Machine Learning

10-701/15-781, Spring 2010

Hidden Markov Model

Eric Xing

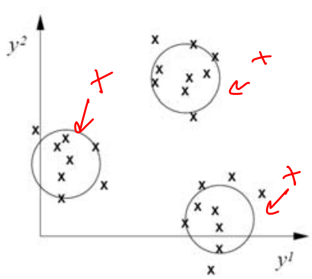
Lecture 11, February 22, 2010

Reading: Chap. 13 CB

Motivation

- Clustering:

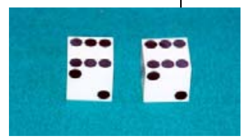
$\{x_n, y_n\}$



$P(y_n | x_n, \mu, \Sigma)$

A somewhat similar problem

An experience in a casino



<u>Game:</u>

1. You bet $1
2. You roll (always with a fair die)
3. Casino player rolls (maybe with fair die, maybe with loaded die)
4. Highest number wins $2

6 . 3

Question:

1

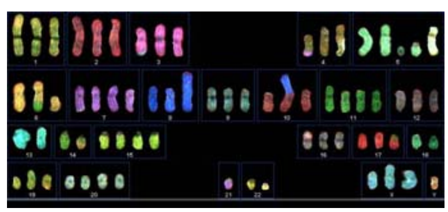
1245526462146146136136661664661636
616366163616515615115146123562344

6

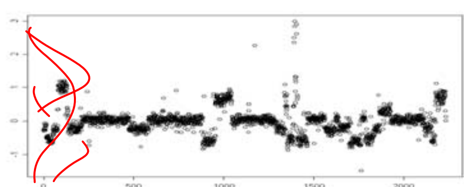
Which die is being used in each play?

Question:

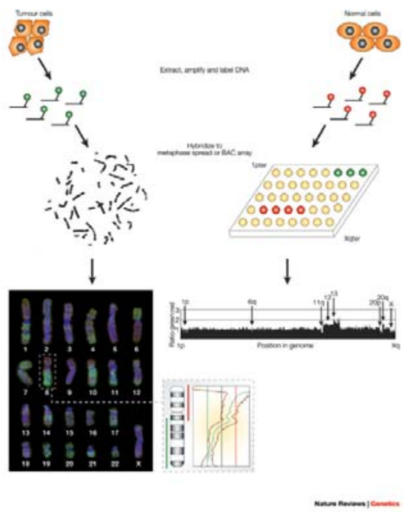
- Naturally, data points arrive one at a time
  - Does the ordering index carry (additional) clustering information besides the data value itself ?
  - Example:
    Chromosomes of tumor cell:



Copy number measurements (known as CGH)



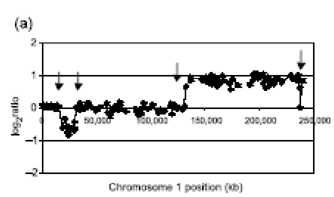
Array CGH (comparative genomic hybridization)



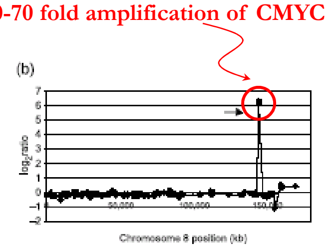
- The basic assumption of a CGH experiment is that the ratio of the binding of test and control DNA is proportional to the ratio of the copy numbers of sequences in the two samples.
- But various kinds of noises make the true observations less easy to interpret …

DNA Copy number aberration types in breast cancer

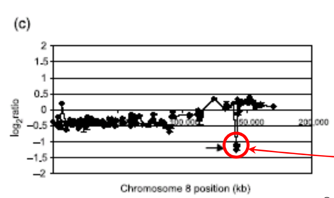
60-70 fold amplification of CMYC region



Copy number profile for chromosome 1 from 600 MPE cell line



Copy number profile for chromosome 8 from COLO320 cell line



Copy number profile for chromosome 8 in MDA-MB-231 cell line

deletion

Question:

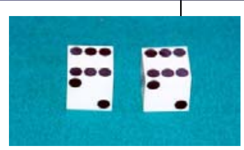
- Sometimes, just data value by itself is hardly clusterable!

