6.867 Machine learning and neural networks

Tommi Jaakkola
MIT AI Lab
tommi@ai.mit.edu

Lecture 17: HMM estimation/inference

Topics

- Hidden markov models
  - forward-backward algorithm
  - estimation via EM

Review: hidden Markov models

- A hidden Markov model (HMM) is model where we generate a sequence of outputs in addition to the Markov state sequence

$$\begin{array}{ccccccc} s_0 & \rightarrow & s_1 & \rightarrow & s_2 & \rightarrow & \ldots \\ \downarrow & & \downarrow & & \downarrow & & \\ O_0 & & O_1 & & O_2 & & \end{array}$$

  A HMM is defined by

  1. number of states m
  2. initial state distribution $P_0(s_0)$
  3. state transition model $P_1(s_{t+1}|s_t)$
  4. output model $P_o(O_t|s_t)$ (discrete or continuous)

- This is a *latent variable model* in the sense that we will only observe the outputs $\{O_0, O_1, \ldots, O_n\}$; the state sequence remains “hidden”

HMM problems

- There are several problems we have to solve
  1. How do we evaluate the probability that our model generated the observation sequence $\{O_0, O_1, \ldots, O_n\}$?
     - forward-backward algorithm
  2. How do we uncover the most likely hidden state sequence corresponding to these observations?
     - dynamic programming
  3. How do we adapt the parameters of the HMM to better account for the observations?
     - the EM-algorithm

Probability of observed data

- In principle computing the probability of the observed sequence involves summing over exponentially many possible hidden state sequences

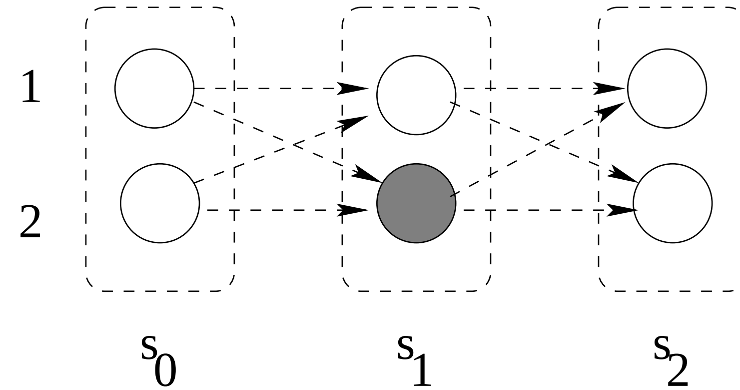
$$P(O_0, \ldots, O_n) = \sum_{s_0, \ldots, s_n} \overbrace{P_0(s_0)P_1(O_0|s_0)\ldots P_1(s_n|s_{n-1})P_o(O_n|s_n)}^{\text{Prob. given a specific hidden state sequence}}$$

$$\begin{array}{ccccccccc} s_0 & \rightarrow & s_1 & \rightarrow & s_2 & \rightarrow \ldots \rightarrow & s_{n-1} & \rightarrow & s_n \\ \downarrow & & \downarrow & & \downarrow & & \downarrow & & \downarrow \\ O_0 & & O_1 & & O_2 & & O_{n-1} & & O_n \end{array}$$

- We can, however, exploit the structure of the model to evaluate the probability much more efficiently

Forward-backward algorithm

$$\begin{tikzcd} s_0 \arrow[r] \arrow[d] & s_1 \arrow[r] \arrow[d] & s_2 \arrow[d] \\ O_0 & O_1 & O_2 \end{tikzcd}$$



$O_0 = heads,\ O_1 = tails,\ O_2 = heads$

- Forward probabilities $\alpha_t(i)$:

$$\begin{aligned} \alpha_t(i) &= P(O_0, \ldots, O_t, s_t = i) \\ \frac{\alpha_t(i)}{\sum_j \alpha_t(j)} &= P(s_t = i | O_0, \ldots, O_t) \end{aligned}$$

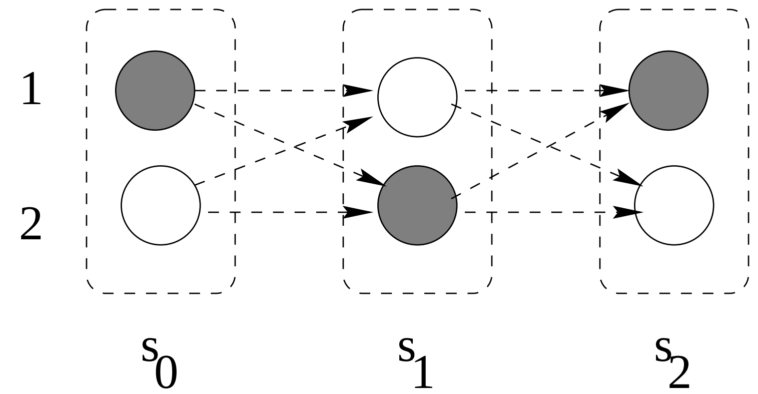
  (tracking etc; discrete state Kalman filter)

