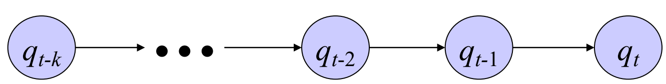
Hidden Markov Model (HMM)

1

Markov Model



1st-order Markov model

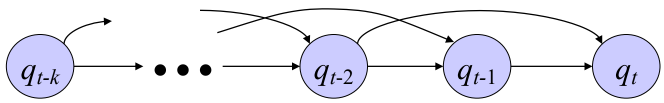
q_t represents the state at time t

$$P(q_t \mid q_{t-1}, q_{t-2}, \ldots) = P(q_t \mid q_{t-1})$$

$$P(q_t, q_{t-1}, q_{t-2}, \ldots) = P(q_t \mid q_{t-1})P(q_{t-1} \mid q_{t-2}) \ldots$$

2

Markov Model

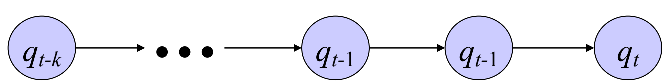


2nd-order Markov model

$P(q_t \mid q_{t-1}, q_{t-2}, \ldots) = P(q_t \mid q_{t-1}, q_{t-2})$

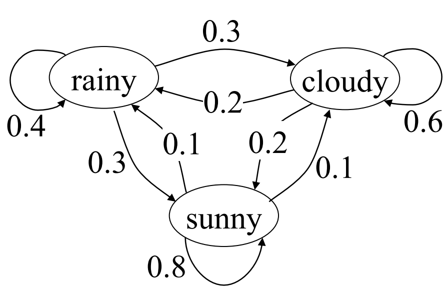
3

Markov Model



A 1st-order Markov model for weather predictor

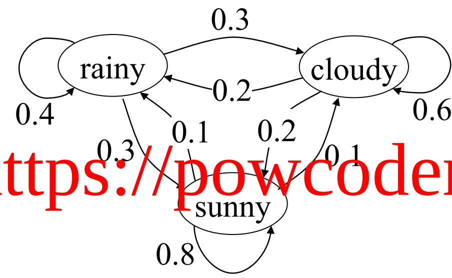
State-transition probability $P(q_t \mid q_{t-1})$



q_{t-1} \ q_t	rainy	cloudy	sunny
rainy	0.4	0.3	0.3
cloudy	0.2	0.6	0.2
sunny	0.1	0.1	0.8

4

Markov Model



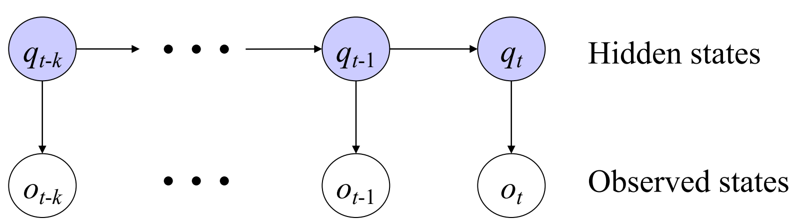
P(rainy → rainy → sunny → cloudy → sunny)?
↑
today

