~~Incomplete data~~

Review

How to minimize $f(\theta)$?

1) gradient descent $\vec{\theta} \leftarrow \vec{\theta} - \eta\left(\frac{\partial f}{\partial \vec{\theta}}\right)$

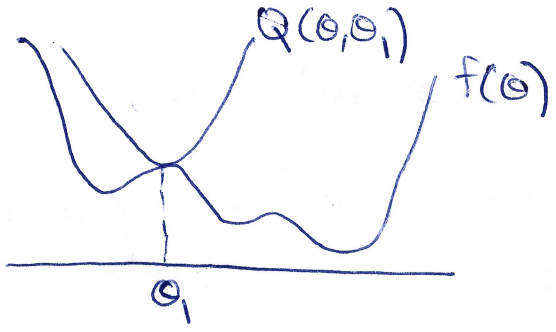
2) Newton's method $\vec{\theta} \leftarrow \vec{\theta} - H^{-1}\left(\frac{\partial f}{\partial \vec{\theta}}\right)$

3) Auxiliary fn: $Q(\vec{\theta}', \vec{\theta})$ such that

$Q(\theta, \theta) = f(\theta)$ and $Q(\theta', \theta) \geq f(\theta)$

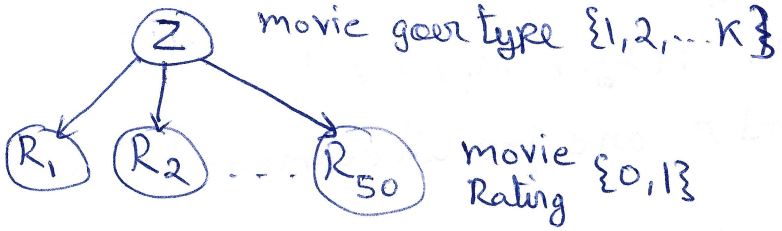
Choose $\theta_{new} = \arg\min_{\theta} Q(\theta, \theta_{old})$

Then $f(\theta_{new}) \leq f(\theta_{old})$



Incomplete data

Ex: movies recommender system



ML estimation for incomplete data

Examples $1, 2, \ldots, T$

Hidden Nodes $H^{(t)}$

Visible Nodes $V^{(t)}$

How to choose CPTs to maximize $\mathcal{L} = \sum_t \log P(V^{(t)})$?

Today: EM algorithm
- Statement & Induction
- Formal derivation.

Next Week: Many concrete examples.

Expectation Maximization (EM) algorithm

- To maximize log likelihood for incomplete data
  - Initialize CPTs to some random (non zero) values
  - Iterate until convergence.

Ⓐ E-step (Inference)

Compute posterior probabilities.

At root nodes $P(X_i = x \mid V^{(t)})$

At nodes with parents $P(X_i = x, Pa_i = \pi \mid V^{(t)})$

for all values of x and π and for all examples $t = 1, 2, \ldots, T$

Ⓑ M-step (Learning)

Root node $P(X_i = x) \leftarrow \frac{1}{T} \sum_{t=1}^{T} P(X_i = x \mid V^{(t)})$

Nodes with parents $P(X_i = x \mid Pa_i = \pi) \leftarrow \dfrac{\sum_t P(X_i = x, Pa_i = \pi \mid V^{(t)})}{\sum_t P(Pa_i = \pi \mid V^{(t)})}$

Intuition for updates

By analogy to complete data case.

Node w/ parents $P_{ML}(X_i = x / Pa_i = \pi) = \frac{count(X_i = x, Pa_i = \pi)}{count(Pa_i = \pi)}$

$$= \frac{\sum_{t=1}^{T} I(X_i^{(t)}, x)\, I(Pa_i^{(t)}, \pi)}{\sum_{t=1}^{T} I(Pa_i^{(t)}, \pi)}$$

Root nodes: $P_{ML}(X_i = x) = count(X_i = x) = \frac{1}{T} \sum_{t=1}^{T} I(X_i^{(t)}, x)$

For incomplete data, we must "fill in" missing values.

Expected statistics under $P(H / V^{(t)})$ from current CPTs substitute for observed ~~data~~ statistics when data is incomplete.

RHS of EM updates reduces to complete data counts at nodes where child & parents are observed.

