11/2 $\boxed{\text{Review}}$ – Markov models

* Random variables $S_t = \{1, 2, \cdots, n\}$ state at time t.
* Belief network $S_1 \to S_2 \to \cdots \to S_{T-1} \to S_T$
* Assumptions
  - finite context $P(S_t | S_1, S_2, \cdots, S_{t-1}) = P(S_t | S_{t-1})$
  - shared CPTs $P(S_t = s' | S_{t-1} = s) = P(S_{t+1} = s' | S_t = s)$
* Weaknesses
  - modeling k^{th} order correlations requires CPTs with $O(n^k)$ elements.
  - assumes that the true state of the world can be observed.

$\boxed{\text{Hidden Markov models (HMMs)}}$

* Random variables

  $S_t \in \{1, 2, \cdots, n\}$ state at time t.

  $O_t \in \{1, 2, \cdots, m\}$ observation at time t.

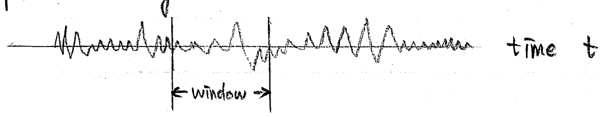
  Observations O_t are noisy, partial reflections of states S_t.

  Ex: toilet training

  $S = \{$ have-to-go, don't-need-to-go, went $\}$

  $O = \{$ neutral, funny walk, intense concentration, squat $\}$

  Ex: speech recognition



  O_t : acoustic measurements on windowed waveform at time t

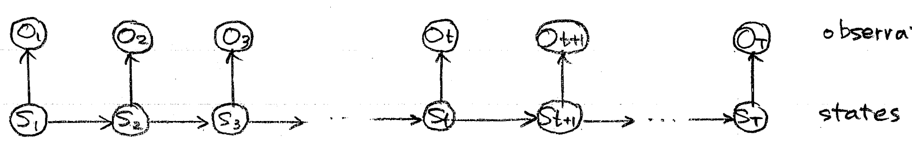
  S_t : unit of language (word, syllable, phoneme)

  Ex: robotics

  O_t : sensor readings

  S_t : location, orientation.

* Belief network



: This is a polytree!

* Assumptions

  finite context $P(S_t | S_1, S_2, \cdots S_{t-1}) = P(S_t | S_{t-1})$

  $P(O_t | S_1, S_2, \cdots S_T) = P(O_t | S_t)$

– shared CPTs

$P(S_{t+1}=s' \mid S_t = s) = P(S_t = s' \mid S_{t-1} = s)$

$P(O_t = o \mid S_t = s) = P(O_{t+1} = o \mid S_{t+1} = s)$

* Joint distribution

$$P(\underbrace{S_1, S_2, \cdots, S_T}_{\vec{S}}, \underbrace{O_1, O_2, \cdots, O_T}_{\vec{O}}) = \underbrace{P(S_1)}_{\text{initial state}} \left[\prod_{t=2}^{T} P(S_t \mid S_{t-1})\right]\left[\prod_{t=1}^{T} P(O_t \mid S_t)\right].$$

* Parameters

$\pi_i = P(S_1 = i)$ initial ~~state~~ distribution.

$a_{ij} = P(S_{t+1} = j \mid S_t = i)$ transition matrix

$b_{ik} = P(O_t = k \mid S_t = i)$ emission matrix

For clarity: $b_{ik} = b_i(k)$.

* Key computations / questions in HMMs.

inference (assume parameters given)

1) How to compute likelihood $P(O_1, O_2, \cdots, O_T)$?

Ex: isolated word recognition.

2) How to compute most likely (hidden) state sequence

$$\underset{\vec{S}}{\arg\max}\, P(S_1, S_2, \cdots, S_T \mid O_1, O_2, \cdots, O_T)$$

Ex: continuous speech recognition.

learning

3) How to estimate parameters $\{\pi_i, a_{ij}, b_{ik}\}$ that maximize $P(O_1, O_2, \cdots, O_T)$?

[or maybe multiple observation sequences].

