Lecture 4: Probabilistic models 2

Helen Yannakoudakis and Oana Cocarascu

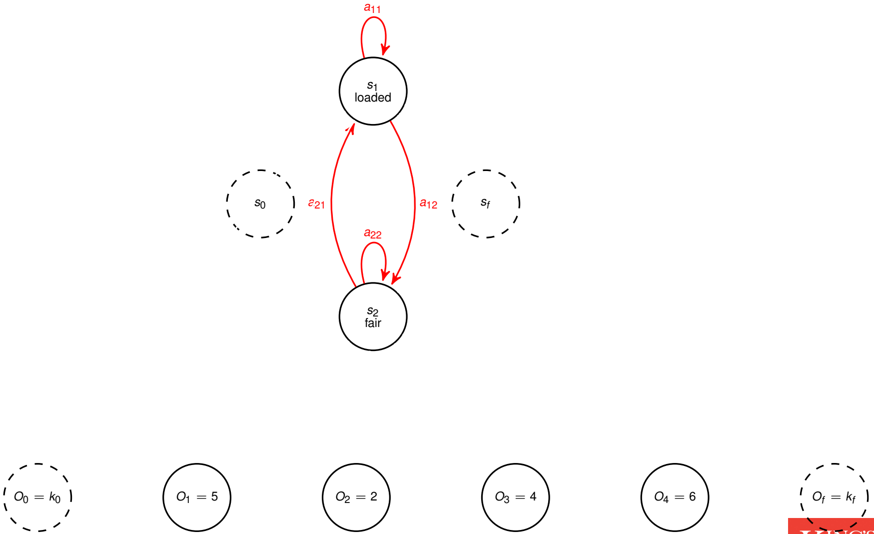
Department of Informatics
King's College London

February 7, 2022

The dice HMM

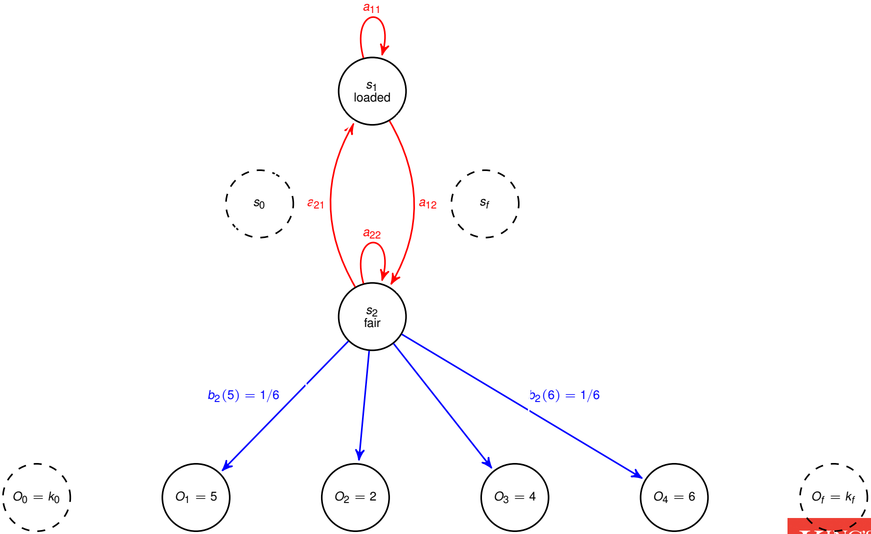
- Imagine a fraudulous croupier in a casino where customers bet on dice outcomes.
- She has two dice – a fair one and a loaded one.
- The fair one has the normal distribution of outcomes – $P(O) = \frac{1}{6}$ for each number 1 to 6.
- The loaded one has a different distribution.
- She secretly switches between the two dice.
- You don't know which dice is currently in use. You can only observe the numbers that are thrown.