  - Unlike continuous vectors, which can take different values in an “infinite” space, and often naturally settle to different cluster just due to value differences, entities with discrete attributes often can not manifest their labels by a one time snapshot of their discrete values alone, sometime additional information is needed …
  - e.g.,

124552646214614613613666166466163661636616361651561511514612356234 4

Suppose you were told about the following story before heading to Vegas…

The Dishonest Casino !!!



A casino has two dice:

- **Fair die**
  P(1) = P(2) = P(3) = P(5) = P(6) = 1/6
- **Loaded die**
  P(1) = P(2) = P(3) = P(5) = 1/10
  P(6) = 1/2

Casino player switches back-&-forth between fair and loaded die once every 20 turns

Puzzles Regarding the Dishonest Casino

GIVEN: A sequence of rolls by the casino player

1245526462146146136136661664661636616366163616515615115146123562344

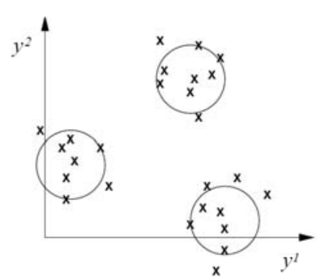
? ? ?

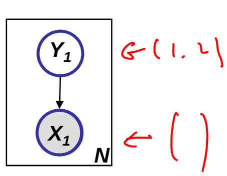
QUESTION

- How likely is this sequence, given our model of how the casino works?
  - This is the **EVALUATION** problem
- What portion of the sequence was generated with the fair die, and what portion with the loaded die?
  - This is the **DECODING** question
- How "loaded" is the loaded die? How "fair" is the fair die? How often does the casino player change from fair to loaded, and back?
  - This is the **LEARNING** question

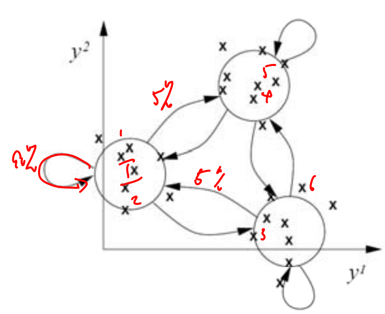
From static to dynamic mixture models

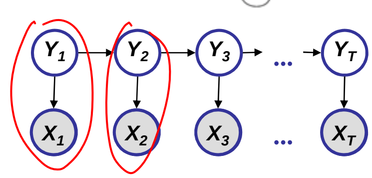
Static mixture





Dynamic mixture

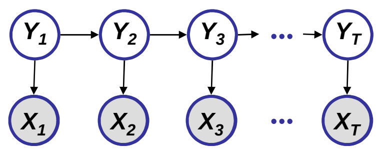




Hidden Markov Models

The underlying source:
genomic entities,
dice,

The sequence:
CGH signal,
sequence of rolls,

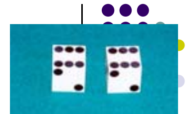


$P(y_{t+1} | y_t, y_{t-1}, \cdots y_{t-2} \cdots y_1)$

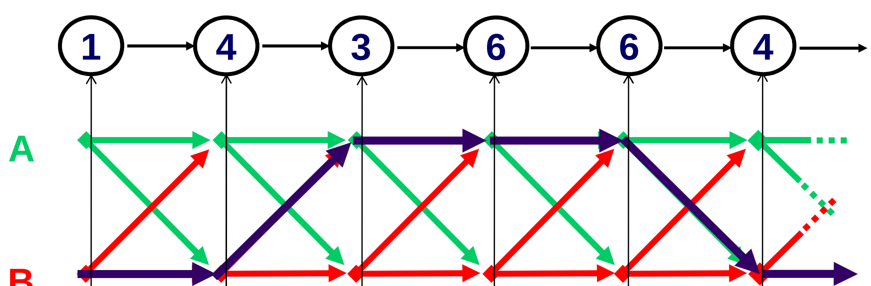
$= P(y_{t+1} | y_t)$

Markov property:

An HMM is a Stochastic Generative Model

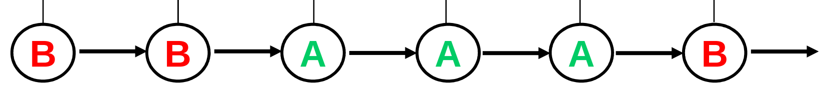


- Observed sequence:



- Hidden sequence (a parse or segmentation):

Y=



Definition (of HMM)

- Observation space
  - **Alphabetic set:** $\mathrm{C}=\{c_1, c_2, \cdots, c_K\}$
  - **Euclidean space:** $\mathbb{R}^d$
- Index set of hidden states
  $\mathrm{I}=\{1, 2, \cdots, M\}$
- Transition probabilities between any two states

  $\boxed{p(y_t^j = 1 \mid y_{t-1}^i = 1) = a_{i,j},}$

  or $p(y_t \mid y_{t-1}^i = 1) \sim \text{Multinomial}(a_{i,1}, a_{i,2}, \ldots, a_{i,M}), \forall i \in \mathrm{I}.$
- Start probabilities

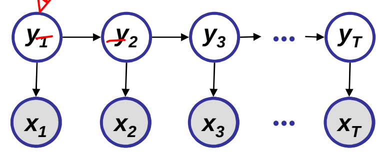
  $p(y_1) \sim \text{Multinomial}(\pi_1, \pi_2, \ldots, \pi_M).$
- Emission probabilities associated with each state

  $p(x_t \mid y_t^i = 1) \sim \text{Multinomial}(b_{i,1}, b_{i,2}, \ldots, b_{i,K}), \forall i \in \mathrm{I}.$

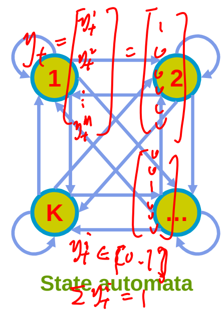
  or in general:

  $p(x_t \mid y_t^i = 1) \sim \mathrm{f}(\cdot \mid \theta_i), \forall i \in \mathrm{I}.$

a_{ij}



Graphical model



State automata

The Dishonest Casino Model

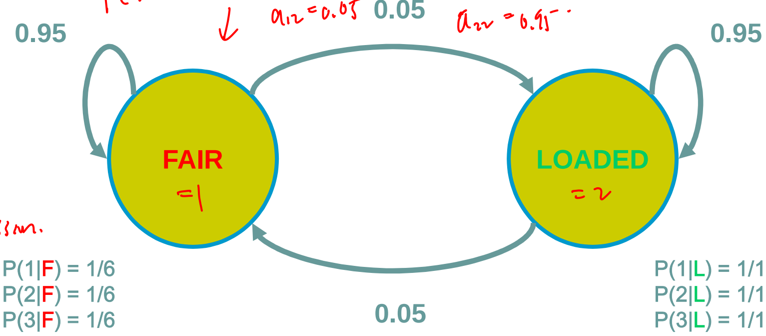
Transition:

$p(y_t \mid y_{t-1})$

$a_{11} = 0.95$
$a_{12} = 0.05$

$a_{21} = .05$
$a_{22} = 0.95$

$\pi_1 = 0.5$
$\pi_2 = 0.5$



Emission.