Assignment Project Exam Help

5

https://powcoder.com
Add WeChat powcoder

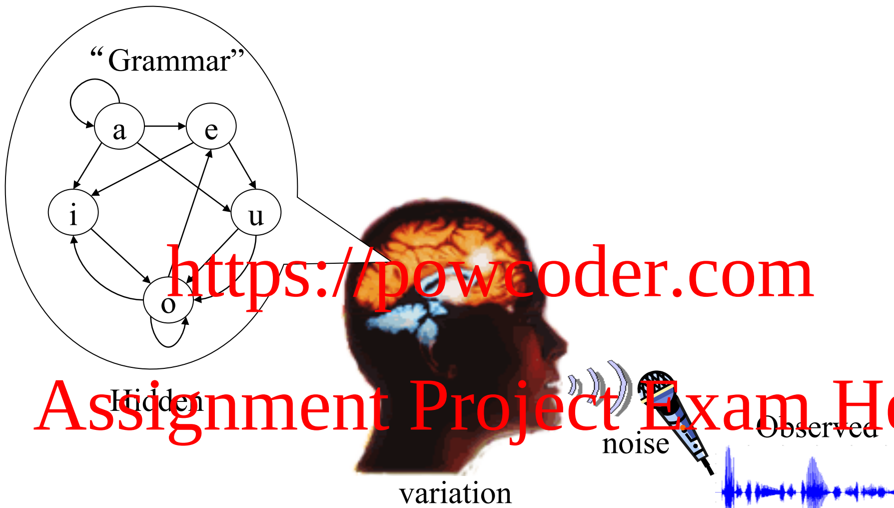
What is an HMM?



- Graphical Model
- Circles indicate states
- Arrows indicate probabilistic dependencies between states

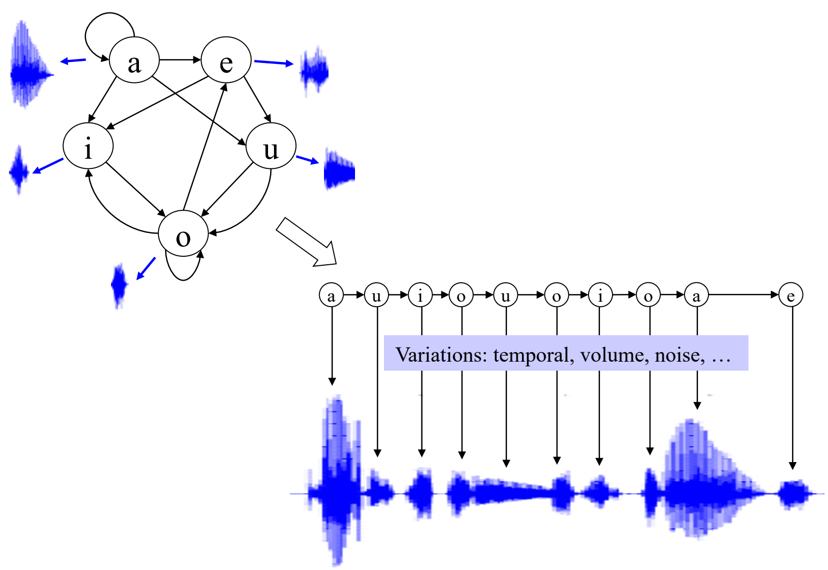
6

An Example



7

An Example



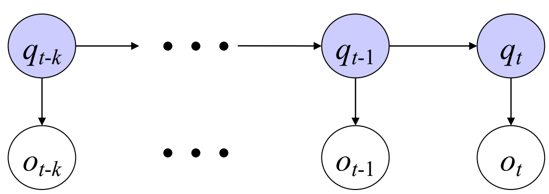
8

Applications of HMM

- Speech recognition
  - hidden: phoneme/words
  - observed: acoustic signals
- Handwriting recognition
  - hidden: words
  - observed: images
- Part-of-speech tagging
  - hidden: part-of-speech tags
  - observed: words
- Machine translation
  - hidden: words in the target language
  - observed: words in the source language
- … …

9

Formulate Discrete HMM



$\lambda = \{\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{\pi}\}$

q_t – The hidden state at time t

$\boldsymbol{S} = \{1, 2, \ldots, N\}$ – all possible values of hidden states.

o_t – The observation at time t

$\boldsymbol{V} = \{1, 2, \ldots, M\}$ – all possible values of observed states.

$\boldsymbol{A} = [a_{ij}]_{i,j}$ – state transition probabilities.

