Machine Learning

4771

Instructors:
Adrian Weller and Ilia Vovsha
Lecture 25

• HMMs with Evidence
• HMM Collect
• HMM Evaluate
• HMM Distribute
• HMM Decode
• HMM Parameter Learning via JTA & EM
Recall HMM Basic Operations

• Would like to do 3 basic things with our HMMs:
  1) Evaluate: given \( y_0, \ldots, y_T & \theta \) compute \( p(y_1, \ldots, y_T) \)
  2) Decode/inference: given \( y_0, \ldots, y_T & \theta \) find MAP \( q_0, \ldots, q_T \)
      or marginals \( p(q_0), \ldots, p(q_T) \)
  3) Max Likelihood Learn: given \( y_0, \ldots, y_T \) learn parameters \( \theta \)

• Typically use Baum-Welch (\( \alpha-\beta \) algo)... JTA is more general:

HMMs easily get Junction Tree
HMMs: JTA Init & Verify

- **Init:**\( \psi(q_0, y_0) = p(q_0)p(y_0 | q_0) \) \( \psi(q_t, q_{t+1}) = p(q_{t+1} | q_t) = \alpha_{q_t, q_{t+1}} \) \( \psi(q_t, y_t) = p(y_t | q_t) \)

- **Collect up from leaves:** doesn’t change zeta separators
  \( \varsigma^*(q_t) = \sum_{y_t} \psi(q_t, y_t) = \sum_{y_t} p(y_t | q_t) = 1 \)
  \( \psi^*(q_{t-1}, q_t) = \frac{\varsigma^*}{\varsigma} \psi(q_{t-1}, q_t) = \psi(q_{t-1}, q_t) \)

- **Collect left-right via phi’s:** changes backbone to marginals
  \( \phi^*(q_0) = \sum_{y_0} \psi(q_0, y_0) = p(q_0) \)
  \( \phi^*(q_t) = \sum_{q_{t-1}} \psi^*(q_{t-1}, q_t) = p(q_t) \)
  \( \psi^*(q_{t-1}, q_t) = \frac{\phi^*}{\phi} \psi(q_{t-1}, q_t) = p(q_0, q_1) \)
  \( \psi^*(q_{t-1}, q_t) = \frac{p(q_{t-1})}{p(q_{t-1} | q_{t-1})} p(q_t | q_{t-1}) = p(q_{t-1}, q_t) \)

- **Distribute:**
  \( \varsigma^{**}(q_t) = \sum_{q_{t-1}} \psi^*(q_{t-1}, q_t) = \sum_{q_{t-1}} p(q_{t-1}, q_t) = p(q_t) \)
  \( \psi^{**}(q_t, y_t) = \frac{\varsigma^{**}}{\varsigma^*} \psi(q_t, y_t) = \frac{p(q_t)}{1} p(y_t | q_t) = p(y_t, q_t) \)

\[ Z = 1 \]
\[ \phi(q_t) = 1 \]
\[ \varsigma(q_t) = 1 \]

...done!
HMMs: JTA with Evidence

• If y sequence is observed (in problems 1, 2, 3) get evidence:

\[ p(q, \bar{y}) = p(q_0) \prod_{t=1}^{T} p(q_t | q_{t-1}) \prod_{t=0}^{T} p(y_t | q_t) \]

• The potentials turn into slices:

\[ \psi(q_0, \bar{y}_0) \quad \psi(q_0, q_1) \quad \phi(q_0) \quad \phi(q_1) \quad \varsigma(q_t) \times \sum_{y_t} \psi(q_t, y_t) \]

\[ \varsigma^*(q_t) = \psi(q_t, \bar{y}_t) = p(\bar{y}_t | q_t) \]

• Next, pick a root, for example rightmost one: \( \psi(q_{T-1}, q_T) \)

• Collect all zeta separators bottom up:

\[ \varsigma^*(q_t) = \psi(q_t, \bar{y}_t) = p(\bar{y}_t | q_t) \]

• Collect leftmost phi separator to the right:

\[ \phi^*(q_0) = \sum_{y_0} \psi(q_0, \bar{y}_0) \delta(y_0 - \bar{y}_0) = p(\bar{y}_0, q_0) \]
HMMs: Collect with Evidence