P(1|F) = 1/6
P(2|F) = 1/6
P(3|F) = 1/6
P(4|F) = 1/6
P(5|F) = 1/6
P(6|F) = 1/6

P(1|L) = 1/10
P(2|L) = 1/10
P(3|L) = 1/10
P(4|L) = 1/10
P(5|L) = 1/10
P(6|L) = 1/2

Three Main Questions on HMMs

1. Evaluation

   GIVEN an HMM $\boldsymbol{M}$, and a sequence $\boldsymbol{x}$,
   FIND Prob ($\boldsymbol{x}$ | $\boldsymbol{M}$)
   ALGO. Forward

2. Decoding

   GIVEN an HMM $\boldsymbol{M}$, and a sequence $\boldsymbol{x}$,
   FIND the sequence $\mathbf{y}$ of states that maximizes, e.g., P($\boldsymbol{y}$ | $\boldsymbol{x}$, $\boldsymbol{M}$),
   or the most probable subsequence of states
   ALGO. Viterbi, Forward-backward

3. Learning

   GIVEN an HMM $\boldsymbol{M}$, with unspecified transition/emission probs.,
   and a sequence $\boldsymbol{x}$,
   FIND parameters $\theta = (\pi_i, a_{ij}, \eta_{ik})$ that maximize P($\boldsymbol{x}$ | θ)
   ALGO. Baum-Welch (EM)

Joint Probability

$P(X\cancel{0} | M)$

X 1245526462146146136136661664661636616366163616515615115146123562344

Y FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF

- When the state-labeling is known, this is easy …

$$P(\mathbf{X}, \mathbf{Y}) \quad ?$$

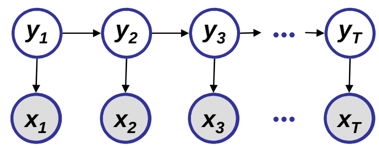
$$= P(x_1 x_2 \cdots x_T , y_1 \cdots y_T)$$

$$P(A, B, C) = P(A) P(B|A) P(C|B, A)$$

Probability of a Parse

$\prod_{node} p(node \mid parents)$

- Given a sequence $\mathbf{x} = x_1 \ldots\ldots x_T$ and a parse $\mathbf{y} = y_1, \ldots\ldots, y_T$,
- To find how likely is the parse: (given our HMM and the sequence)



$y_1 = 1$, 2 $\quad p(2|1)$ $\quad 0.5$

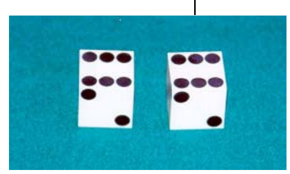
$$
\begin{aligned}
\underline{p(\mathbf{x}, \mathbf{y})} &= p(x_1 \ldots\ldots x_T, y_1, \ldots\ldots, y_T) \qquad \text{(Joint probability)} \\
&= p(y_1)\, p(x_1 \mid y_1)\, p(y_2 \mid y_1)\, p(x_2 \mid y_2) \ldots p(y_T \mid y_{T-1})\, p(x_T \mid y_T) \\
&= p(y_1)\, P(y_2 \mid y_1) \ldots p(y_T \mid y_{T-1}) \times p(x_1 \mid y_1)\, p(x_2 \mid y_2) \ldots p(x_T \mid y_T)
\end{aligned}
$$

- Marginal probability: $p(\mathbf{x}) = \sum_{\mathbf{y}} p(\mathbf{x}, \mathbf{y}) = \sum_{y_1} \sum_{y_2} \cdots \sum_{y_N} \pi_{y_1} \prod_{t=2}^{T} p(y_t \mid y_{t-1}) \prod_{t=1}^{T} p(x_t \mid y_t)$
- Posterior probability: $p(\mathbf{y} \mid \mathbf{x}) = p(\mathbf{x}, \mathbf{y}) / p(\mathbf{x})$

17

Example: the Dishonest Casino

- Let the sequence of rolls be:
  - x = 1, 2, 1, 5, 6, 2, 1, 6, 2, 4



- Then, what is the likelihood of
  - y = Fair, Fair, Fair, Fair, Fair, Fair, Fair, Fair, Fair, Fair?