$a_{ij} = P(q_{t+1} = j \mid q_t = i), \quad a_{ij} \geq 0, \quad 1 \leq i, j \leq N. \ \Sigma_j a_{ij} = 1.$

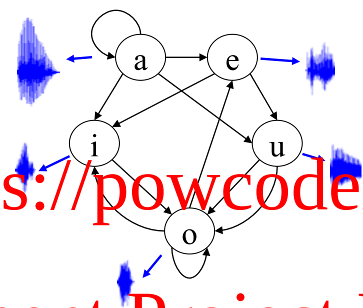
$\boldsymbol{B} = [b_{ik}]_{i,k}$ – emission probabilities.

$b_{ik} = P(o_t = k \mid q_t = i), \quad b_{ik} \geq 0, \quad 1 \leq i \leq N \ \& \ 1 \leq k \leq M. \ \Sigma_k b_{ik} = 1.$

$\boldsymbol{\pi} = \{\pi_i\}$ – initial state probabilities. $\pi_i = P(q_1 = i)$. $\Sigma_i \pi_i = 1$.

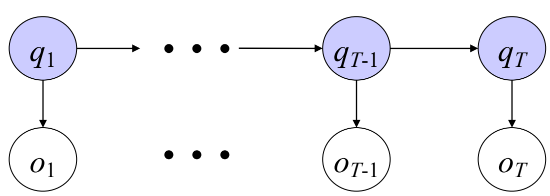
10

$\lambda = \{\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{\pi}\}$



11

Three Basic Problems for HMMs

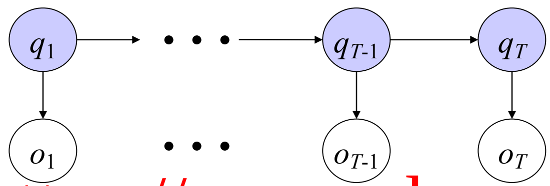


$\lambda = \{\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{\pi}\}$

1. Given the observation sequence $O = o_1 o_2 \ldots o_T$ and the model λ, how to efficiently compute $P(O;\lambda)$.

12

Three Basic Problems for HMMs

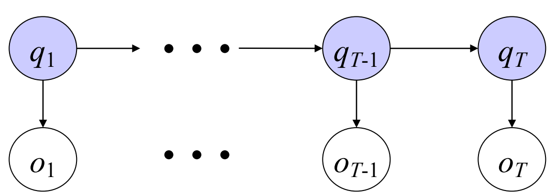


$\lambda = \{\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{\pi}\}$

2. Given the observation sequence $O = o_1 o_2 \ldots o_T$ and the model λ, how to compute the hidden state sequence $Q = q_1 q_2 \ldots q_T$ which best "explains" the observations (i.e., compute the most likely hidden state sequence)?

13

Three Basic Problems for HMMs

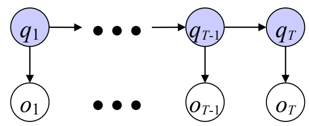


$\lambda = \{\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{\pi}\}$

3. How to adjust the model parameters $\lambda = \{\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{\pi}\}$ to maximize $P(O; \lambda)$. (adjust the model to fit the data best?)

14

Problem 1: Probability of an Observation Sequence



$\lambda = \{\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{\pi}\}$

Given $O = o_1 o_2 \ldots o_T$ and λ, compute $P(O;\lambda)$.

$$P(O;\lambda) = \sum_Q P(O,Q;\lambda) = \sum_Q P(O|Q;\lambda)P(Q;\lambda)$$

$$P(O|Q;\lambda) = \prod_{t=1}^{T} P(o_t|q_t;\lambda) \qquad P(Q;\lambda) = \pi_{q_1} a_{q_1 q_2} a_{q_2 q_3} \ldots a_{q_{T-1} q_T}$$

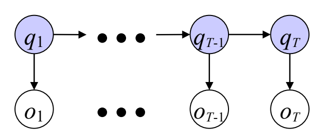
However, there are N^T possible hidden state sequences!

There is an efficient way – dynamic programming.