- Now, we will collect (*) along the backbone left to right
- Update each clique with its left and bottom separators:

\[
\psi^* (q_t, q_{t+1}) = \frac{\phi^* (q_t)}{1} \zeta^* (q_{t+1}) \psi (q_t, q_{t+1}) = \phi^* (q_t) p (\bar{y}_{t+1} | q_{t+1}) \alpha_{q_{t}, q_{t+1}}
\]

\[
\phi^* (q_{t+1}) = \sum_{q_t} \psi^* (q_t, q_{t+1}) = \sum_{q_t} \phi^* (q_t) p (\bar{y}_{t+1} | q_{t+1}) \alpha_{q_{t}, q_{t+1}}
\]

- Keep going along chain until right most node
- Note: above formula for phi is recursive, could use as is.
- Recall we had \( \phi^* (q_0) = p (\bar{y}_0, q_0) \)
- Hence \( \psi^* (q_0, q_1) = p (\bar{y}_0, q_0) p (\bar{y}_1 | q_1) p (q_1 | q_0) \)
  \[= p (\bar{y}_0, q_0) p (\bar{y}_1 | q_1, q_0, \bar{y}_0) p (q_1 | q_0, \bar{y}_0) \quad \text{conditional indep} \]
  \[= p (\bar{y}_0, q_0) p (\bar{y}_1, q_1 | q_0, \bar{y}_0) = p (\bar{y}_0, \bar{y}_1, q_0, q_1) \]
HMMs: Evaluate with Evidence

Hence \[ \phi^*(q_1) = \sum_{q_0} \psi^*(q_0, q_1) = \sum_{q_0} p(\bar{y}_0, \bar{y}_1, q_0, q_1) = p(\bar{y}_0, \bar{y}_1, q_1) \]

Continuing, we obtain...

\[ \psi^*(q_{t-1}, q_t) = p(\bar{y}_0, \ldots, \bar{y}_t, q_{t-1}, q_t) \]

\[ \phi^*(q_t) = \sum_{q_{t-1}} \psi^*(q_{t-1}, q_t) = \sum_{q_{t-1}} p(\bar{y}_0, \ldots, \bar{y}_t, q_{t-1}, q_t) = p(\bar{y}_0, \ldots, \bar{y}_t, q_t) \]

These are the marginals AND observed data up to the point \( t \)
HMMs: Evaluate with Evidence

• If we are solving the first HMM problem, likelihood:
  1) **Evaluate**: given \( y_0, \ldots, y_T \) & \( \theta \), compute \( p(y_0, \ldots, y_T | \theta) \)
• We are already almost done! Collect is enough.

• As we collect to the root (rightmost node), we finally get:

\[
\psi^* (q_{T-1}, q_T) = p(\overline{y}_0, \ldots, \overline{y}_T, q_{T-1}, q_T)
\]

* marginal AND all observed data

• Can compute the likelihood just by summing this root \( \psi^* \)

\[
p(\overline{y}_0, \ldots, \overline{y}_T) = \sum_{q_{T-1},q_T} p(\overline{y}_0, \ldots, \overline{y}_T, q_{T-1}, q_T) = \sum_{q_{T-1},q_T} \psi^* (q_{T-1}, q_T)
\]

• What should we expect in the distribute phase?
HMMs: Distribute with Evidence

- Now, we distribute (**) along the backbone right to left
- Have first ** for root (stays the same): $\psi^*(q_{T-1}, q_T) = \psi^*(q_{T-1}, q_T)$
- Start distributing from there:

\[
\phi^{**}(q_t) = \sum_{q_{t+1}} \psi^{**}(q_t, q_{t+1})
\]

\[
\psi^{**}(q_t, q_{t+1}) = \frac{\phi^{**}(q_{t+1})}{\phi^*(q_{t+1})} \psi^*(q_t, q_{t+1})
\]

\[
\psi^{**}(q_{T-2}, q_{T-1}) = \frac{\phi^{**}(q_{T-1})}{\phi^*(q_{T-1})} p(\overline{y}_0, \ldots, \overline{y}_{T-1}, q_{T-2}, q_{T-1})
\]

\[
\psi^{**}(q_T, y_T) = \frac{\zeta^{**}(q_T)}{p(\overline{y}_T | q_T) p(\overline{y}_T | q_T)}\]

\[
= p(\overline{y}_0, \ldots, \overline{y}_T, q_T)
\]

why?
HMMs: Distribute with Evidence

Examine the general case:

Assume $\psi^{**}(q_t, q_{t+1}) = p(\overline{y}_0, \ldots, \overline{y}_T, q_t, q_{t+1})$ [true for $t + 1 = T$]

Then $\phi^{**}(q_t) = \sum_{q_{t+1}} \psi^{**}(q_t, q_{t+1}) = p(\overline{y}_0, \ldots, \overline{y}_T, q_t)$

$\psi^{**}(q_{t-1}, q_t) = \frac{\phi^{**}(q_t)}{\phi^*(q_t)} \psi^*(q_{t-1}, q_t) = \frac{p(\overline{y}_0, \ldots, \overline{y}_T, q_t)}{p(\overline{y}_0, \ldots, \overline{y}_T, q_t)} p(\overline{y}_0, \ldots, \overline{y}_T, q_t, q_{t-1}, q_t)
= p(\overline{y}_{t+1}, \ldots, \overline{y}_T | \overline{y}_0, \ldots, \overline{y}_T, q_t) p(\overline{y}_0, \ldots, \overline{y}_T, q_{t-1}, q_t)
= p(\overline{y}_{t+1}, \ldots, \overline{y}_T | \overline{y}_0, \ldots, \overline{y}_T, q_{t-1}, q_t) p(\overline{y}_0, \ldots, \overline{y}_T, q_{t-1}, q_t)
= p(\overline{y}_0, \ldots, \overline{y}_T, \overline{y}_T, q_{t-1}, q_t)$  conditional independence, see below

Hence result holds for all $t$
HMMs: Marginals & MaxDecoding

- After JTA is finished, we have the following:
  \[ \phi^{**}(q_t) \propto p(q_t | \bar{y}_1, \ldots, \bar{y}_T) \]
  \[ s^{**}(q_{t+1}) \propto p(q_{t+1} | \bar{y}_1, \ldots, \bar{y}_T) \]
  \[ \psi^{**}(q_t, q_{t+1}) \propto p(q_t, q_{t+1} | \bar{y}_1, \ldots, \bar{y}_T) \]

  Normalize to get these conditionals.

- We have solved part of the HMM Problem:
  2) **Decode**: given \( y_0, \ldots, y_T \) and \( \theta \) find \( p(q_0), \ldots, p(q_T) \) and \( q_0, \ldots, q_T \)

- The separators define a distribution over the hidden states
  - e.g. the probability the audio \( y_t \) was due to phoneme \( q_t \)
  - We can also decode to find the most likely path \( q_0 \ldots q_T \)

- Here, we use the ArgMax JTA algorithm

- Run JTA but replace sums with max

- Then, find biggest entry in separators:
  \[ \hat{q}_t = \arg \max_{q_t} \phi^{**}(q_t) \quad \forall t = 0 \ldots T \]
HMMs: EM Learning

- Finally 3) Max Likelihood: given $y_0, ..., y_T$ learn parameters $\theta$
- Recall max likelihood: 
  $$\hat{\theta} = \arg \max_\theta \log p(y | \theta)$$
- If observe $q$, it’s easy to maximize the complete likelihood:
  $$l(\theta) = \log \left( p(q, y) \right)$$
  $$= \log \left( p(q_0) \prod_{t=1}^T p(q_t | q_{t-1}) \prod_{t=0}^T p(y_t | q_t) \right)$$
  $$= \log p(q_0) + \sum_{t=1}^T \log p(q_t | q_{t-1}) + \sum_{t=0}^T \log p(y_t | q_t)$$
  $$= \log \prod_{i=1}^M [\pi_i]^{q^i_0} + \sum_{t=1}^T \log \prod_{i=1}^M \prod_{j=1}^M [\alpha_{ij}]^{q^i_{t-1}q^j_t} + \sum_{t=0}^T \log \prod_{i=1}^M \prod_{j=1}^N [\eta_{ij}]^{q^i_t y^j_t}$$