The dice HMM



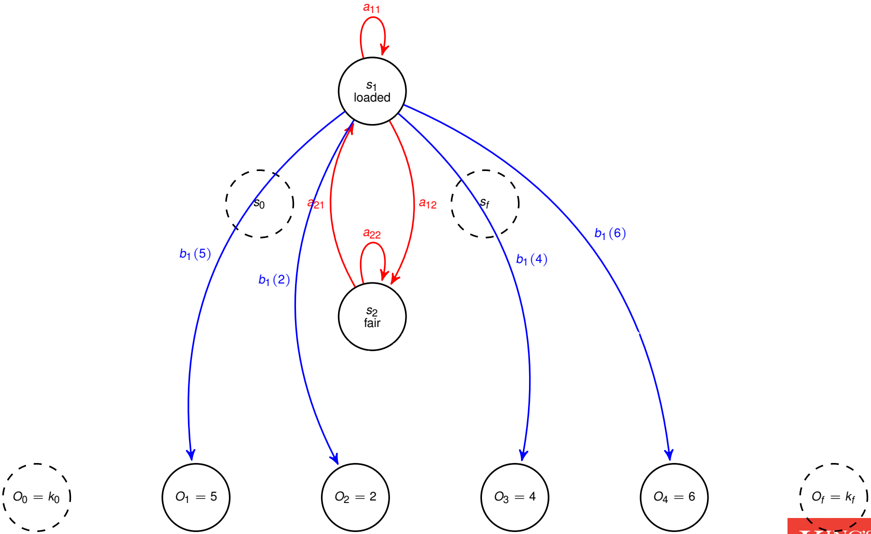
- There are two states (fair and loaded), and two special states (start s_0 and end s_f).
- Distribution of observations differs between the states.

The dice HMM



- There are two states (fair and loaded), and two special states (start s_0 and end s_f).
- Distribution of observations differs between the states.

The dice HMM



- There are two states (fair and loaded), and two special states (start s_0 and end s_f).
- Distribution of observations differs between the states.

The dice HMM

- Labelled HMM learning, i.e. it has
  - Input: dual tape of state and observation (dice outcome) sequences X and O.

	F	F	F	F	L	L	L	F	F	F	F	L	L	L	L	F	F	F	L	L	
	1	3	4	5	6	6	5	1	2	3	1	4	3	5	4	1	2	6	1	2	

  - Output: HMM parameters A, B.

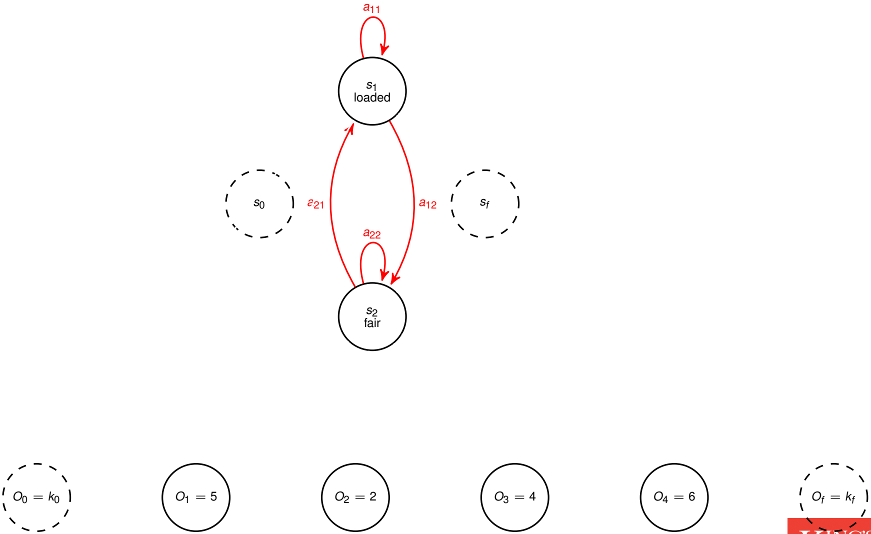
The dice HMM

- Labelled HMM learning, i.e. it has
  - Input: dual tape of state and observation (dice outcome) sequences X and O.

(s_0)	F	F	F	F	L	L	L	F	F	F	F	L	L	L	L	F	F	F	L	L	(s_f)
(k_0)	1	3	4	5	6	6	5	1	2	3	1	4	3	5	4	1	2	6	1	2	(k_f)

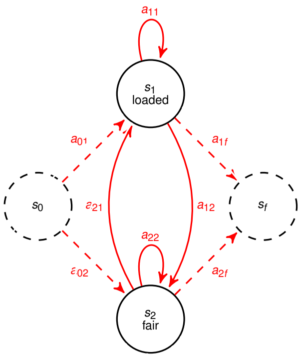
  - Output: HMM parameters A, B.