15

The Forward Procedure

$$P(o_1, \ldots, o_T; \lambda) = \sum_{q_T=1}^{N} P(o_1, \ldots, o_T, q_T; \lambda)$$



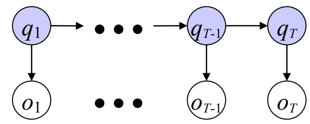
$\lambda = \{\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{\pi}\}$

$$\boxed{P(o_1, \ldots, o_T, q_T)} = P(o_1, \ldots, o_T|q_T)P(q_T)$$
$$= P(o_1, \ldots, o_{T-1}|q_T)P(o_T|q_T)P(q_T)$$
$$= P(o_1, \ldots, o_{T-1}, q_T)P(o_T|q_T)$$
$$= \sum_{q_{T-1}=1}^{N} P(o_1, \ldots, o_{T-1}, q_{T-1}, q_T)\, P(o_T|q_T)$$
$$= \sum_{q_{T-1}=1}^{N} P(o_1, \ldots, o_{T-1}, q_T|q_{T-1})\, P(q_{T-1})P(o_T|q_T)$$
$$= \sum_{q_{T-1}=1}^{N} P(o_1, \ldots, o_{T-1}\, |q_{T-1}, q_T)\, P(\, q_T|q_{T-1})P(q_{T-1})P(o_T|q_T)$$
$$= \sum_{q_{T-1}=1}^{N} P(o_1, \ldots, o_{T-1}\, |q_{T-1})\, P(\, q_T|q_{T-1})P(q_{T-1})P(o_T|q_T)$$
$$= \sum_{q_{T-1}=1}^{N} \boxed{P(o_1, \ldots, o_{T-1}\, , q_{T-1})} \underbrace{P(\, q_T|q_{T-1})}_{\text{Transition}} \underbrace{P(o_T|q_T)}_{\text{Emission}}$$

16

The Forward Procedure

$$P(o_1,\dots,o_t;\lambda) = \sum_{q_t=1}^{N} P(o_1,\dots,o_t,q_t;\lambda)$$



$\lambda = \{\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{\pi}\}$

$$P(\underline{o_1,\dots,o_t,q_t;\lambda}) = \sum_{q_{t-1}=1}^{N} P(\underline{o_1,\dots,o_{t-1},q_{t-1};\lambda})P(q_t|q_{t-1})P(o_t|q_t;\lambda)$$

Define the forward variable $\alpha_t(j) = P(o_1,\dots,o_t,q_t = j;\lambda)$

$$= b_{jo_t} \sum_{i=1}^{N} \alpha_{t-1}(i)a_{ij}$$

emission probability
$b_{jo_t} = P(o_t|q_t = j)$

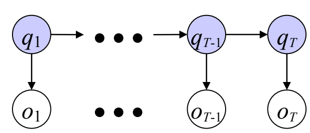
transition probability
$a_{ij} = P(q_t = j|q_{t-1} = i)$