(1) Computing likelihood

$$P(O_1, O_2, \cdots, O_T) = \sum_{\vec{S}} P(S_1, S_2, \cdots, S_T, O_1, \cdots, O_T) \quad \text{marginalization}$$

$$= \sum_{\vec{S}} P(S_1) \prod_{t=2}^{T} P(S_t \mid S_{t-1}) \prod_{t=1}^{T} P(O_t \mid S_t).$$

* Efficient recursion.

$$P(O_1, O_2, \cdots, O_t, O_{t+1}, S_{t+1} = j)$$

$$= \sum_{i=1}^{n} P(O_1, O_2, \cdots, O_{t+1}, S_{t+1} = j, S_t = i) \quad \text{marginalization}$$

$$= \sum_{i=1}^{n} P(O_1, O_2, \cdots, O_t, S_t = i)\, P(S_{t+1} = j \mid S_t = i, \underbrace{O_1, \cdots, O_t}_{\text{conditional independence}})\, P(O_{t+1} \mid S_{t+1} = j, \underbrace{S_t = i, O_1, \cdots, O_t}_{\text{conditional independence}})$$

product rule

$$= \sum_{i=1}^{n} \underbrace{P(O_1, O_2, \cdots, O_t, S_t = i)}_{\text{recursion}}\, \underbrace{P(S_{t+1} = j \mid S_t = i)\, P(O_{t+1} \mid S_{t+1} = j)}_{\text{CPTs of HMM}}$$

* Shorthand notation.

$\alpha_{it} \triangleq P(O_1, O_2, \cdots, O_t, S_t = i)$ [$n \times T$ matrix]

$\alpha_{j\,t+1} = \sum_{i=1}^{n} \alpha_{it}\, a_{ij}\, \underbrace{b_j(O_{t+1})}_{b_{j O_{t+1}}}$ "forward algorithm"

sum last column to get likelihood.

base case : $\alpha_{i1} = P(O_1, S_1 = i) = P(S_1 = i)\, P(O_1 | S_1 = i) = \pi_i\, b_i(O_1)$

* likelihood

$$P(O_1, \cdots, O_T) = \sum_{i=1}^{n} P(O_1, O_2, \cdots, O_T, S_T = i) \quad \text{marginalization.}$$
$$= \sum_{i=1}^{n} \alpha_{iT}$$

* Warning : for long sequence, watch out for underflow.

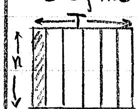
(2) Computing most likely state sequence

$$S^* = \{S_1^*, S_2^*, \cdots, S_T^*\}$$
$$= \underset{S}{\operatorname{argmax}}\ P(S_1, S_2, \cdots, S_T \mid O_1, O_2, \cdots, O_T)$$

constant with respect to hidden states.

$$= \underset{S}{\operatorname{argmax}} \left[P(S_1, S_2, \cdots, S_T, O_1, O_2, \cdots, O_T) \,/\, P(O_1, O_2, \cdots, O_T) \right]$$
$$= \underset{S}{\operatorname{argmax}}\ P(S_1, S_2, \cdots, S_T, O_1, O_2, \cdots, O_T)$$

Define $l_{it}^* = \max_{\{S_1, S_2, \cdots, S_{t-1}\}} \log P(O_1, O_2, \cdots, O_t, S_1, S_2, \cdots, S_{t-1}, S_t = i)$



= log probability of most likely t-step "path" that ends in state i at time t for observations $\{O_1, O_2, \cdots, O_t\}$

* Form recursion

(i) base case ($t = 1$)

$$l_{i1}^* = \log P(S_1 = i, O_1) = \log [P(S_1 = i)\, P(O_1 | S_1 = i)]$$
$$= \log \pi_i + \log b_i(O_1).$$

(ii) from time t to time $t+1$:

product rule + conditional independence

$$l_{j\,t+1}^* = \max_{S_1, \cdots, S_t} \log P(S_1, S_2, \cdots, S_t, S_{t+1} = j, O_1, O_2, \cdots, O_{t+1})$$
$$= \max_{S_1, \cdots, S_{t-1}} \max_{i} \left[\log P(S_1, \cdots, S_{t-1}, S_t = i, O_1, \cdots, O_t)\, P(S_{t+1} = j | S_t = i) \times P(O_{t+1} | S_{t+1} = j) \right]$$

i representing state S_t

$$= \max_{i} \left[\left\{ \max_{S_1, \cdots, S_{t-1}} \log P(S_1, \cdots, S_{t-1}, S_t = i, O_1, \cdots, O_t) \right\} + \log P(S_{t+1} = j \mid S_t = i) \right] + \log P(O_{t+1} \mid S_{t+1} = j)$$
$$= \max_{i} \left[l_{it}^* + \log a_{ij} \right] + \log b_j(O_{t+1}).$$

* How to derive S^* from l^* ?