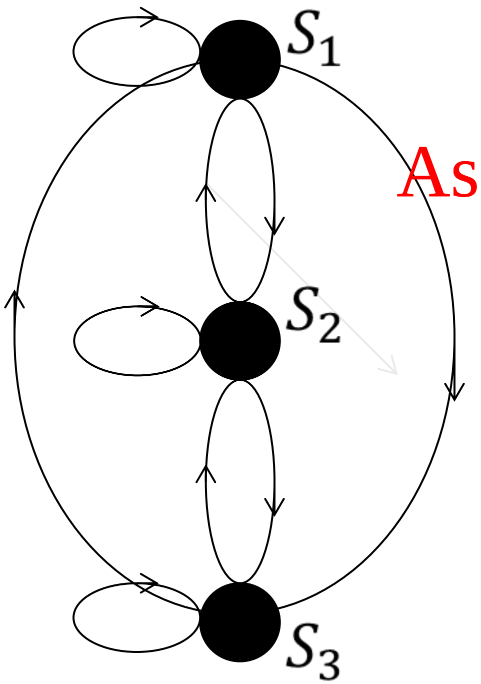
$y_1 = I_2 \quad y_2 = I_3$



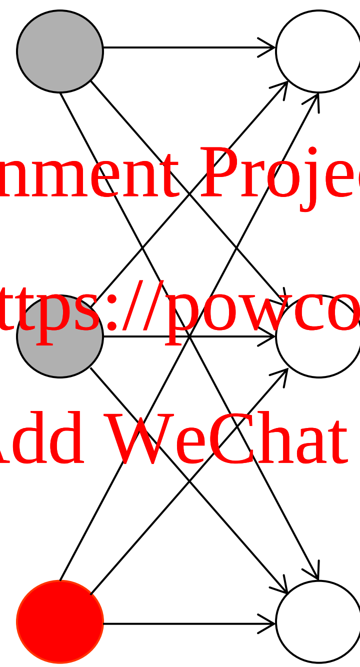
$\alpha_1(1) = 0.05$

$\alpha_1(2) = P_0(2) \times B_2(2) = 0.02$

Example



$y_1 = I_2 \quad y_2 = I_3$

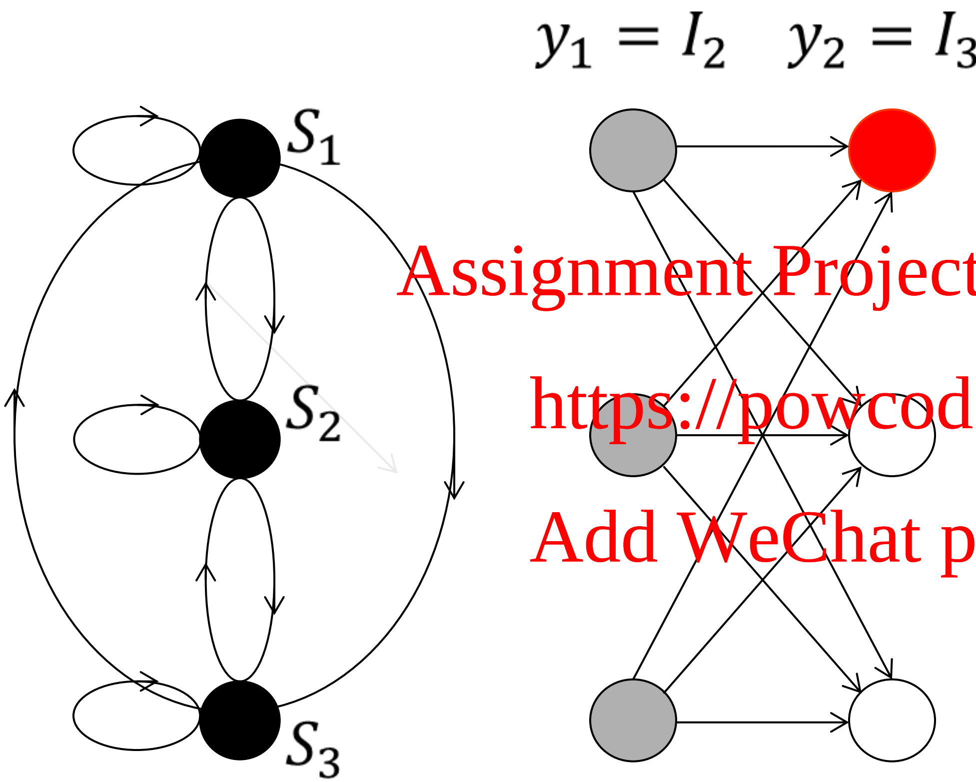


$\alpha_1(1) = B_1(2) = 0.05$

$\alpha_1(2) = B_2(2) = 0.02$

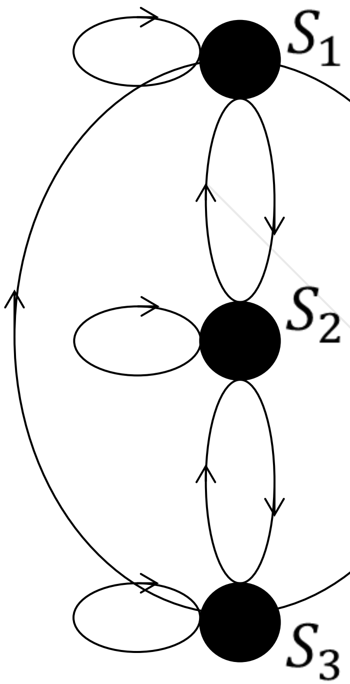
$\alpha_1(3) = P_0(3)B_3(2) = 0.06$

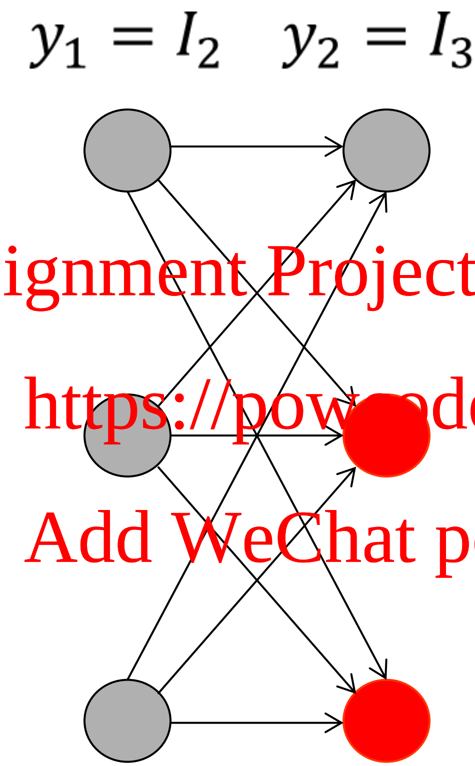
Example



$$\alpha_2(1) =$$
$$\alpha_1(1) \times a_{11} \times B_1(3)$$
$$+\alpha_1(2) \times a_{21} \times B_1(3)$$
$$+\alpha_1(3) \times a_{31} \times B_1(3) =$$
$$0.05 \times 0.5 \times 0.1$$
$$+0.02 \times 0.1 \times 0.1$$
$$+0.06 \times 0.4 \times 0.1 =$$
$$0.0025 + 0.0002 + 0.0024 =$$
$$\underline{0.0051}$$

Example





Similarly:

$\alpha_2(2) = 0.0012,$
$\alpha_2(3) = 0.0201$

Hence,

$$P(y_1 y_2) = P(y_1, y_2, x_2 = S_1) + P(y_1, y_2, x_2 = S_2) + P(y_1, y_2, x_2 = S_3) = \alpha_2(1) + \alpha_2(2) + \alpha_2(3) = 0.0051 + 0.0012 + 0.0201$$