    (say initial probs $a_{0Fair} = ½$, $a_{0Loaded} = ½$)

$P(x, y)$

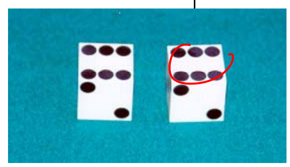
½ × P(1 | Fair) P(Fair | Fair) P(2 | Fair) P(Fair | Fair) … P(4 | Fair) =

1/6 $\quad$ .95 $\quad$ 1/6 $\quad$ 0.95

½ × (**1/6**)^{10} × (**0.95**)^{9} = .00000000521158647211 = 5.21 × 10^{-9}

18

Example: the Dishonest Casino



- So, the likelihood the die is fair in all this run is just 5.21×10^{-9}
- OK, but what is the likelihood of
  - π = Loaded, Loaded, Loaded, Loaded, Loaded, Loaded, Loaded, Loaded, Loaded, Loaded?

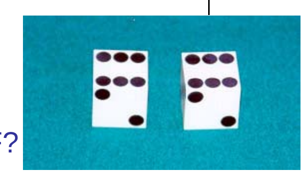
½ × P(1 | Loaded) P(Loaded | Loaded) … P(4 | Loaded) =

-5×10^{-9}

$\frac{1}{2} \times (1/10)^8 \times (1/2)^2 (0.95)^9 = .00000000078781176215 = 0.79 \times 10^{-9}$

- Therefore, it is after all 6.59 times more likely that the die is fair all the way, than that it is loaded all the way

Example: the Dishonest Casino



- Let the sequence of rolls be:
  - x = 1, 6, 6, 5, 6, 2, 6, 6, 3, 6
- Now, what is the likelihood π = F, F, …, F?
  - $\frac{1}{2} \times (1/6)^{10} \times (0.95)^9 = 0.5 \times 10^{-9}$, same as before
- What is the likelihood y = L, L, …, L?

$\frac{1}{2} \times (1/10)^4 \times (1/2)^6 (0.95)^9 = .00000049238235134735 = 5 \times 10^{-7}$

- So, it is 100 times more likely the die is loaded

Marginal Probability

$p(x, y)$

$p(x)$

1245526462146146136136661664661636616366163616515615115146123562344

FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF

- What if state-labeling Y is not observed

$$P(\mathbf{X}) \quad ?$$

$$= \sum_{y} p(x, y)$$

The Forward Algorithm

$\alpha_T^k = p(x_1 \cdots x_T, y_T)$

$p(x) = p(x_1 \cdots x_T)$

- We want to calculate $P(\mathbf{x})$, the likelihood of $\mathbf{x}$, given the HMM
  - Sum over all possible ways of generating $\mathbf{x}$:

$$p(\mathbf{x}) = \sum_{\mathbf{y}} p(\mathbf{x},\mathbf{y}) = \sum_{y_1}\sum_{y_2}\cdots\sum_{y_T} \pi_{y_1}\prod_{t=2}^{T} a_{y_{t-1},y_t}\prod_{t=1}^{T} p(x_t \mid y_t) \to O(M^T)$$

  - To avoid summing over an exponential number of paths $\mathbf{y}$, define $\to O(M^2)$

$$\alpha(y_t^k = 1) = \alpha_t^k \stackrel{\text{def}}{=} P(x_1,\ldots,x_t, y_t^k = 1) \quad \text{(the forward probability)}$$

  - The recursion:

$$\alpha_t^k = p(x_t \mid y_t^k = 1)\sum_i \alpha_{t-1}^i a_{i,k}$$

$$P(\mathbf{x}) = \sum_k \boxed{\alpha_T^k}$$

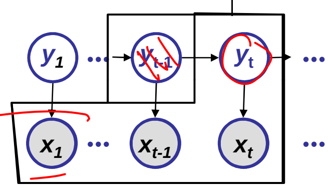
t $\quad t+1 \quad t+2$

α_t^k

$\begin{bmatrix} \alpha_t^1 \\ \alpha_t^2 \\ \vdots \\ \alpha_t^M \end{bmatrix} \quad \begin{bmatrix} \alpha_{t+1}^1 \\ \\ \\ \\ \end{bmatrix} \quad \begin{bmatrix} \\ \\ \\ \\ \end{bmatrix}$