The dice HMM



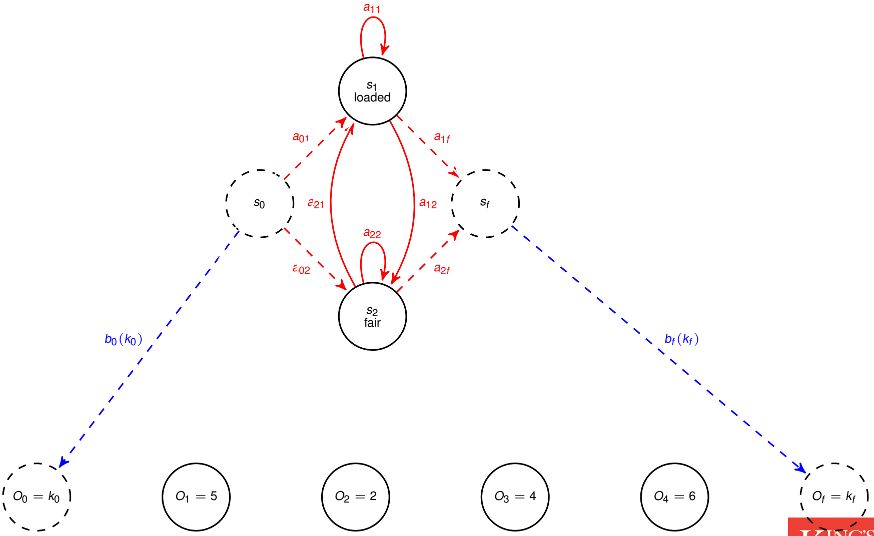
- There are two states (fair and loaded), and two special states (start s_0 and end s_f).
- Distribution of observations differs between the states.

The dice HMM



- There are two states (fair and loaded), and two special states (start s_0 and end s_f).
- Distribution of observations differs between the states.

The dice HMM



- There are two states (fair and loaded), and two special states (start s_0 and end s_f).
- Distribution of observations differs between the states.

Parameter estimation of HMM parameters A, B

(s_0)	F	F	F	F	L	L	L	F	F	F	F	L	L	L	L	F	F	F	L	L	(s_f)
(k_0)	1	3	4	5	6	6	5	1	2	3	1	4	3	5	4	1	2	6	1	2	(k_f)

- Transition matrix A consists of transition probabilities a_{ij}

$$a_{ij} = P(X_{t+1} = s_j | X_t = s_i) \sim \frac{count_{trans}(X_t = s_i, X_{t+1} = s_j)}{count_{trans}(X_t = s_i)}$$

- Emission matrix B consists of emission probabilities $b_i(k_j)$

$$b_i(k_j) = P(O_t = k_j | X_t = s_i) \sim \frac{count_{emission}(O_t = k_j, X_t = s_i)}{count_{emission}(X_t = s_i)}$$

Emission probabilities	k0	1	2	3	4	5	6	kf
F	0.00	0.36	0.18	0.18	0.09	0.09	0.09	0.00
L	0.00	0.11	0.11	0.11	0.22	0.22	0.22	0.00
s0	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
sf	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00

Transition probabilities	to F	to L	to s0	to sf
From s0	1.00	0.00	0.00	0.00
From F	0.73	0.27	0.00	0.00
From L	0.22	0.67	0.00	0.11
From sf	0.00	0.00	0.00	0.00

Parameter estimation of HMM parameters A, B

(s_0)	F	F	F	F	L	L	L	F	F	F	F	L	L	L	L	F	F	F	L	L	(s_f)
(k_0)	1	3	4	5	6	6	5	1	2	3	1	4	3	5	4	1	2	6	1	2	(k_f)

- Transition matrix A consists of transition probabilities a_{ij}

$$a_{ij} = P(X_{t+1} = s_j | X_t = s_i) \sim \frac{count_{trans}(X_t = F, X_{t+1} = L)}{count_{trans}(X_t = F)}$$

- Emission matrix B consists of emission probabilities $b_i(k_j)$

$$b_i(k_j) = P(O_t = k_j | X_t = s_i) \sim \frac{count_{emission}(O_t = k_j, X_t = s_i)}{count_{emission}(X_t = s_i)}$$

Emission probabilities	k0	1	2	3	4	5	6	kf
F	0.00	0.36	0.18	0.18	0.09	0.09	0.09	0.00
L	0.00	0.11	0.11	0.11	0.22	0.22	0.22	0.00
s0	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
sf	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00

Transition probabilities	to F	to L	to s0	to sf
From s0	1.00	0.00	0.00	0.00
From F	0.73	0.27	0.00	0.00
From L	0.22	0.67	0.00	0.11
From sf	0.00	0.00	0.00	0.00

Parameter estimation of HMM parameters A, B

(s_0)	F	F	F	F	L	L	L	F	F	F	F	L	L	L	L	F	F	F	L	L	(s_f)
(k_0)	1	3	4	5	6	6	5	1	2	3	1	4	3	5	4	1	2	6	1	2	(k_f)