$$\underline{= \mathbf{0.0264}}$$

HMM Parameter Estimation

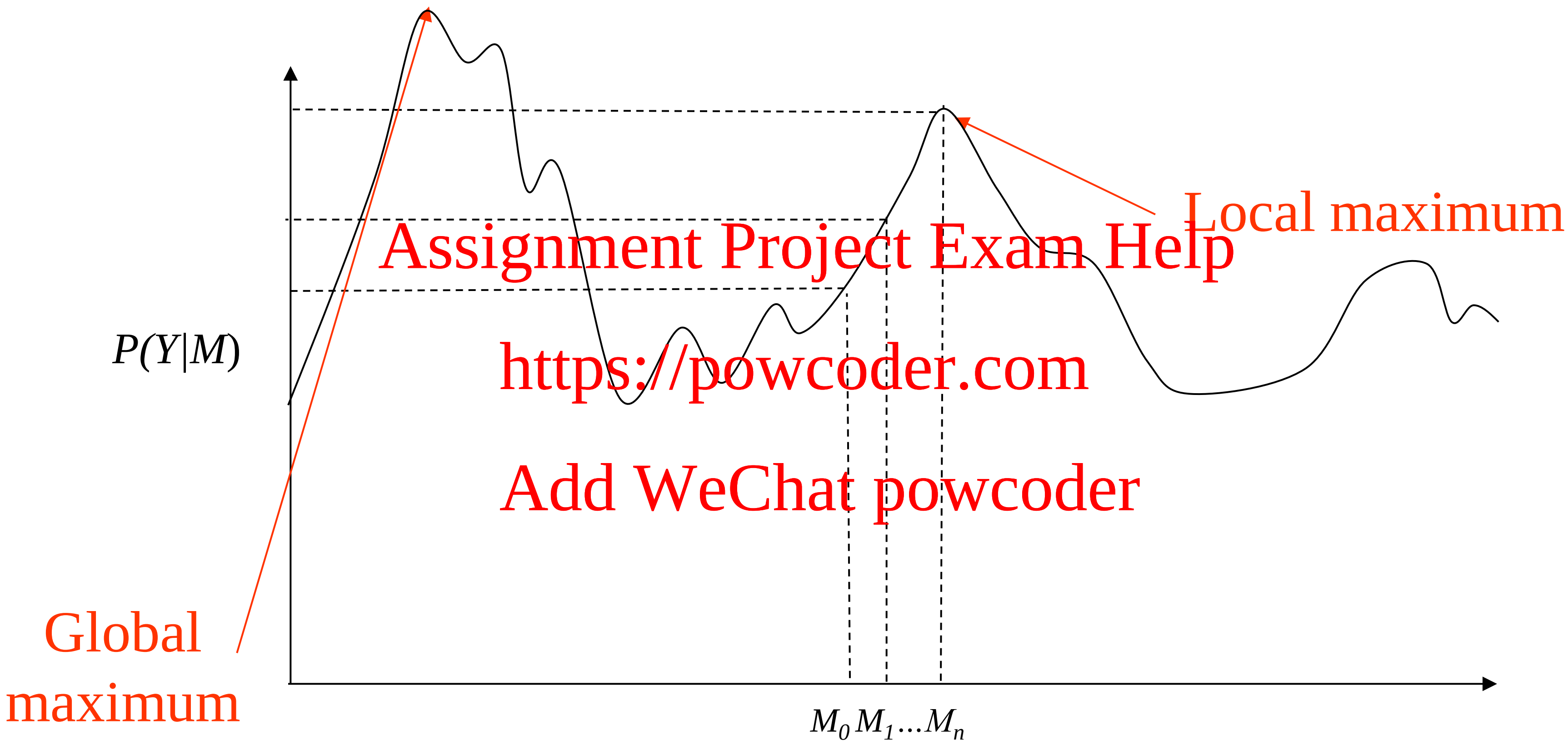
- Given a HMM and a sequence *y* we can calculate *P*(*y*)
- But where does the HMM come from? In other words how do we estimate the HMM's parameters?
- This is done from data, using an algorithm similar to the E-M algorithm for estimating the parameters of a GMM
- The HMM training algorithm is called the Baum-Welch algorithm
- Like the E-M algorithm, it involves making an initial estimate and then iteratively improving the estimate until convergence. Hence it is only <u>locally</u> optimal

HMM training

1. Make an initial estimate of the HMM – M_0
2. Obtain a large set of training data Y
3. Set $i=1$
4. Apply the Baum-Welch algorithm to Y and M_{i-1} to get a new model M_i such that $P(Y|M_i) \geq P(Y|M_{i-1})$
5. If $|P(Y|M_i) - P(Y|M_{i-1})| \leq \varepsilon$ then stop, else
    1. $i = i+1$
    2. Go back to step 4.

UNIVERSITY OF BIRMINGHAM

Local optimality



Summary

- Hidden Markov Models
- Calculating the probability of an observation sequence
- The forward probability calculation
- HMM training