- Backward propabilities $\beta_t(i)$:

$$\beta_t(i) = P(O_{t+1}, \ldots, O_n | s_t = i)$$

  (evidence about the current state from future observations)

Recursive forward updates



$O_0 = heads$, $O_1 = tails$, $O_2 = heads$

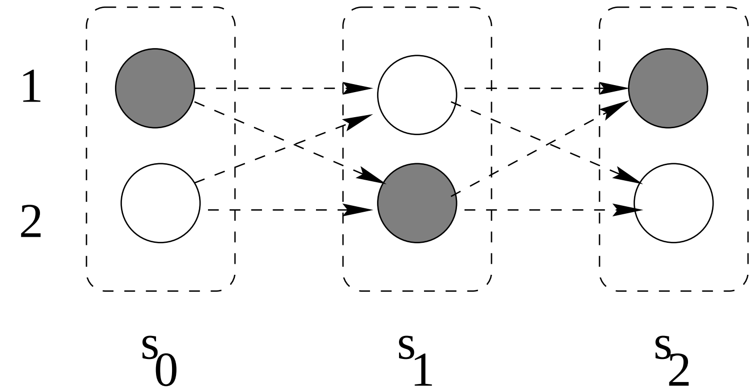
- Forward recursion: $\alpha_t(i) = P(O_0, \ldots, O_t, s_t = i)$

$$\begin{aligned}
\alpha_0(1) &= P_0(1)\, P_o(heads|1) \\
\alpha_0(2) &= P_0(2)\, P_o(heads|2) \\
\alpha_1(1) &= \Big[\alpha_0(1) P_1(1|1) + \alpha_0(2) P_1(1|2)\Big]\, P_o(tails|1) \\
\alpha_1(2) &= \Big[\alpha_0(1) P_1(2|1) + \alpha_0(2) P_1(2|2)\Big]\, P_o(tails|2)
\end{aligned}$$

- More generally:

$$\begin{aligned}
\alpha_0(i) &= P_0(s_0 = i)\, P_o(O_0|s_0 = i) \\
\alpha_t(i) &= \Big[\sum_j \alpha_{t-1}(j)\, P_1(s_t = i|s_{t-1} = j)\Big]\, P_o(O_t|s_t = i)
\end{aligned}$$

Recursive backward updates



$O_0 = heads,\ O_1 = tails,\ O_2 = heads$

- Backward recursion: $\beta_t(i) = P(O_{t+1}, \ldots, O_n | s_t = i)$

$$\begin{aligned}
\beta_2(1) &= 1 \\
\beta_2(2) &= 1 \\
\beta_1(1) &= P_1(1|1)P_o(heads|1)\beta_2(1) + P_1(2|1)P_o(heads|2)\beta_2(2) \\
\beta_1(2) &= P_1(1|2)P_o(heads|1)\beta_2(1) + P_1(2|2)P_o(heads|2)\beta_2(2)
\end{aligned}$$

- More generally:

$$\begin{aligned}
\beta_n(i) &= 1 \\
\beta_{t-1}(i) &= \sum_j P_1(s_t = j | s_{t-1} = i) P_o(O_t | s_t = j)\, \beta_t(j)
\end{aligned}$$

Forward/backward probabilities

$$\begin{array}{ccccccccc}
s_0 & \longrightarrow & s_1 & \longrightarrow & s_2 & \longrightarrow \ldots \longrightarrow & s_{n-1} & \longrightarrow & s_n \\
\downarrow & & \downarrow & & \downarrow & & \downarrow & & \downarrow \\
O_0 & & O_1 & & O_2 & & O_{n-1} & & O_n
\end{array}$$

- The forward/backward probabilities

$$\begin{aligned}
\alpha_t(i) &= P(O_0, \ldots, O_t, s_t = i) \\
\beta_t(i) &= P(O_{t+1}, \ldots, O_n | s_t = i)
\end{aligned}$$

permit us to evaluate various posterior probabilities

For example, the probability of generating the observations and going through state i at time t is

$$P(O_0, \ldots, O_n, s_t = i) = \alpha_t(i)\beta_t(i)$$

Summing over the possible states at time t gives back

$$P(O_0, \ldots, O_n) = \sum_j \alpha_t(j)\beta_t(j), \ \text{ for any } t = 0, \ldots, n$$