- Transition matrix A consists of transition probabilities a_{ij}

$$a_{ij} = P(X_{t+1} = s_j | X_t = s_i) \sim \frac{count_{trans}(X_t = F, X_{t+1} = L)}{count_{trans}(X_t = F)}$$

- Emission matrix B consists of emission probabilities $b_i(k_j)$

$$b_i(k_j) = P(O_t = k_j | X_t = s_i) \sim \frac{count_{emission}(O_t = 6, X_t = L)}{count_{emission}(X_t = L)}$$

Emission probabilities	k0	1	2	3	4	5	6	kf
F	0.00	0.36	0.18	0.18	0.09	0.09	0.09	0.00
L	0.00	0.11	0.11	0.11	0.22	0.22	0.22	0.00
s0	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
sf	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00

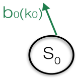
Transition probabilities	to F	to L	to s0	to sf
From s0	1.00	0.00	0.00	0.00
From F	0.73	0.27	0.00	0.00
From L	0.22	0.67	0.00	0.11
From sf	0.00	0.00	0.00	0.00

Fundamental tasks with HMMs

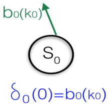
- Problem 1 (Labelled Learning)
  - Given a parallel observation and state sequence O and X, learn the HMM parameters A and B. → looked at this
- Problem 2 (Unlabelled Learning)
  - Given an observation sequence O (and only the set of emitting states S_e), learn the HMM parameters A and B.
- Problem 3 (Likelihood)
  - Given an HMM $\mu = (A, B)$ and an observation sequence O, determine the likelihood $P(O|\mu)$.
- Problem 4 (Decoding)
  - Given an observation sequence O and an HMM $\mu = (A, B)$, discover the best hidden state sequence X. → now to this

Viterbi algorithm, initialisation

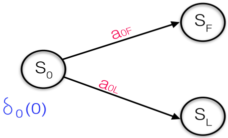
Viterbi algorithm, initialisation



Viterbi algorithm, initialisation

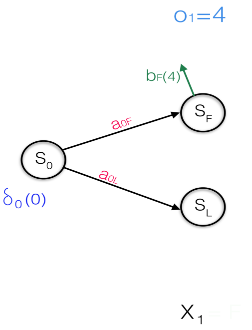


Viterbi algorithm, main step

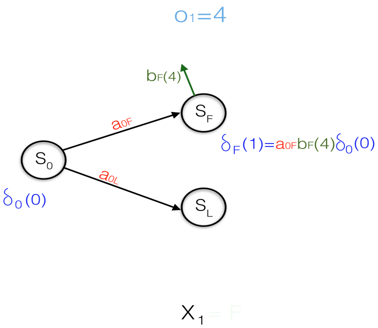


X_1

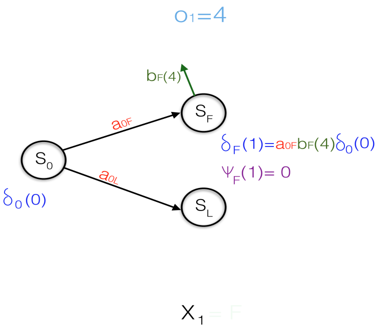
Viterbi algorithm, main step: observation is 4