The Forward Algorithm – derivation

- Compute the forward probability:



$$\alpha_t^k = P(x_1,\ldots,x_{t-1},x_t,y_t^k=1)$$

$$= \sum_{y_{t-1}} P(x_1 \cdots x_{t-1},\ y_{t-1}^k,\ y_t^k=1)$$

$$= \sum_{y_{t-1}} P(x_1,\ldots,x_{t-1},y_{t-1}) P(y_t^k=1 \mid y_{t-1},\cancel{x_1,\ldots,x_{t-1}}) P(x_t \mid y_t^k=1,\cancel{x_1,\ldots,x_{t-1},y_{t-1}})$$

$$= \sum_{y_{t-1}} \underbrace{P(x_1,\ldots,x_{t-1},y_{t-1})}_{\alpha_{t-1}} P(y_t^k=1 \mid y_{t-1}) P(x_t \mid y_t^k=1)$$

$$= P(x_t \mid y_t^k=1) \sum_i P(x_1,\ldots,x_{t-1},y_{t-1}^i=1) P(y_t^k=1 \mid y_{t-1}^i=1)$$

$$= P(x_t \mid y_t^k=1) \sum_i \alpha_{t-1}^i a_{i,k}$$

Chain rule: $P(A,B,C) = P(A)P(B \mid A)P(C \mid A,B)$

23

The Forward Algorithm

- We can compute α_t^k for all k, t, using dynamic programming!

Initialization:

$$\alpha_1^k = P(x_1 \mid y_1^k=1)\pi_k$$

$$\alpha_1^k = P(x_1, y_1^k=1)$$
$$= P(x_1 \mid y_1^k=1) P(y_1^k=1)$$
$$= P(x_1 \mid y_1^k=1)\pi_k$$

Iteration:

$$\alpha_t^k = P(x_t \mid y_t^k=1) \sum_i \alpha_{t-1}^i a_{i,k}$$

α_1 α_2

$$\begin{bmatrix}\alpha_1^1 \\ \alpha_1^2 \\ \vdots \\ \alpha_1^k\end{bmatrix} \quad \begin{bmatrix}\alpha_2^1 \\ \alpha_2^2 \\ \vdots\end{bmatrix}$$

Termination:

$$P(\mathbf{x}) = \sum_k \alpha_T^k$$

α_T

$$\begin{bmatrix}\alpha_T^1 \\ \alpha_T^2 \\ \vdots \\ \alpha_T^k\end{bmatrix}$$

$O(k^2 T)$

24

Three Main Questions on HMMs

1. Evaluation

GIVEN an HMM ***M***, and a sequence ***x***,
FIND Prob (***x*** | ***M***)
ALGO. Forward

2. Decoding

GIVEN an HMM ***M***, and a sequence ***x*** ,
FIND the sequence $\mathbf{y}$ of states that maximizes, e.g., P(***y*** | ***x*** , ***M***),
or the most probable subsequence of states
ALGO. Viterbi, Forward-backward

3. Learning

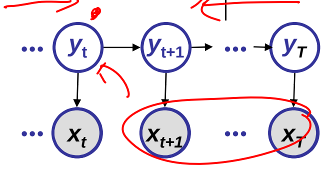
GIVEN an HMM ***M***, with unspecified transition/emission probs.,
and a sequence ***x***,
FIND parameters $\theta = (\pi_i, a_{ij}, \eta_{ik})$ that maximize P(***x*** | θ)
ALGO. Baum-Welch (EM)

The Backward Algorithm

$P(y_t \mid x) = \frac{P(y_t, x)}{P(x)}$

- We want to compute $P(y_t^k = 1 \mid \mathbf{x})$,

  the posterior probability distribution on the t^{th} position, given $\mathbf{x}$