$P(o_1,\dots,o_t;\lambda) = ?$

17

The Forward Procedure

$$P(o_1,\dots,o_t;\lambda) = \sum_{q_t=1}^{N} P(o_1,\dots,o_t,q_t;\lambda)$$



$\lambda = \{\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{\pi}\}$

$$\alpha_t(j) = b_{jo_t} \sum_{i=1}^{N} \alpha_{t-1}(i)a_{ij}$$

$$P(o_1,\dots,o_t;\lambda) = \sum_{j=1}^{N} \alpha_t(j)$$

$P(o_1,\dots,o_T;\lambda) =?$

(1) Initialize $\alpha_1(i) = \pi_i b_{io_1}$ $\quad 1 \le i \le N$

(2) Induction $(2 \le t \le T;\ 1 \le j \le N)$

$$\alpha_t(j) = b_{jo_t} \sum_{i=1}^{N} \alpha_{t-1}(i)a_{ij}$$

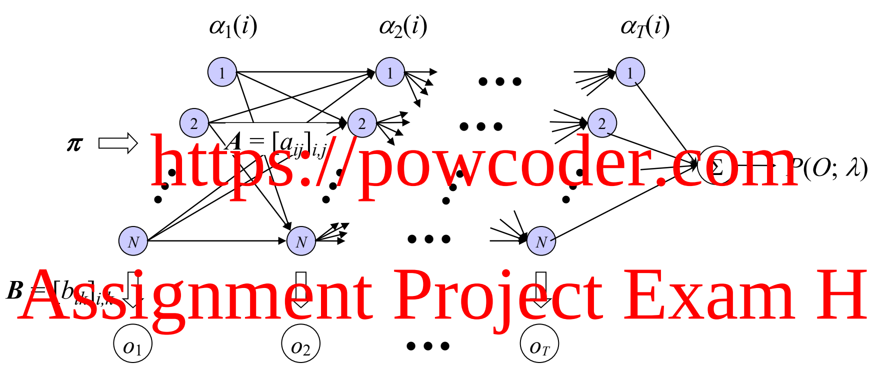
(3) Termination

$$P(O;\lambda) = P(o_1,\dots,o_T;\lambda) = \sum_{i=1}^{N} \alpha_T(i)$$

18

The Forward Procedure

$\lambda = \{\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{\pi}\}$ $\quad P(O;\lambda) = \sum_{i=1}^{N} \alpha_T(i)$ $\quad \alpha_t(j) = b_{jo_t} \sum_{i=1}^{N} \alpha_{t-1}(i) a_{ij}$

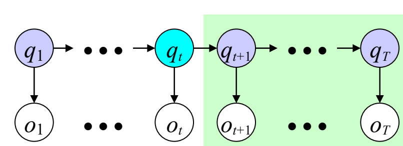


19

The Backward Procedure

Analogous to the forward variable, define a backward variable

$$\beta_t(i) = P(o_{t+1}, \ldots, o_T | q_t = i; \lambda)$$



(1) Initialize $\beta_T(i) = 1 \quad 1 \le i \le N$

(2) Induction $(T\text{-}1 \ge t \ge 1;\ 1 \le j \le N)$

$$\beta_t(i) = \sum_{j=1}^{N} a_{ij} b_{jo_{t+1}} \beta_{t+1}(j)$$

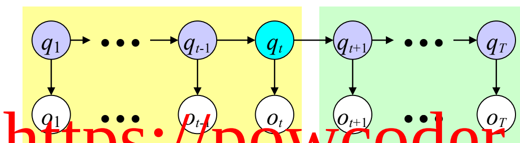
(3) Termination

$$P(O;\lambda) = P(o_1, \ldots, o_T; \lambda) = \sum_{i=1}^{N} \pi_i b_{io_1} \beta_1(i)$$

20

The Forward-Backward Procedure

$$P(O;\lambda) = \sum_{i=1}^{N} \alpha_T(i) \qquad\qquad P(O;\lambda) = \sum_{i=1}^{N} \pi_i \beta_1(i)$$

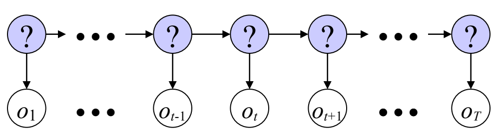


$$\alpha_t(i) = P(o_1, \dots, o_t, q_t = i;\lambda) \qquad \beta_t(i) = P(o_{t+1}, \dots, o_T | q_t = i;\lambda)$$