Viterbi algorithm, main step: observation is 4

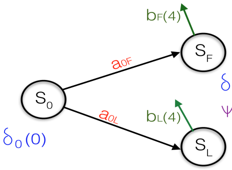


Viterbi algorithm, main step: observation is 4



Viterbi algorithm, main step: observation is 4

$o_1 = 4$

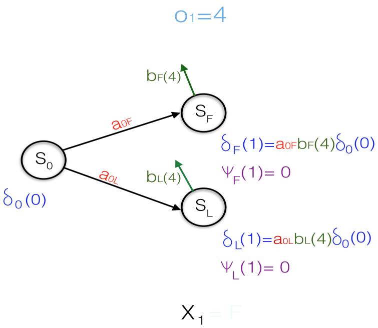


$\delta_F(1) = a_{0F} b_F(4) \delta_0(0)$

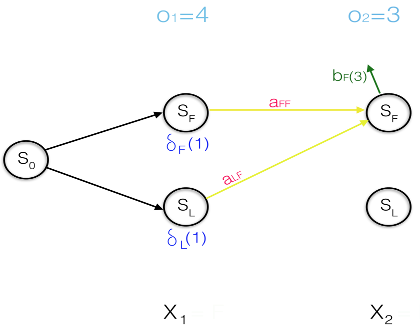
$\Psi_F(1) = 0$

X_1

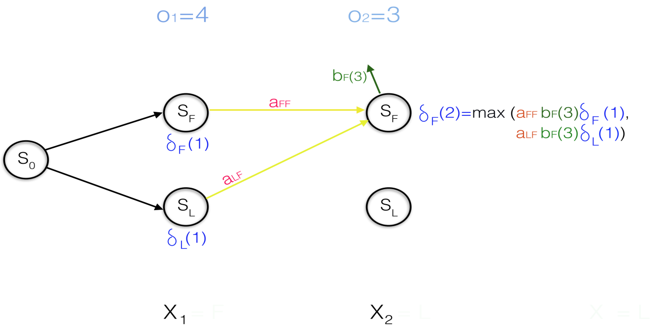
Viterbi algorithm, main step: observation is 4



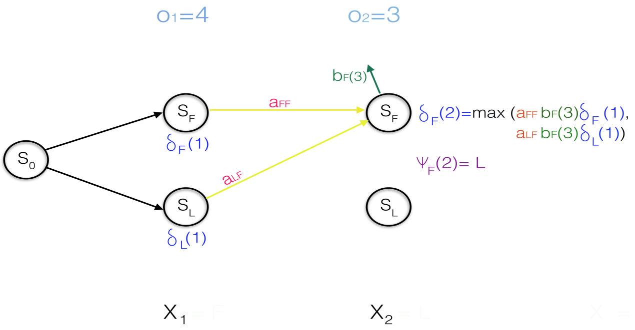
Viterbi algorithm, main step: observation is 3



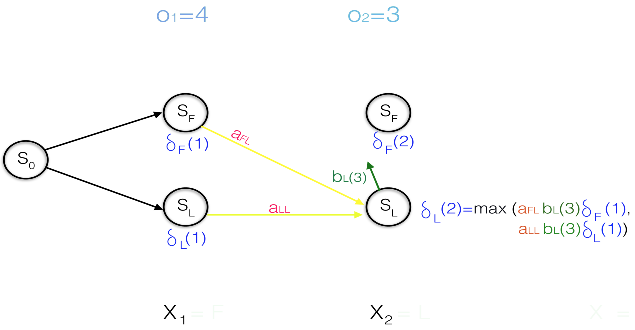
Viterbi algorithm, main step: observation is 3



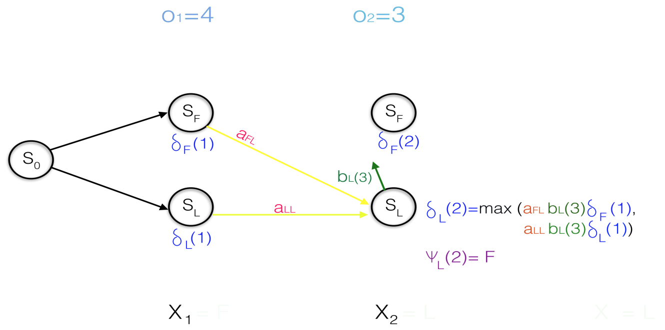
Viterbi algorithm, main step: observation is 3



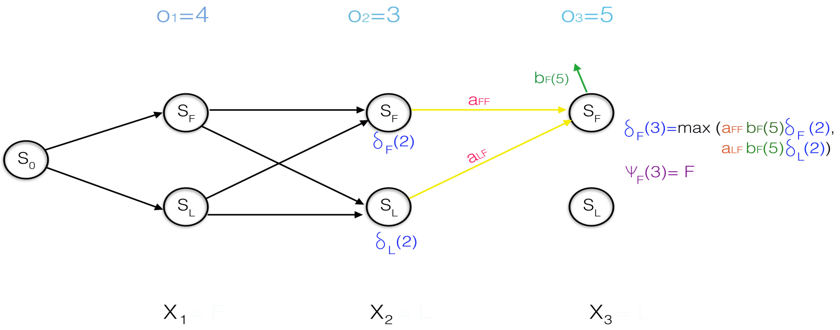
Viterbi algorithm, main step: observation is 3