- We start by computing

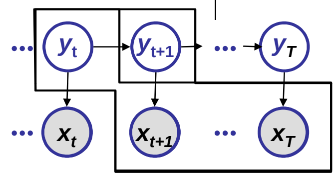
$$P(y_t^k = 1, \mathbf{x}) = P(x_1, \ldots, x_t, y_t^k = 1, x_{t+1}, \ldots, x_T)$$
$$= P(x_1, \ldots, x_t, y_t^k = 1) P(x_{t+1}, \ldots, x_T \mid \cancel{x_1, \ldots, x_t}, y_t^k = 1)$$
$$= \underbrace{P(x_1 \ldots x_t, y_t^k = 1)}_{\text{Forward, } \alpha_t^k} \underbrace{P(x_{t+1} \ldots x_T \mid y_t^k = 1)}_{\text{Backward, } \beta_t^k = P(x_{t+1}, \ldots, x_T \mid y_t^k = 1)}$$

- The recursion:

$$\beta_t^k = \sum_i a_{k,i} \, p(x_{t+1} \mid y_{t+1}^i = 1) \beta_{t+1}^i$$

The Backward Algorithm – derivation

- Define the backward probability:



$$\beta_t^k = P(x_{t+1},\ldots,x_T \mid y_t^k = 1)$$
$$= \sum_{y_{t+1}} P(x_{t+1},\ldots,x_T, y_{t+1} \mid y_t^k = 1)$$
$$= \sum_i P(y_{t+1}^i = 1 \mid y_t^k = 1) p(x_{t+1} \mid y_{t+1}^i = 1, y_t^k = 1) P(x_{t+2},\ldots,x_T \mid x_{t+1}, y_{t+1}^i = 1, y_t^k = 1)$$
$$= \sum_i P(y_{t+1}^i = 1 \mid y_t^k = 1) p(x_{t+1} \mid y_{t+1}^i = 1) P(x_{t+2},\ldots,x_T \mid y_{t+1}^i = 1)$$
$$= \sum_i a_{k,i} p(x_{t+1} \mid y_{t+1}^i = 1) \beta_{t+1}^i$$

Chain rule: $P(A, B, C \mid \alpha) = P(A \mid \alpha) P(B \mid A, \alpha) P(C \mid A, B, \alpha)$

The Backward Algorithm

- We can compute β_t^k for all k, t, using dynamic programming!

Initialization:

$$\beta_T^k = 1, \ \forall k$$

Iteration:

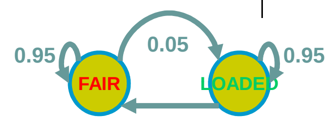
$$\beta_t^k = \sum_i a_{k,i} P(x_{t+1} \mid y_{t+1}^i = 1) \beta_{t+1}^i$$

Termination:

$$P(\mathbf{x}) = \sum_k \alpha_1^k \beta_1^k$$

Example:

x = 1, 2, 1, 5, 6, 2, 1, 6, 2, 4



P(1|F) = 1/6
P(2|F) = 1/6
P(3|F) = 1/6
P(4|F) = 1/6
P(5|F) = 1/6
P(6|F) = 1/6

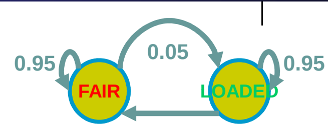
P(1|L) = 1/10
P(2|L) = 1/10
P(3|L) = 1/10
P(4|L) = 1/10
P(5|L) = 1/10
P(6|L) = 1/2

$$\alpha_t^k = P(x_t \mid y_t^k = 1)\sum_i \alpha_{t-1}^i a_{i,k}$$

$$\beta_t^k = \sum_i a_{k,i} P(x_{t+1} \mid y_{t+1}^i = 1)\beta_{t+1}^i$$

x = 1, 2, 1, 5, 6, 2, 1, 6, 2, 4

Alpha (actual)		Beta (actual)	
0.0833	0.0500	0.0000	0.0000
0.0136	0.0052	0.0000	0.0000
0.0022	0.0006	0.0000	0.0000
0.0004	0.0001	0.0000	0.0000
0.0001	0.0000	0.0001	0.0001
0.0000	0.0000	0.0007	0.0006
0.0000	0.0000	0.0045	0.0055
0.0000	0.0000	0.0264	0.0112
0.0000	0.0000	0.1633	0.1033
0.0000	0.0000	1.0000	1.0000



