27/6

PMAI - R11 / EM

PGM: Representation / Inference / Learning

Parameters / structure

Complete Data / Missing Data

The EM idea: if we know the value of the missing data we could find MLE $\hat{\theta}$, maximizing the log-likelihood.

- E: Try to complete (guess) the missing data based on observed data and current params $\theta^{(t)}$ — Inference
- M: Find new params $\theta^{(t+1)}$ maximizing the log-likelihood w. the filled-in data.

* We do it in a soft manner, computing and maximizing the expected log likelihood.

- maximizing w.r.t. θ
- expected w.r.t. $P(H \mid O, \theta^{(t)})$
- expected sufficient statistics.

Formally:

1. Initialize $\theta^{(0)}$ to some values
2. Calculate

$$Q(\theta; \theta^{(t)}) = \sum_m \sum_{x_{h[m]}} P(x_{h[m]} \mid x_{o[m]}; \theta^{(t)}) \cdot \log P(x_{h[m]}, x_{o[m]}; \theta)$$

3. Find $\theta^{(t+1)} = \arg\max_\theta Q(\theta; \theta^{(t)})$

$$= \mathbb{E}_{P(x_h \mid x_o; \theta^{(t)})} \log P(x_h, x_o; \theta)$$

The expected complete-data log-likelihood w.r.t the missing data posterior (using current parameters).

EM - Proof of convergence

(show that $l(\theta^{(t+1)} : D) \geq l(\theta^{(t)} : D)$)

For any $q(x_h)$ (and specifically $q(x_h) = P(x_h | x_o ; \theta^{(t)})$)

for a single sample $\langle x_o, x_h \rangle[m]$

$$\rightarrow \sum_{x_h} q(x_h) \log P(x_h, x_o ; \theta)$$

$$= \sum_{x_h} q(x_h) \log \left[P(x_o ; \theta) P(x_h | x_o ; \theta) \right] \quad \leftarrow \text{chain rule}$$

$$= \sum_{x_h} q(x_h) \log P(x_o ; \theta) + \sum_{x_h} q(x_h) \log P(x_h | x_o ; \theta)$$

$$= \log P(x_o ; \theta) + \sum_{x_h} q(x_h) \log \frac{P(x_h | x_o ; \theta)}{q(x_h)} + \sum_{x_h} q(x_h) \log q(x_h)$$

$$= l(x_o ; \theta) - KL\left(q(x_h) \,||\, P(x_h | x_o ; \theta)\right) - H(q(x_h))$$

(minus)

$\Downarrow$

$Q(\theta^{(t+1)} ; \theta^{(t)}) \geq Q(\theta^{(t)} ; \theta^{(t)})$ by maximization.

$\downarrow$

$$0 \leq Q(\theta^{(t+1)} ; \theta^{(t)}) - Q(\theta^{(t)} ; \theta^{(t)})$$

$$= l(D ; \theta^{(t+1)}) - l(D ; \theta^{(t)})$$

$$- KL\left(P(H|D ; \theta^{(t)}) \,||\, P(H|D ; \theta^{(t+1)})\right) + KL\left(P(H|D ; \theta^{(t)}) \,||\, P(H|D ; \theta^{(t)})\right)$$

(same)

$$- \cancel{H(q(H))} + \cancel{H(q(H))}$$

$\downarrow$

$$l(D ; \theta^{(t+1)}) \geq l(D ; \theta^{(t)}) + C$$

($C \geq 0$)

□

Note: $Q(\theta ; \theta^{(t)}) = \sum_m \sum_{x_h[m]} P(x_h[m] | x_o[m] ; \theta^{(t)}) \log P(x_h[m], x_o[m] ; \theta)$

(θ: new, $\theta^{(t)}$: old; $P(x_h[m] | x_o[m] ; \theta^{(t)}) = q(x_h)$)

($\downarrow$ In $Q(\theta^{(t+1)} ; \theta^{(t)})$ and $Q(\theta^{(t)} ; \theta^{(t)})$, $q(H)$ is the same and θ changes.

$$= E_q \log P(\theta : D^+) = E_q \, l(\theta : D^+)$$

Exercise: Incremental EM - CPD at a time

Instead of updating all params in each M step, we select one parameter x_i and update only its CPD: $\theta_{x_i|Pa_i}$. Does the likelihood increase in each iteration? i.e.

$$\ell(\theta^{(t+1)}:D) \geq \ell(\theta^{(t)}:D)$$

prove your answer.