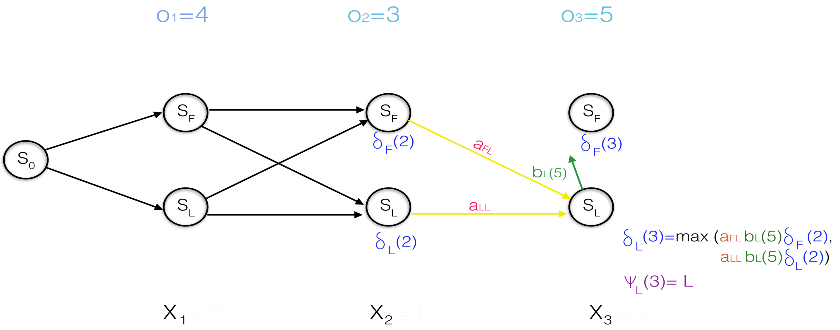
Viterbi algorithm, main step: observation is 3



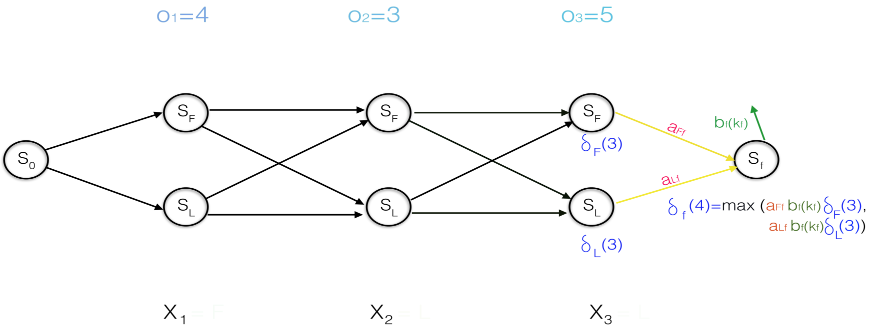
Viterbi algorithm, main step: observation is 5



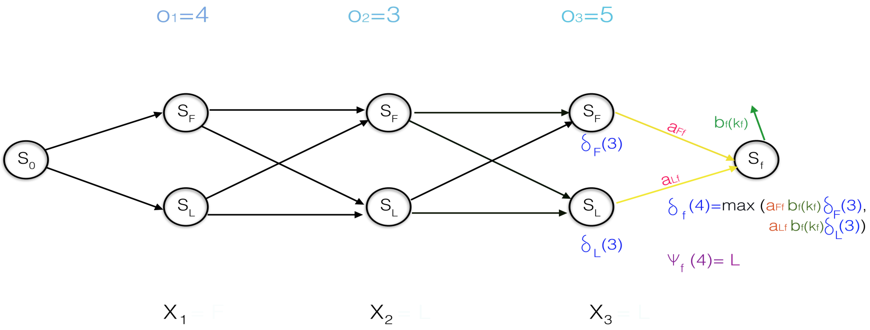
Viterbi algorithm, main step: observation is 5



