10/21 * ML estimation for complete data

$P_{ML}(X_i = x \mid pa_i = \pi) = \text{count}(X_i=x, pa_i=\pi) / \text{count}(pa_i=\pi)$

* ML estimation for incomplete data

Examples $t = 1, 2, \cdots, T$

Hidden nodes $H^{(t)}$

Visible nodes $V^{(t)}$

$\boxed{\text{Review}}$

* EM algorithm

E-step : compute posterior probabilities

$P(X_i = x, pa_i = \pi \mid V^{(t)})$ inference

M-step : update CPTs

$P(X_i = x \mid pa_i = \pi) \longleftarrow \sum_t P(X_i = x, pa_i = \pi \mid V^{(t)}) / \sum_t P(pa_i = \pi \mid V^{(t)})$

Iterate until convergence. Note that RHS depends on current CPT estimates.

* Properties

1) no learning rate or tuning parameters

2) monotonic convergence

- each iteration improves log-likelihood $\mathcal{L} = \sum_t \log P(V^{(t)})$

Detour - numerical optimization

How to minimize $f(\vec{\theta})$?

1) gradient descent : $\vec{\theta}_{t+1} = \vec{\theta}_t - \eta \frac{\partial f}{\partial \vec{\theta}}$ learning rate $\eta > 0$

2) Newton's method : $\vec{\theta}_{t+1} = \vec{\theta}_t - H^{-1} \frac{\partial f}{\partial \vec{\theta}}$ (not always converge ; expensive to compute H)

3) Auxiliary function : $Q(\vec{\theta}, \vec{\theta}')$

Suppose $Q(\vec{\theta}, \vec{\theta}')$ satisfies : (i) $Q(\vec{\theta}, \vec{\theta}) = f(\vec{\theta})$ for all $\vec{\theta}$.

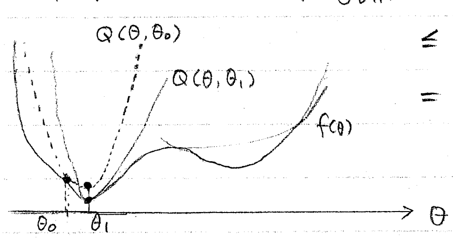
(ii) $Q(\vec{\theta}, \vec{\theta}') \geq f(\vec{\theta})$ for all $\vec{\theta}, \vec{\theta}'$

Consider update rule : $\vec{\theta}_{t+1} = \underset{\vec{\theta}}{\arg\min}\ Q(\vec{\theta}, \vec{\theta}_t)$

It follows that : $f(\vec{\theta}_{t+1}) \leq Q(\vec{\theta}_{t+1}, \vec{\theta}_t)$ by property (ii)

$\leq Q(\vec{\theta}_t, \vec{\theta}_t)$ because $\vec{\theta}_{t+1} = \underset{\vec{\theta}}{\arg\min}\ Q(\vec{\theta}, \vec{\theta}_t)$

$= f(\vec{\theta}_t)$



Properties : - no learning rate

- monotonic improvement

- convergence to local stationary point (local minimum in general)

How to derive auxiliary function for ML estimation?

* Key inequality

Let $P(X)$ and $\tilde{P}(X)$ be different distributions over $X = \{X_1, X_2, \cdots, X_n\}$

$$\log \tilde{P}(V) = \log \left[\tilde{P}(h,v) / \tilde{P}(h|v) \right] \quad \text{for any instantiation } h \in H \text{ of hidden nodes}$$

$$= \sum_h P(h|v) \log \left[\tilde{P}(h,v) / \tilde{P}(h|v) \right]$$

$$= \sum_h P(h|v) \{ \log \tilde{P}(h,v) - \log \tilde{P}(h|v) + \log P(h|v) - \log P(h|v) \}$$

$$= \sum_h P(h|v) \{ \log \tilde{P}(h,v) - \log P(h|v) + \log \frac{P(h|v)}{\tilde{P}(h|v)} \}$$

$$= \sum_h P(h|v) \{ \log \tilde{P}(h,v) - \log P(h|v) \} + KL(P(h|v), \tilde{P}(h|v))$$

$$\log \tilde{P}(V) \geq \sum_h P(h|v) \{ \log \tilde{P}(h,v) - \log P(h|v) \}$$

* Relation to EM algorithm

Imagine $P(X) = P_{OLD}(X; \theta)$ with old CPTs θ

Imagine $\tilde{P}(X) = P_{NEW}(X; \theta')$ with new CPTs θ'

How to derive update rule $\theta \rightarrow \theta'$, $P(X) \rightarrow \tilde{P}(X)$?