$$P(O;\lambda) = \sum_{i=1}^{N} \alpha_t(i)\beta_t(i)$$

21

Problem 2: Optimal State Sequence

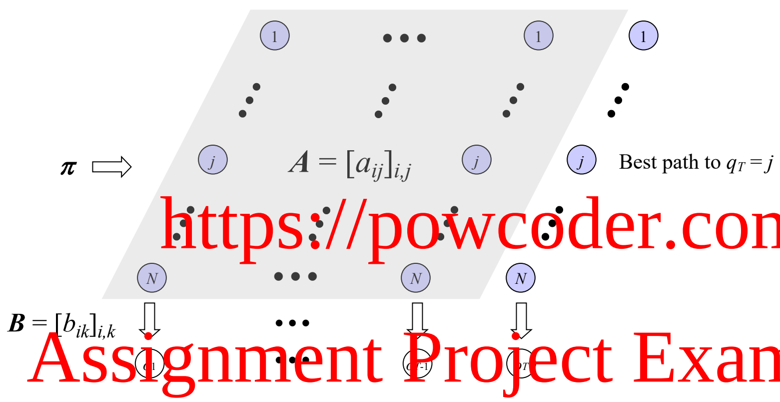


$$Q^* = \operatorname*{argmax}_Q P(Q|O;\lambda) = \operatorname*{argmax}_Q P(Q,O;\lambda)$$

$$O = o_1 o_2 \dots o_T \qquad Q = q_1 q_2 \dots q_T$$

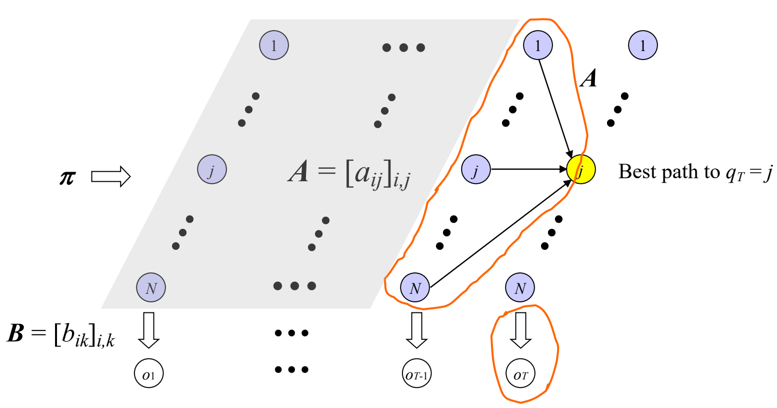
22

The Viterbi Algorithm



23

The Viterbi Algorithm



$$\max P(o_1, \dots, o_T, q_1, \dots, q_{T-1}, q_T = j; \lambda)$$

$$= \max_i \left[P(o_1, \dots, o_{T-1}, q_1, \dots, q_{T-1} = i; \lambda) a_{ij} \right] b_{j o_T}$$

24

The Viterbi Algorithm

$$\max P(o_1, \ldots, o_T, q_1, \ldots, q_{T-1}, q_T = j; \lambda)$$
$$= \max_i \left[P(o_1, \ldots, o_{T-1}, q_1, \ldots, q_{T-1} = i; \lambda) a_{ij}\right] b_{jo_T}$$

Define $\delta_{t+1}(j)$ – The highest probability of landing in state j at time t+1 after seeing the observations up to time t+1

$$\delta_{t+1}(j) = f(\delta_t(i)) = [\max \delta_t(i) a_{ij}] b_{jo_{t+1}}$$

Define the precedence $\psi_{t+1}(j) = \underset{i}{\operatorname{argmax}}\, \delta_t(i) a_{ij}$

25

The Viterbi Algorithm

1. Initialization

$$\delta_1(j) = \pi_j b_{jo_1} \quad \psi_1(j) = 0 \quad 1 \le j \le N$$

2. *for* (t = 1 *to* T – 1; j = 1 to N)

$$\delta_{t+1}(j) = [\max \delta_t(i) a_{ij}] b_{jo_{t+1}}$$
$$\psi_{t+1}(j) = \underset{i}{\operatorname{argmax}}\, \delta_t(i) a_{ij}$$