Forward/backward probabilities cont'd

$$\begin{array}{ccccccccccc} s_0 & \rightarrow \ldots \rightarrow & s_t & \rightarrow & s_{t+1} & \rightarrow \ldots \rightarrow & s_n \\ \downarrow & & \downarrow & & \downarrow & & \downarrow \\ O_0 & & O_t & & O_{t+1} & & O_n \end{array}$$

$$\begin{aligned} \alpha_t(i) &= P(O_0, \ldots, O_t, s_t = i) \quad \text{current estimate about } s_t \\ \beta_t(i) &= P(O_{t+1}, \ldots, O_n | s_t = i) \quad \text{future evidence about } s_t \end{aligned}$$

- Using these probabilities we can compute the posterior probability that the HMM was in a particular state i at time t

$$P(s_t = i | O_0, \ldots, O_n) = \frac{\alpha_t(i)\beta_t(i)}{\sum_j \alpha_t(j)\beta_t(j)} \stackrel{def}{=} \gamma_t(i)$$

Forward/backward probabilities cont'd

$$
\begin{array}{ccccccc}
s_0 & \rightarrow \ldots \rightarrow & s_t & \rightarrow & s_{t+1} & \rightarrow \ldots \rightarrow & s_n \\
\downarrow & & \downarrow & & \downarrow & & \downarrow \\
O_0 & & O_t & & O_{t+1} & & O_n
\end{array}
$$

$$
\begin{aligned}
\alpha_t(i) &= P(O_0, \ldots, O_t, s_t = i) \quad \text{current estimate about } s_t \\
\beta_{t+1}(j) &= P(O_{t+2}, \ldots, O_n | s_{t+1} = j) \quad \text{future evidence about } s_{t+1}
\end{aligned}
$$

- We can also compute the posterior probability that the system was in state i at time t AND transitioned to state j at time $t+1$:

$$
\begin{aligned}
& P(s_t = i, s_{t+1} = j | O_0, \ldots, O_n) \\
&= \frac{\alpha_t(i) \overbrace{P_1(s_{t+1} = j | s_t = i) P_o(O_{t+1} | s_{t+1} = j)}^{\text{fixed } i \rightarrow j \text{ transition, one observation}} \beta_{t+1}(j)}{\sum_j \alpha_t(j) \beta_t(j)} \\
&\stackrel{def}{=} \xi_t(i, j),
\end{aligned}
$$

where $t = 0, \ldots, n-1$.

The EM algorithm for HMMs

Assume we have L observation sequences $O_0^{(l)}, \ldots, O_{n_l}^{(l)}$

E-step: compute the posterior probabilities

$$\begin{array}{rl} \gamma_t^{(l)}(i) & \text{for all } l,\ i, \text{ and } t\ (t = 0, \ldots, n_l) \\ \xi_t^{(l)}(i,j) & \text{for all } l,\ i, \text{ and } t\ (t = 0, \ldots, n_l - 1) \end{array}$$

M-step:

The initial state distribution can be updated according to the expected fraction of times the sequences started from a specific state i

$$\hat{P}_0(i) \leftarrow \frac{1}{L} \sum_{l=1}^{L} \gamma_0^{(l)}(i)$$

M-step cont'd

To update the transtion probabilities, we first define the expected number of transitions from i to j

$$\hat{N}(i,j) \;=\; \sum_{l=1}^{L}\sum_{t=0}^{n-1}\xi_t^{(l)}(i,j)$$

- The maximum likelihood estimate of the transition probabilities is then a ratio of these soft counts

$$\hat{P}_1(j|i) \;\leftarrow\; \frac{\hat{N}(i,j)}{\sum_{j'}\hat{N}(i,j')} = \frac{\#\text{ transitions } i \rightarrow j}{\#\text{ visits to } i}$$

- What about the output distributions?

M-step cont'd

- If the outputs are discrete, we define the expected number of times a particular observations say $O = k$ was generated from a specific state i

$$\hat{N}_O(i,k) = \sum_{l=1}^{L} \sum_{t=0}^{n_l} \gamma_t^{(l)}(i)\, \delta(O_t^{(l)}, k)$$

  where $\delta(O_t^{(l)}, k) = 1$ if $O_t^{(l)} = k$ and zero otherwise.

  The ML estimate is the ratio of this to the expected number of visits to i

$$\hat{P}_O(k|i) \;\leftarrow\; \frac{\hat{N}_O(i,k)}{\sum_{k'} \hat{N}_O(i,k')} = \frac{\#\text{ outputs } k \text{ in state } i}{\#\text{ visits to } i}$$

M-step cont'd

- If the outputs are continuous (e.g., multi-variate Gaussian), we have to solve a weighted maximum likelihood estimation problem

  Separately for each i we maximize:

$$\sum_{l=1}^{L} \sum_{t=0}^{n_l} \gamma_t^{(l)}(i) \log P(O_t^{(l)}|i)$$