Properties of EM algorithm

1) No learning rate
2) Monotonic convergence: each iteration improves $\ell = \sum_t \log P(V^{(t)})$

Derivation of EM

- Key inequality: let $P(X)$ & $\tilde{P}(X)$ be different distributions over

$$X = \{X_1, X_2, \ldots, X_n\}$$

Let V be subset of observed nodes.

$$\log \tilde{P}(V) = \log \frac{\tilde{P}(H=h,V)}{\tilde{P}(H=h|V)} \quad \text{for any setting } h \text{ of unobserved nodes.}$$

$$= \sum_h P(h|V) \log \frac{\tilde{P}(H=h,V)}{\tilde{P}(H=h|V)}$$

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

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

This is KL distance between posteriors P & $\hat{P}$

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

For ML estimation,

Imagine $P(X)$ is from BN with current (old) CPTs (like $\vec{\theta}$)

Imagine $\tilde{P}(X)$ is from BN with updated (new) CPTs (like $\vec{\theta}'$)

Formal statement of EM algorithm

E-step: Compute auxiliary function.

$$Q(\tilde{P}, P) = \sum_{t=1}^{T} \left[\sum_h P(h|V^{(t)}) \log \frac{\tilde{P}(h,V^{(t)})}{P(h|V^{(t)})} \right]$$

($\tilde{P}$: new CPTs; P: old CPTs)

To validate:

(i) $$Q(P,P) = \sum_t \sum_h P(h|V^{(t)}) \log \frac{P(h,V^{(t)})}{P(h|V^{(t)})}$$

$$= \sum_t \sum_h P(h|V^{(t)}) \log P(V^{(t)})$$

$$= \sum_t \log P(V^{(t)}) = \text{old}$$

(ii) $Q(\tilde{P}, P) \leq \sum_t \log \tilde{P}(V^{(t)}) = \ell(\tilde{P}) = \ell_{new}$
(by key inequality)

M step Maximize $Q(\tilde{P}, P)$

choose CPTs $\tilde{P}(X_i = x \mid Pa_i = \pi)$ to maximize it.

Suppose we can choose CPTs in this way

Then $\ell_{new} = \sum_t \log \tilde{P}(V^{(t)})$

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

$$\geq \sum_t \sum_h P(h \mid V^{(t)}) \log \frac{P(h, V^{(t)})}{P(h \mid V^{(t)})}$$ b/c update rule maximizes above RHS w.r.t $\hat{P}$

$$= \sum_t \sum_h P(h \mid V^{(t)}) \log P(V^{(t)})$$

$$= \sum_t \log P(V^{(t)})$$

$$= \ell_{old}.$$

Derive M-step updates

- Maximize $Q(\tilde{P}, P) = \sum_t \sum_h P(h \mid V^{(t)}) \log \tilde{P}(h, V^{(t)})$

$$- \sum_t \sum_h P(h \mid V^{(t)}) \log P(h \mid V^{(t)})$$

$$= \sum_t \sum_h P(h \mid V^{(t)}) \log \prod_{i=1}^{n} \tilde{P}(X_i = x \mid Pa_i = \pi) \Big|_{\substack{H=h \\ V=V^{(t)}}} - \cdots\cdots$$

$$= \sum_{i=1}^{n} \sum_h \sum_t P(h \mid V^{(t)}) \log \tilde{P}(X_i = x \mid Pa_i = \pi)$$

$$= \sum_{i=1}^{n} \sum_{x} \sum_{\pi} \sum_{t} P(X_i = x, pa_i = \pi / v^{(t)}) \log \tilde{P}(X_i = x | pa_i = \pi)$$

Maximize $Q(\tilde{P}, P) = \sum_i \sum_\pi \sum_x \left\{ \sum_t P(X_i = x, pa_i = \pi / v^{(t)}) \right\} \log \tilde{P}(X_i = x | pa_i = \pi)$

Recall complete data,

maximize $l_{complete} = \sum_i \sum_\pi \sum_x count(X_i = x, pa_i = \pi) \log \tilde{P}(X_i = x | pa_i = \pi)$

$$= \sum_i \sum_\pi \sum_x \left\{ \sum_{t=1}^{T} I(X_i^{(t)}, x) I(pa_i^{(t)}, \pi) \right\} \log \tilde{P}(X_i = x | pa_i = \pi)$$

$$P_{ML}(X_i = x | pa_i = \pi) = \frac{\sum_t I(X_i^{(t)}, x) I(pa_i^{(t)}, \pi)}{\sum_t I(pa_i^{(t)}, \pi)}$$

To maximize $Q(\tilde{P}, P)$

$$\tilde{P}(X_i = x | pa_i = \pi) \leftarrow \frac{\sum_t P(X_i = x, pa_i = \pi / v^{(t)})}{\sum_t P(pa_i = \pi / v^{(t)})}$$

Matches update rule for EM