Machine Learning

4771

Instructor: Tony Jebara
Topic 25

• Review: HMMs Basic Operations
• HMMs via the Junction Tree Algorithm & EM
• Other Graphical Models
• Structure Learning
Review: HMMs JTA Evidence

- Now, if observe sequence (problems 1, 2, 3) have evidence
  \[ p(q, \bar{y}) = p(q_0) \prod_{t=1}^{T} p(q_t | q_{t-1}) \prod_{t=0}^{T} p(\bar{y}_t | q_t) \]

- The potentials turn into slices:
  \[ \psi(q_0, y_0) \quad \psi(q_0, q_1) \quad \phi(q_0) \quad \phi(q_1) \quad \psi(q_1, \bar{y}_1) \quad s^*(q_t) \times \sum_{y_t} \psi(q_t, y_t) \quad \psi(q_1, \bar{y}_1) \]
  \[ s^*(q_t) = \psi(q_t, \bar{y}_t) = p(\bar{y}_t | q_t) \]

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

- Collect all zeta separators bottom up:
  \[ s^*(q_t) = \psi(q_t, \bar{y}_t) = p(\bar{y}_t | q_t) \]

- Collect leftmost phi separator to the right:
  \[ \phi^*(q_0) = \sum_{y_0} \psi(q_0, \bar{y}_0) \delta(y_0 - \bar{y}_0) = p(\bar{y}_0, q_0) \]
Review: HMMs JTA Collect

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

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

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

- Keep going along chain until right most node
- Note: above formula for phi is recursive, could use as is.
- Property: recall we had \( \phi^* (q_0) = p(\overline{y}_0, q_0) \)
\[
\phi^* (q_1) = \sum_{q_0} p(\overline{y}_0, q_0) p(\overline{y}_1 | q_1) p(q_1 | q_0) = p(\overline{y}_0, \overline{y}_1, q_1)
\]
\[
\phi^* (q_2) = \sum_{q_1} p(\overline{y}_0, \overline{y}_1, q_1) p(\overline{y}_2 | q_2) p(q_2 | q_1) = p(\overline{y}_0, \overline{y}_1, \overline{y}_2, q_2)
\]
\[
\phi^* (q_{t+1}) = \sum_{q_t} p(\overline{y}_0, \ldots, \overline{y}_t, q_t) p(\overline{y}_{t+1} | q_{t+1}) p(q_{t+1} | q_t) = p(\overline{y}_0, \ldots, \overline{y}_{t+1}, q_{t+1})
\]
HMMs: Evaluate

• Say we are solving the first HMM problem:
  1) **Evaluate**: given \( y_1, \ldots, y_T \) compute likelihood \( p(y_1, \ldots, y_T) \)
• If we want to compute the likelihood, we are already done!
• We really just need to do collect (not even distribute).
• From previous slide we had:
  \[
  \phi^*(q_{t+1}) = \sum_{q_t} p(y_0, \ldots, y_t, q_t) p(y_{t+1} | q_{t+1}) p(q_{t+1} | q_t) = p(y_0, \ldots, y_{t+1}, q_{t+1})
  \]
• As we collect to the root (rightmost node), we finally get:
  \[
  \phi^*(q_T) = p(y_0, \ldots, y_T, q_T)
  \]

• Can compute the likelihood just by marginalizing this \( \phi^* \)
  \[
  p(y_0, \ldots, y_T) = \sum_{q_T} p(y_0, \ldots, y_T, q_T) = \sum_{q_T} \phi^*(q_T)
  \]
• So, adding up the entries in last \( \phi^* \) gives us the likelihood
HMMs: JTA Distribute

- Back to collecting... say just finished collecting to the root with our last update formula:

\[ \psi^* (q_{T-1}, q_T) = \frac{\phi^* (q_{T-1})}{1} \frac{\zeta^* (q_T)}{1} \psi (q_{T-1}, q_T) = \phi^* (q_{T-1}) \ p(y_T | q_T) \ a_{q_{T-1} \ q_T} \]