endfor

3. Termination $q_T^* = \underset{1 \le i \le N}{\operatorname{argmax}}\, \delta_T(i)$

4. State sequence backtracking $q_t^* = \psi_{t+1}(q_{t+1}^*)$

26

Consider $Q^* = \underset{Q}{\operatorname{argmax}} P(Q|O;\lambda)$

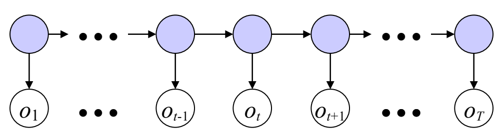
$q_1^* = \underset{q_1}{\operatorname{argmax}} P(q_1|O;\lambda)$

$q_1' = \underset{q_1}{\operatorname{argmax}} P(q_1|o_1;\lambda)$

Question: $Q^*(1) \stackrel{?}{=} q_1^* \stackrel{?}{=} q_1'$

27

Problem 3: Learning Parameters of HMMs



$$\lambda^* = \underset{\lambda}{\operatorname{argmax}} P(O;\lambda)$$

Easy if the hidden states are known.

The EM algorithm

28

The EM Algorithm

– The forward-backward (or Baum-Welch) algorithm

O – Observations $\{O_1, O_2, \dots O_D\}$

Q – The corresponding hidden states $\{Q_1, Q_2, \dots Q_D\}$

$$O_d = o_{d1} o_{d2} \dots o_{dT_d} \qquad Q_d = q_{d1} q_{d2} \dots q_{dT_d}$$

If $d \neq e$, T_d may not $= T_e$

Expectation step: $P(\mathbf{Q}|\mathbf{O}; \lambda^{(n)})$

Maximization step: $\lambda^{(n+1)} = \underset{\lambda}{\operatorname{argmax}}\, E\left[\log P(\mathbf{O}, \mathbf{Q}; \lambda) \,|\mathbf{O}; \lambda^{(n)}\right]$

29

Re-estimate Parameters

Expectation step: $P(\mathbf{Q}|\mathbf{O}; \lambda^{(n)})$

Maximization step:

$\pi_i^{(n+1)}$ = expected frequency in state i at $t = 1$.

$$a_{ij}^{(n+1)} = \frac{\text{expected number of transition from state } i \text{ to state } j}{\text{expected number of transition from state } i}$$

$$b_{ik}^{(n+1)} = \frac{\text{expected number of times in state } i \text{ and observing } k}{\text{expected number of times in } i}$$

30

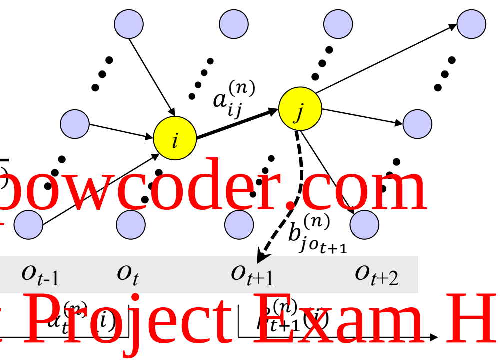
Transition Probabilities

Given the observations and the current parameters $\lambda^{(n)}$:

(a) The probability of traversing an arc: From state i at time t to state j at time t+1

$$\xi_t^{(n)}(i,j) = P(q_t = i, q_{t+1} = j \,|O; \lambda^{(n)})$$

$$= \frac{P(q_t = i, q_{t+1} = j \,, O; \lambda^{(n)})}{\sum_{r=1}^{N}\sum_{s=1}^{N} P(q_t = r, q_{t+1} = s, O; \lambda^{(n)})}$$