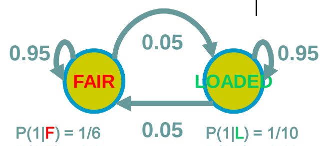
P(1|F) = 1/6
P(2|F) = 1/6
P(3|F) = 1/6
P(4|F) = 1/6
P(5|F) = 1/6
P(6|F) = 1/6

P(1|L) = 1/10
P(2|L) = 1/10
P(3|L) = 1/10
P(4|L) = 1/10
P(5|L) = 1/10
P(6|L) = 1/2

$$\alpha_t^k = P(x_t \mid y_t^k = 1)\sum_i \alpha_{t-1}^i a_{i,k}$$

$$\beta_t^k = \sum_i a_{k,i} P(x_{t+1} \mid y_{t+1}^i = 1)\beta_{t+1}^i$$

x = 1, 2, 1, 5, 6, 2, 1, 6, 2, 4



P(1|F) = 1/6
P(2|F) = 1/6
P(3|F) = 1/6
P(4|F) = 1/6
P(5|F) = 1/6
P(6|F) = 1/6

P(1|L) = 1/10
P(2|L) = 1/10
P(3|L) = 1/10
P(4|L) = 1/10
P(5|L) = 1/10
P(6|L) = 1/2

Alpha (logs)		Beta (logs)	
-2.4849	-2.9957	-16.2439	-17.2014
-4.2969	-5.2655	-14.4185	-14.9922
-6.1201	-7.4896	-12.6028	-12.7337
-7.9499	-9.6553	-10.8042	-10.4389
-9.7834	-10.1454	-9.0373	-9.7289
-11.5905	-12.4264	-7.2181	-7.4833
-13.4110	-14.6657	-5.4135	-5.1977
-15.2391	-15.2407	-3.6352	-4.4938
-17.0310	-17.5432	-1.8120	-2.2698
-18.8430	-19.8129	0	0

$$\alpha_t^k = P(x_t \mid y_t^k = 1)\sum_i \alpha_{t-1}^i a_{i,k}$$

$$\beta_t^k = \sum_i a_{k,i} P(x_{t+1} \mid y_{t+1}^i = 1)\beta_{t+1}^i$$

What is the probability of a hidden state prediction?

What is the probability of a hidden state prediction?

- A single state:

$$P(y_t|\mathbf{X})$$

- What about a hidden state sequence ?

$$P(y_1, \ldots, y_T|\mathbf{X})$$

Posterior decoding

- We can now calculate

$$P(y_t^k = 1 \mid \mathbf{x}) = \frac{P(y_t^k = 1, \mathbf{x})}{P(\mathbf{x})} = \frac{\alpha_t^k \beta_t^k}{P(\mathbf{x})}$$

- Then, we can ask
  - What is the most likely state at position *t* of sequence $\mathbf{x}$:

$$k_t^* = \arg\max_k P(y_t^k = 1 \mid \mathbf{x})$$

  - Note that this is an MPA of a single hidden state, what if we want to a MPA of a whole hidden state sequence?
  - Posterior Decoding: $\left\{ y_t^{k_t^*} = 1 : t = 1 \cdots T \right\}$
  - This is different from MPA of a whole sequence of hidden states
  - This can be understood as *bit error rate* vs. *word error rate*

Example:
MPA of *X* ?
MPA of *(X, Y)* ?

x	*y*	*P(x,y)*
0	*0*	*0.35*
0	*1*	*0.05*
1	*0*	*0.3*
1	*1*	*0.3*