- 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 going to the left from there:

\[ \phi^{**} (q_t) = \sum_{q_{t+1}} \psi^{**} (q_t, q_{t+1}) \]
\[ \zeta^{**} (q_{t+1}) = \sum_{q_t} \psi^{**} (q_t, q_{t+1}) \]
\[ \psi^{**} (q_t, q_{t+1}) = \frac{\phi^{**} (q_{t+1})}{\phi^* (q_{t+1})} \psi^* (q_t, q_{t+1}) \]
HMMs: Decode & Evaluate

• Next HMM Problem:
  2) **Decode**: given $y_1, \ldots, y_T$ compute best $q_1, \ldots, q_T$ or $p(q_t)$

• Now that JTA is finished, we have the following:
  \[
  \phi^{**}(q_t) \propto p(q_t \mid \overline{y}_1, \ldots, \overline{y}_T) \quad \psi^{**}(q_{t+1}) \propto p(q_{t+1} \mid \overline{y}_1, \ldots, \overline{y}_T)
  \]
  \[
  \psi^{**}(q, q_{t+1}) \propto p(q, q_{t+1} \mid \overline{y}_1, \ldots, \overline{y}_T)
  \]

• The separators define a distribution over the hidden states
• This tells us the probability the audio $y_t$ was phoneme $q_t$
• Decode: get most likely path $q_0 \ldots q_T$
• Do ArgMax JTA algorithm instead
• 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 $y_1, \ldots, y_T$ learn parameters $\theta$
- Recall max likelihood: $\hat{\theta} = \arg \max \theta \log p(y | \theta)$
- If we had complete likelihood it would be easy:

$$l(\theta) = \log(p(q, y))$$

$$= \log(p(q_0) \prod_{t=1}^{T} p(q_t | q_{t-1}) \prod_{t=0}^{T} p(\bar{y}_t | q_t))$$

$$= \log p(q_0) + \sum_{t=1}^{T} \log p(q_t | q_{t-1}) + \sum_{t=0}^{T} \log p(\bar{y}_t | q_t)$$

$$= \log \prod_{i=1}^{M} \pi_i^{q_i^0} + \sum_{t=1}^{T} \log \prod_{i=1}^{M} \prod_{j=1}^{M} a_{ij}^{q^i_{t-1}q^i_t} + \sum_{t=0}^{T} \log \prod_{i=1}^{M} \prod_{j=1}^{N} \eta_{ij}^{q^i_t y^j_t}$$

$$= \sum_{i=1}^{M} q_i^0 \log \pi_i + \sum_{t=1}^{T} \sum_{i,j=1}^{M} q^i_{t-1} q^j_t \log a_{ij} + \sum_{t=0}^{T} \sum_{i=1}^{M} \sum_{j=1}^{N} q^i_t y^j_t \log \eta_{ij}$$

Introduce Lagrange & take derivatives

$$\hat{\pi}_i = q_i^0$$

$$\hat{a}_{ij} = \frac{\sum_{t=0}^{T-1} q^i_t q^j_{t+1}}{\sum_{k=1}^{M} \sum_{t=0}^{T-1} q^i_t q^k_{t+1}}$$

$$\hat{\eta}_{ij} = \frac{\sum_{t=0}^{T} q^i_t y^j_t}{\sum_{k=1}^{N} \sum_{t=0}^{T} q^i_t y^k_t}$$

$$\sum_{i=1}^{M} \pi_i = 1 \quad \sum_{j=1}^{M} a_{ij} = 1 \quad \sum_{j=1}^{N} \eta_{ij} = 1$$
HMMs: EM Learning

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

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

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

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

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

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

- M-step is maximizing as before:

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

- What are \( E\{\} \)'s?
HMMs: EM Learning

• But, we don’t observe the q’s, incomplete...
  \[ p(\bar{y} | \theta) = \sum_q p(q, \bar{y} | \theta) = \sum_{q_0} \ldots \sum_{q_T} p(q_0) \prod_{t=1}^{T} p(q_t | q_{t-1}) \prod_{t=0}^{T} p(\bar{y}_t | q_t) \]