* Formal statement of EM

(i) E step

Compute auxiliary function

$$Q(\theta, \theta') = \sum_t \sum_h P(h|V^{(t)}) \log \tilde{P}(h, V^{(t)}) - \sum_t \sum_h P(h|V^{(t)}) \log P(h|V^{(t)})$$

old and new CPTs

expected value of $\log \tilde{P}(h, V^{(t)})$

(ii) M-step

Maximize $\sum_t \sum_h P(h|V^{(t)}) \log \tilde{P}(h, V^{(t)})$ in terms of new CPTs $\tilde{P}(X_i = x | pa_i = \pi)$

* Convergence proof

Suppose we choose new CPTs in this way.

$$\mathcal{L}_{new} = \sum_t \log \tilde{P}(V^{(t)})$$

$$\geq \sum_t \{ \sum_h P(h|V^{(t)}) \log \tilde{P}(h, V^{(t)}) - \sum_h P(h|V^{(t)}) \log P(h|V^{(t)}) \} \quad \text{(key inequality)}$$

$$\geq \sum_t \{ \sum_h P(h|V^{(t)}) \log P(h, V^{(t)}) - \sum_h P(h|V^{(t)}) \log P(h|V^{(t)}) \} \quad \text{because of how } \tilde{P} \text{ is chosen in the M-step.}$$

$$= \sum_t \sum_h P(h|V^{(t)}) \log \left[\frac{P(h, V^{(t)})}{P(h|V^{(t)})} \right]$$

$$= \sum_t \Big[\overbrace{\sum_h P(h|V^{(t)})}^{1} \Big] \log P(V^{(t)})$$

$$= \sum_t \log P(V^{(t)}) = \mathcal{L}_{OLD}$$

$\therefore \boxed{\mathcal{L}_{new} \geq \mathcal{L}_{OLD}}$

- M-step guarantees monotonic improvement
- Stronger guarantee than gradient ascent.

* Derivation of M-step for discrete BNs with lookup CPTs.

maximize $\sum_t \sum_h P(h \mid V^{(t)}) \log \tilde{P}(h, V^{(t)})$

$$= \sum_t \sum_h P(h \mid V^{(t)}) \log \prod_i \tilde{P}(X_i \mid pa_i) \Big|_{X=\{h, V^{(t)}\}}$$

nodes in BN

$$= \sum_t \sum_h \sum_i P(h \mid V^{(t)}) \log \tilde{P}(X_i \mid pa_i) \Big|_{X=\{h, V^{(t)}\}}$$

$$= \sum_t \sum_i \sum_\pi \sum_x P(X_i = x, pa_i = \pi \mid V^{(t)}) \log \tilde{P}(X_i = x \mid pa_i = \pi)$$

sum over possible x of node X_i

sum over possible π of parent configurations.

$$= \sum_i \sum_\pi \sum_x \underbrace{\left[\sum_t P(X_i = x, pa_i = \pi \mid V^{(t)}) \right]}_{\text{expected count}} \log \tilde{P}(X_i = x \mid pa_i = \pi) \quad : \text{regrouping}$$

Just like ML for complete data case, with expected count replacing count$(X_i = x, pa_i = \pi)$.

Solution (M-step) of EM:

$$\tilde{P}(X_i = x \mid pa_i = \pi) = \frac{\sum_t P(X_i = x, pa_i = \pi \mid V^{(t)})}{\sum_{x'} \sum_t P(X_i = x', pa_i = \pi \mid V^{(t)})}$$

Denominator simplifies to: $\sum_t P(pa_i = \pi \mid V^{(t)})$