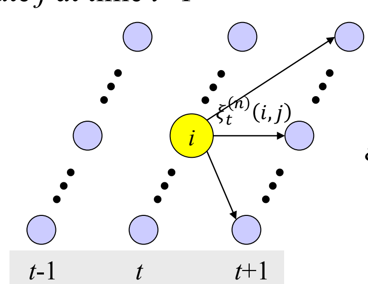
$$\xi_t^{(n)}(i,j) = \frac{\alpha_t^{(n)}(i) \times a_{ij}^{(n)} b_{j o_{t+1}}^{(n)} \times \beta_{t+1}^{(n)}(j)}{\sum_{r=1}^{N}\sum_{s=1}^{N} \alpha_t^{(n)}(r) \times a_{rs}^{(n)} b_{s o_{t+1}}^{(n)} \times \beta_{t+1}^{(n)}(s)}$$

31

Transition Probabilities

Given the observations and the current parameters $\lambda^{(n)}$:

(a) The probability of traversing an arc: From state i at time t to state j at time t+1



$$\xi_t^{(n)}(i,j) = P(q_t = i, q_{t+1} = j \,|O; \lambda^{(n)})$$

(b) The probability of being in state i at time t: $\gamma_t^{(n)}(i) = \sum_{j=1}^{N} \xi_t^{(n)}(i,j)$

32

Re-estimate Parameters

Given the observations and the current parameters $\lambda^{(n)}$:

(a) The probability of traversing an arc: $\xi_t^{(n)}(i,j)$

(b) The probability of being in state *i* at time *t*: $\gamma_t^{(n)}(i)$

- The probability that *i* is a start state π_i

$$\pi_i^{(n+1)} = \text{expected frequency in state } i \text{ at } (t=1) = \gamma_1^{(n)}(i)$$

33

Re-estimate Parameters

Given the observations and the current parameters $\lambda^{(n)}$:

(a) The probability of traversing an arc: $\xi_t^{(n)}(i,j)$

(b) The probability of being in state *i* at time *t*: $\gamma_t^{(n)}(i)$

- The probability that *i* is a start state π_i

$$\pi_i^{(n+1)} = \gamma_1^{(n)}(i)$$

- Transition Probabilities

$$a_{ij}^{(n+1)} = \frac{\text{expected number of transition from state } i \text{ to state } j}{\text{expected number of transition from state } i} = \frac{\sum_{t=1}^{T-1} \xi_t^{(n)}(i,j)}{\sum_{m=1}^{T-1} \gamma_m^{(n)}(i)}$$

34

Re-estimate Parameters

Given the observations and the current parameters $\lambda^{(n)}$:

(a) The probability of traversing an arc: $\xi_t^{(n)}(i,j)$

(b) The probability of being in state *i* at time *t*: $\gamma_t^{(n)}(i)$

- The probability that *i* is a start state π_i

$$\pi_i^{(n+1)} = \gamma_1^{(n)}(i)$$

- Transition Probabilities

$$a_{ij}^{(n+1)} = \frac{\sum_{t=1}^{T-1} \xi_t^{(n)}(i,j)}{\sum_{m=1}^{T-1} \gamma_m^{(n)}(i)}$$

- Emission Probabilities

$$b_{ik}^{(n+1)} = \frac{\text{expected number of times in state } i \text{ and observing } k}{\text{expected number of times in } i}$$

35

Re-estimate Parameters

Given the observations and the current parameters $\lambda^{(n)}$:

(a) The probability of traversing an arc: $\xi_t^{(n)}(i,j)$

(b) The probability of being in state *i* at time *t*: $\gamma_t^{(n)}(i)$

- The probability that *i* is a start state π_i

$$\pi_i^{(n+1)} = \gamma_1^{(n)}(i)$$

- Transition Probabilities

$$a_{ij}^{(n+1)} = \frac{\sum_{t=1}^{T-1} \xi_t^{(n)}(i,j)}{\sum_{m=1}^{T-1} \gamma_m^{(n)}(i)}$$

- Emission Probabilities

$$b_{ik}^{(n+1)} = \frac{\sum_{t=1}^{T-1} \gamma_t^{(n)}(i|o_t = k)}{\sum_{m=1}^{T-1} \gamma_m^{(n)}(i)}$$

36