• EM: Max expected complete likelihood given current p(q)
  \[ E\{l(\theta)\} = E_{p(q_0, \ldots, q_T | y)} \{\log p(q, y)\} = \sum_{q_0} \ldots \sum_{q_T} p(q | y) \log p(q, y) \]
  \[ = E\left\{\sum_{i=1}^{M} q_i^0 \log \pi_i + \sum_{t=1}^{T} \sum_{i,j=1}^{M} q_i^t q_j^{t-1} \log a_{ij} + \sum_{t=0}^{T} \sum_{i=1}^{M} \sum_{j=1}^{N} q_i^t y_j^t \log \eta_{ij}\right\} \]
  \[ = \sum_{i=1}^{M} E\{q_0^i\} \log \pi_i + \sum_{t=1}^{T} \sum_{i,j=1}^{M} E\{q_i^t q_j^{t-1}\} \log a_{ij} + \sum_{t=0}^{T} \sum_{i=1}^{M} \sum_{j=1}^{N} E\{q_i^t\} y_j^t \log \eta_{ij} \]

• M-step is maximizing as before:
  \[ \hat{\pi}_i = E\{q_0^i\} \quad \hat{a}_{ij} = \frac{\sum_{t=0}^{T-1} E\{q_i^t q_{j}^{t+1}\}}{\sum_{k=1}^{M} \sum_{t=0}^{T-1} E\{q_i^t q_k^{t+1}\}} \quad \hat{\eta}_{ij} = \frac{\sum_{t=0}^{T} E\{q_i^t\} y_j^t}{\sum_{k=1}^{N} \sum_{t=0}^{T} E\{q_i^t\} y_k^t} \]

• 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...

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

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

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

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

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

M-step is maximizing as before:

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

What are E{}’s?

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

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

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

- Instead of table for emissions, have Gaussian:

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

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

- Clique initialization:

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

- **M-step** is maximizing as before:

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

- Can thus handle continuous time series:

  but is this the right way??

\[
0 \quad 10 \quad 20 \quad 30 \quad 40 \quad 50
\]

\[
-1 \quad 0 \quad 0.5 \quad 1
\]
Other Graphical Models

- Could be completely observed or hidden (HMMs, need EM)
- Hierarchical HMMs:
  - different scales of time
  - bottom = audio, phoneme, ...
  - word, part-of-speech, ...top
- Medical Diagnostics:
  - bi-partite graph
  - top = diseases
  - bottom = symptoms
- Society of HMMs:
  - multiple audio speakers interacting at a party
- Caveat: some need more algorithms, loopy, EM++, JTA++
Structure Learning

- We could even learn graph structure itself...
- We did this in homework by cross-validation...
- If no hidden variables, & tree structure, can be efficient!
- Use (max likelihood) Chow-Liu Algorithm:

\[ \text{Tree}^* = \arg \max_{T \in \text{Trees}} \max_{\theta_T} p(\text{data} \mid T, \theta_T) \]

1) Start with nodes unconnected
2) For each pair of (discrete) variables \((x_i, x_j)\) compute pdf tables by counting the data
3) If \( \hat{p}(x_i, x_j) \approx \hat{p}(x_i) \hat{p}(x_j) \) not really worth connecting them so Measure how far apart factorized is to joint via KL
   \[ \text{Mutual Information} (i, j) = KL(\hat{p}(x_i, x_j) \mid \mid \hat{p}(x_i) \hat{p}(x_j)) \]
4) Want tree that max’s total KL
5) Use Kruskal to get maximum weight spanning tree where:

\[ W_{ij} = \sum_x \sum_x \hat{p}(x_i, x_j) \log \frac{\hat{p}(x_i, x_j)}{\hat{p}(x_i) \hat{p}(x_j)} \]