Answer: Yes

Proof:

We saw that:

$$\ell(\theta^{(t+1)}:D) - \ell(\theta^{(t)}:D) \geq \underset{P(H|D;\theta^{(t)})}{\mathbb{E}}\left[\ell(\theta^{(t+1)}:D^+\right] - \underbrace{\underset{P(H|D;\theta^{(t)})}{\mathbb{E}}\left[\ell(\theta^{(t)}:D^+\right]}_{Q(\theta^{(t+1)};\theta^{(t)}) - Q(\theta^{(t)};\theta^{(t)})}$$

↘

We need to show that $Q(\theta^{(t+1)};\theta^{(t)}) \geq Q(\theta^{(t)};\theta^{(t)})$

In the move from $\theta^{(t)}$ to $\theta^{(t+1)}$ we maximized

$$Q(\theta:\theta^{(t)})$$

under constraint that all params except $\theta_{x_i|Pa_i}$ are unchanged $\Rightarrow Q(\theta^{(t+1)};\theta^{(t)}) \geq Q(\theta^{(t)};\theta^{(t)})$.

□

[this is also true if we are maximizing

$$\underset{\theta_{x_i|Pa_i}}{\operatorname{argmax}} \underset{P(H|D;\theta^{(t)})}{\mathbb{E}} P(x_i[\xi_{x_i}];\theta_{x_i|Pa_i})$$

since other factors are unchanged]

EM Algorithm using Expected Sufficient Statistics

Algorithm EM (G, θ^0, D): (for table CPDs)

for $t = 1 \ldots$

$\bar{M}_t = ESS(G, \theta^t, D)$ // E

for $i = 1 \ldots n$

for x_i, u_i in $Val(X_i \cup Pa_i)$

$$\theta^{t+1}_{x_i|u_i} = \frac{\bar{M}_t[x_i, u_i]}{\bar{M}_t[u_i]}$$ // M

return θ^t

Procedure $ESS(G, \theta, D)$:

$\bar{M} = 0$

for $m = 1 \ldots M$ →

Perform inference* on (G, θ) using evidence $o[m]$

for $i = 1 \ldots n$

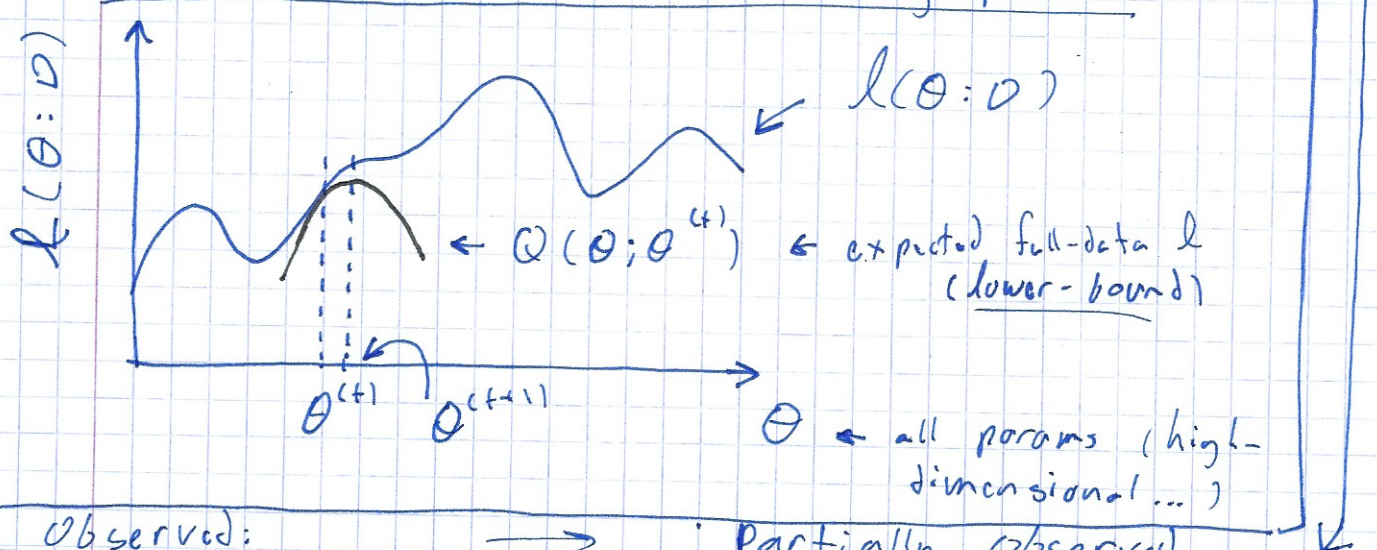
for x_i, u_i in $Val(X_i \cup Pa_i)$

$\bar{M}[x_i, u_i] \mathrel{+}= P(x_i, u_i \mid o[m])$ ← ! (hidden → for sensor life)

return $\bar{M}$

* Inference e.g. construct a C.T. w. evidence, for calculating all marginals $P(x_i, u_i | o)$

The EM likelihood hill-climbing process:



Fully Observed: → Partially Observed

$$M[x,u] = \sum_m 1\{\xi[m] \langle X, U \rangle = (x,u)\}$$

$$\bar{M}[x,u] = \sum_m P(x, u \mid o[m], \theta)$$

27/6

PMAI R11 / Normalized Importance Sampling

PGM → Inference → Approximate → sampling-based
→ Importance Sampling → NIS

Estimate the expectation of some $f(x)$ w.r.t. P by sampling from some Q. (and using weights based on Q and un-normalized version of P: $\tilde{P}(x) \propto P(x)$).