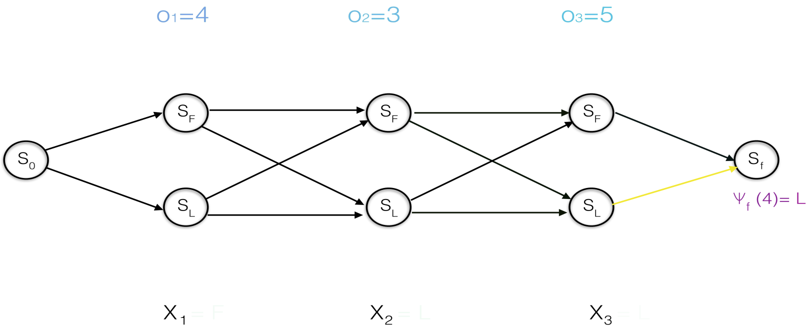
Viterbi algorithm, termination



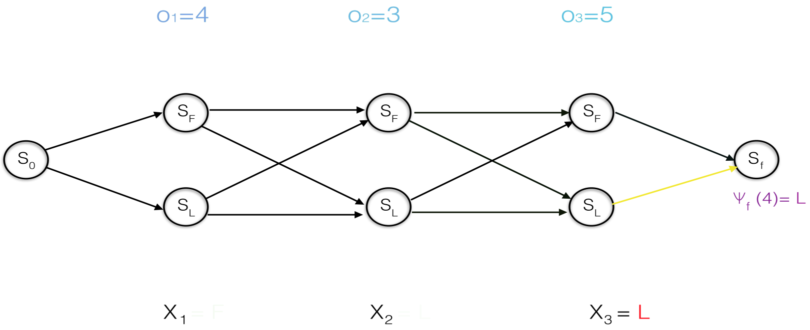
Viterbi algorithm, termination



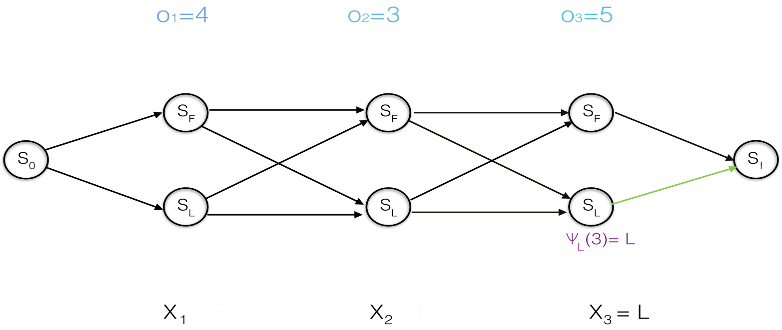
Viterbi algorithm, backtracing



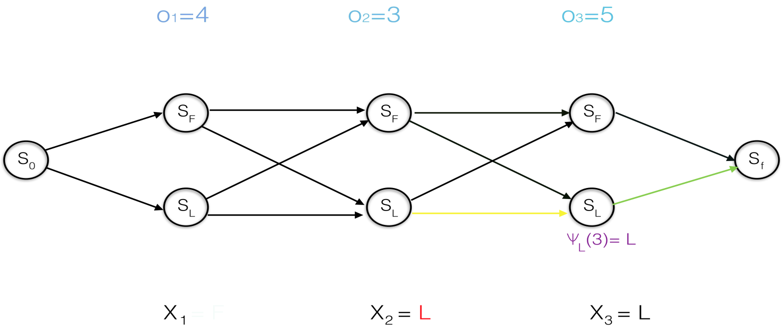
Viterbi algorithm, backtracing



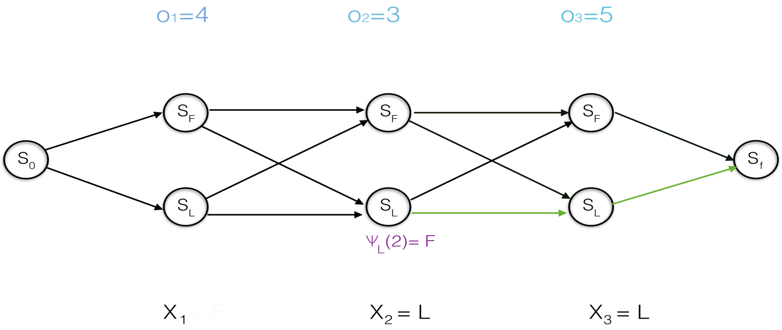
Viterbi algorithm, backtracing



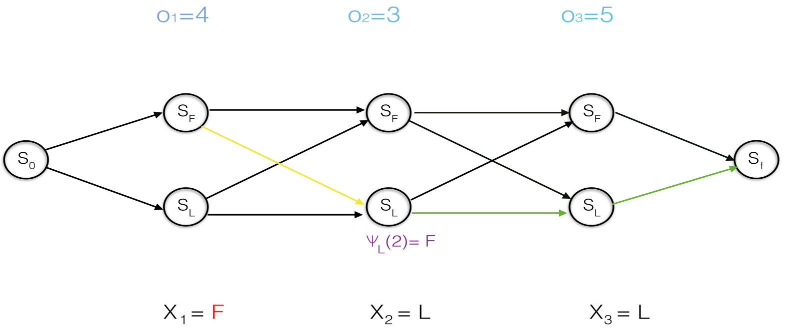
Viterbi algorithm, backtracing



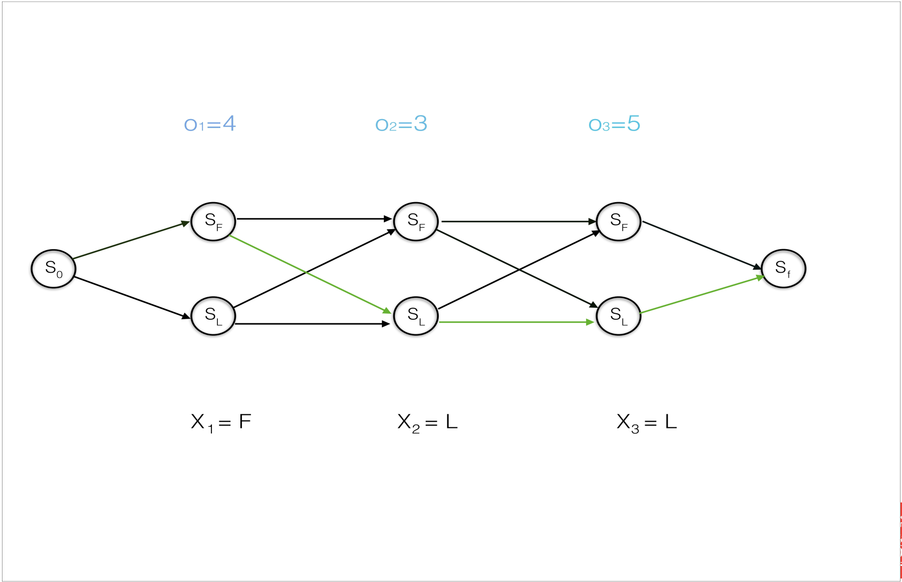
Viterbi algorithm, backtracing



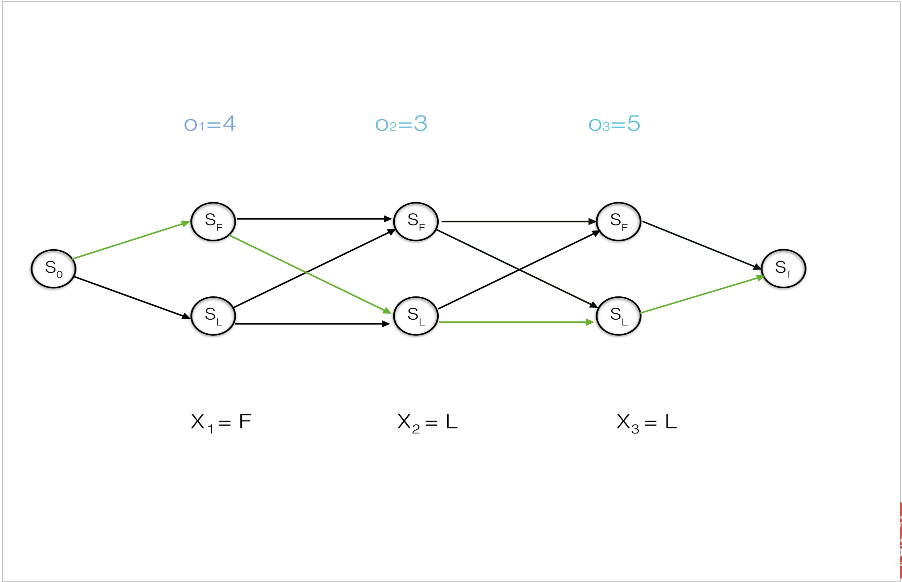
Viterbi algorithm, backtracing



Viterbi algorithm, backtracing



Viterbi algorithm, backtracing



Precision and Recall

- *How would we calculate the precision and recall for a the predicted sequence? Why is it useful to know the precision for the prediction of a sequence?*

Precision and Recall

- We can measure system success using accuracy.
- But sometimes it's only one type of instances that we find interesting.
- We don't want a summary measure that averages over interesting and non-interesting instances, as accuracy does.
- In those cases, we use precision, recall and F-measure.