Introduce Lagrangian & take derivatives

$$\hat{\pi}_i = q^i_0$$
$$\hat{\alpha}_{ij} = \frac{\sum_{t=0}^{T-1} q^i_t q^{j}_{t+1}}{\sum_{k=1}^M \sum_{t=0}^{T-1} q^i_t q^{k}_{t+1}}$$
$$\hat{\eta}_{ij} = \frac{\sum_{t=0}^T q^i_t y^j_t}{\sum_{k=1}^N \sum_{t=0}^T q^i_t y^k_t}$$
HMMs: EM Learning

• But, we don’t observe the q’s, incomplete...

\[ p(\bar{y} | \theta) = \sum_q p(q, \bar{y} | \theta) = \sum_{q_0} \cdots \sum_{q_T} p(q_0) \prod_{t=1}^{T} p(q_t | q_{t-1}) \prod_{t=0}^{T} p(y_t | q_t) \]

• EM: Max expected complete likelihood given current p(q)

\[ E \{ l(\theta) \} = E_{p(q_0, \ldots, q_T | y)} \left\{ \log p(q, y) \right\} = \sum_{q_0} \cdots \sum_{q_T} p(q | y) \log p(q, y) \]
\[ = E \left\{ \sum_{i=1}^{M} q_0^i \log \pi_i + \sum_{t=1}^{T-1} \sum_{i,j=1}^{M} q_t^i q_{t+1}^j \log \alpha_{ij} + \sum_{t=0}^{T} \sum_{i=1}^{M} \sum_{j=1}^{N} q_t^i y_t^j \log \eta_{ij} \right\} \]
\[ = \sum_{i=1}^{M} E \{ q_0^i \} \log \pi_i + \sum_{t=1}^{T} \sum_{i,j=1}^{M} E \{ q_{t-1}^i q_t^j \} \log \alpha_{ij} + \sum_{t=0}^{T} \sum_{i=1}^{M} \sum_{j=1}^{N} E \{ q_t^i \} y_t^j \log \eta_{ij} \]

• M-step is maximizing as before:

\[ \hat{\pi}_i = E \{ q_0^i \} \]
\[ \hat{\alpha}_{ij} = \frac{\sum_{t=0}^{T-1} E \{ q_t^i q_{t+1}^j \}}{\sum_{k=1}^{M} \sum_{t=0}^{T-1} E \{ q_t^i q_{t+1}^k \}} \]
\[ \hat{\eta}_{ij} = \frac{\sum_{t=0}^{T} E \{ q_t^i y_t^j \}}{\sum_{k=1}^{N} \sum_{t=0}^{T} E \{ q_t^i y_t^k \}} \]

• What are E{}’s?
HMMs: EM Learning

• But, we don’t observe the q’s, incomplete...

\[
p(y | \theta) = \sum q \ p(q, y | \theta) = \sum q_0 \cdots \sum q_T \ p(q_0) \prod_{t=1}^{T} p(q_t | q_{t-1}) \prod_{t=0}^{T} p(y_t | q_t)
\]

• EM: Max expected complete likelihood given current \( p(q) \)

\[
E \left\{ l(\theta) \right\} = E_{p(q_0, \ldots, q_T | y)} \left\{ \log p(q, y) \right\} = \sum q_0 \cdots \sum q_T \ p(q | y) \log p(q, y)
\]

\[
= E \left\{ \sum_{i=1}^{M} q_0^i \log \pi_i + \sum_{t=1}^{T} \sum_{i,j=1}^{M} q_t^i q_{t-1}^j \log \alpha_{ij} + \sum_{t=0}^{T} \sum_{i=1}^{M} \sum_{j=1}^{N} q_t^i y_t^j \log \eta_{ij} \right\}
\]

\[
= \sum_{i=1}^{M} E \left\{ q_0^i \right\} \log \pi_i + \sum_{t=1}^{T} \sum_{i,j=1}^{M} E \left\{ q_t^i q_{t-1}^j \right\} \log \alpha_{ij} + \sum_{t=0}^{T} \sum_{i=1}^{M} \sum_{j=1}^{N} E \left\{ q_t^i \right\} y_t^j \log \eta_{ij}
\]

• M-step is maximizing as before:

\[
\hat{\pi}_i = E \left\{ q_0^i \right\} \quad \hat{\alpha}_{ij} = \frac{\sum_{t=0}^{T-1} E \left\{ q_t^i q_{t+1}^j \right\}}{\sum_{k=1}^{M} \sum_{t=0}^{T-1} E \left\{ q_t^i q_t^k \right\}} \quad \hat{\eta}_{ij} = \frac{\sum_{t=0}^{T} E \left\{ q_t^i \right\} y_t^j}{\sum_{k=1}^{N} \sum_{t=0}^{T} E \left\{ q_t^i \right\} y_t^k}
\]