Derivation:

$$\mathbb{E}_P[f(x)] = \mathbb{E}_Q\left[f(x)\frac{P(x)}{Q(x)}\right] = \frac{1}{Z}\mathbb{E}_Q\left[f(x)\frac{\tilde{P}(x)}{Q(x)}\right]$$

$$= \frac{1}{Z}\mathbb{E}_Q[f(x)\,w(x)]$$

$$\mathbb{E}_Q[w(x)] = \mathbb{E}_Q\left[\frac{\tilde{P}(x)}{Q(x)}\right] = \sum_x \tilde{P}(x) = Z \qquad \left(P(x) = \frac{\tilde{P}(x)}{Z}\right)$$

$$\mathbb{E}_P[f(x)] = \left(\frac{\mathbb{E}_Q\left[f(x)\frac{\tilde{P}(x)}{Q(x)}\right]}{\mathbb{E}_Q[w(x)]} =\right) \frac{\mathbb{E}_Q[f(x)\,w(x)]}{\mathbb{E}_Q[w(x)]}$$

$$\hat{\mathbb{E}}_D^{NIS}[f(x)] = \frac{\frac{1}{m}\sum_{m=1}^{m} f(x[m])\,w(x[m])}{\frac{1}{m}\sum_{m=1}^{m} w(x[m])}, \quad w(x[m]) = \frac{\tilde{P}(x[m])}{Q(x[m])}$$

---

* "Sampling on a tree ..."

$$P(x|e) = \frac{P(x,e)}{P(e)} = \frac{P(x,e)}{\sum_x P(x,e)}$$ ← doesn't sum to 1 (joint w. some vals of e).

↳ $\tilde{P}(x|e) = P(x,e)$

↳

$$w[m] = \frac{P(x[m], e[m])}{Q(x[m], e[m] \mid e[m])}$$ ← un-normalized

← normalized

$$\hat{P}(X = x | E = e) = \left(\sum_m w[m]\,\mathbb{1}\{x[m] = x\}\right) / \left(\sum_m w[m]\right)$$

27/6

PMAI - R11 - Gibbs Sampling

PGM → Inference → Approximate → Sampling Based → Gibbs sampling

1. Initialize a sample w. $E=e$ and some values for the other variables.
2. for $t = 1 \dots T$ // for T samples
   $x(t) = x(t-1)$
   for x_i in $X \setminus E$
   sample $x_i^{(t)}$ from $P(x_i \mid x_{-i}^{(t)})$
3. return $x[0], \dots, x[T]$

→ We saw in PGM that $P(x_i \mid x_{-i})$ is easy.

67800: Probabilistic Methods in AI **Spring 2017/18**

Recitation 11: Inference and Learning Review

Teaching Assistant: Eitan Richardson

Prove that $P(X \mid \mathbf{e})$ is stationary w.r.t Gibbs process.

Proof:
A stationary distribution for a Markov chain $\mathcal{T}$:

$$\pi(X = x') = \sum_{x} \pi(X = x)\, \mathcal{T}(x \to x')$$

In our case: $\pi(X) = P(X \mid e)$ and $\mathcal{T}(x \to x') = P(x'_i \mid x_{-i}, \mathbf{e})$, for a random variable x_i.
We have:

$$\begin{aligned}
\sum_{x_i} P(x_i, x_{-i} \mid \mathbf{e}) P(x'_i \mid x_{-i}, \mathbf{e}) &= \sum_{x_i} P(x_i \mid x_{-i}, \mathbf{e}) P(x_{-i} \mid \mathbf{e}) \frac{P(x'_i, x_{-i} \mid \mathbf{e})}{P(x_{-i} \mid \mathbf{e})} \\
&= \sum_{x_i} P(x_i \mid x_{-i}, \mathbf{e}) P(x'_i, x_{-i} \mid \mathbf{e}) \\
&= P(x'_i, x_{-i} \mid \mathbf{e}) \\
&= P(x' \mid \mathbf{e})
\end{aligned}$$

■

* for a chosen i. All other variables treated as constant.

[1] *Original LaTeX template courtesy of UC Berkeley.*