- These metrics are imported from the field of information retrieval, where the difference beween interesting and non-interesting examples is particularly high.
- Accuracy doesn't work well when the types of instances are unbalanced.

KING'S College LONDON

Precision and Recall

Measure precision of L (P_L), recall of L (R_L) and F-measure of L (F_L).

System says:

Truth is:	F	L	Total
F	a	b	a+b
L	c	d	c+d
Total	a+c	b+d	a+b+c+d

- Precision of L: $P_L = \frac{d}{b+d} = \frac{\text{no of correctly predicted L}}{\text{number of predicted L}}$
- Recall of L: $R_L = \frac{d}{c+d} = \frac{\text{no of correctly predicted L}}{\text{true number of L}}$
- F-measure of L: $F_L = \frac{2P_L R_L}{P_L + R_L}$
- Accuracy: $A = \frac{a+d}{a+b+c+d}$

Significance testing

- *Would you mind explaining the 2-tailed test in detail? What is the meaning of ”direction” here? (Two-Tailed vs. One-Tailed Tests p73)*

Binomial Distribution $B(N, q)$

- Call the probability of a negative outcome q (here $q = 0.5$)
- Probability of observing $X = k$ negative events out of N:

$$P_q(X = k|N) = \binom{N}{k} q^k (1-q)^{N-k}$$

- At most k negative events:

$$P_q(X \leqslant k|N) = \sum_{i=0}^{k} \binom{N}{i} q^i (1-q)^{N-i}$$