• What are E{}’s?

\[
E_{p(x)} \{ x^i \} = \sum_x p(x) x^i = \sum_x p(x) \delta(x = x^i) = p(x^i)
\]
HMMs: EM Learning

• But, we don’t observe the q’s, incomplete...

\[
p(y | \theta) = \sum_q p(q, y | \theta) = \sum_{q_0} \cdots \sum_{q_T} p(q_0) \prod_{t=1}^T p(q_t | q_{t-1}) \prod_{t=0}^T p(y_t | q_t)
\]

• EM: Max expected complete likelihood given current p(q)

\[
E \{l(\theta)\} = \mathcal{E}_{p(q_0, \ldots, q_T | y)} \{ \log p(q, y) \} = \sum_{q_0} \cdots \sum_{q_T} p(q | y) \log p(q, y)
\]

\[
= E \left\{ \sum_{t=1}^M q_0^i \log \pi_i + \sum_{t=1}^T \sum_{i, j=1}^M q_t^i q_{t-1}^j \log \alpha_{ij} + \sum_{t=0}^T \sum_{i=1}^M \sum_{j=1}^N q_t^i y_t^j \log \eta_{ij} \right\}
\]

\[
= \sum_{i=1}^M E \{q_0^i\} \log \pi_i + \sum_{t=1}^T \sum_{i, j=1}^M E \{q_{t-1}^i q_t^j\} \log \alpha_{ij} + \sum_{t=0}^T \sum_{i=1}^M \sum_{j=1}^N E \{q_t^i\} y_t^j \log \eta_{ij}
\]

• M-step is maximizing as before:

\[
\hat{\pi}_i = E \{q_0^i\} \quad \hat{\alpha}_{ij} = \frac{\sum_{t=0}^{T-1} E \{q_t^i q_{t+1}^j\}}{\sum_{k=1}^M \sum_{t=0}^{T-1} E \{q_t^i q_{t+1}^k\}} \quad \hat{\eta}_{ij} = \frac{\sum_{t=0}^T E \{q_t^i\} y_t^j}{\sum_{k=1}^N \sum_{t=0}^T E \{q_t^i\} y_t^k}
\]

• What are E{}’s?

\[
E_{p(x)} \{x^i\} = \sum_x p(x) x^i = \sum_x p(x) \delta(x = x^i) = p(x^i)
\]

• Our JTA \(\psi\) & \(\phi\) marginals! (JTA is the E-Step for given \(\theta\))

\[
E \{q_t^i q_{t+1}^j\} = p(q_t = i, q_{t+1} = j | \bar{y}) \quad E \{q_t^i\} = p(q_t = i | \bar{y})
\]
HMMs: Gaussian Emissions

- Instead of table for emissions, have Gaussian:

\[
p(\bar{y} | \theta) = \sum_q p(q, \bar{y} | \theta) = \sum_{q_0} \cdots \sum_{q_T} p(q_0) \prod_{t=1}^{T} p(q_t | q_{t-1}) \prod_{t=0}^{T} p(\bar{y}_t | q_t)
\]

where

\[
p(\bar{y}_t | q_t) = N(\bar{y}_t | \mu_{q_t}, I)
\]

- Clique initialization:

\[
\psi(q_t, \bar{y}_t) = \psi(q_t) = N(\bar{y}_t | \mu_{q_t}, I)
\]

- M-step is maximizing as before:

\[
\hat{\pi}_i = E\{q_0^i\} \quad \hat{\alpha}_{ij} = \frac{\sum_{t=0}^{T-1} E\{q_t^i q_{t+1}^j\}}{\sum_{k=1}^{M} \sum_{t=0}^{T-1} E\{q_t^i q_{t+1}^k\}} \quad \hat{\mu}_i = \frac{\sum_{t=0}^{T} E\{q_t^i\}\bar{y}_t}{\sum_{t=0}^{T} E\{q_t^i\}}
\]

- Can thus handle continuous time series as in speech recognition