Review

Lecture 7.

Approximate Inference

Query node Q
Evidence node E
How to estimate $P(Q|E)$

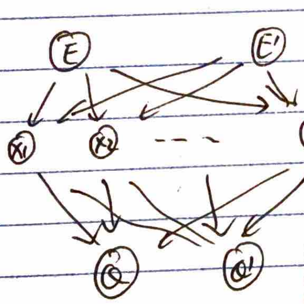
Stochastic Sampling

1) Rejection sampling — slow
2) Likelihood Weighting — faster
3) MCMC — fastest. (today)

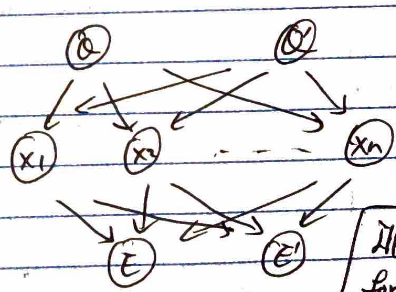
Likelihood Weighting (LW)

$$P(Q=q, Q'=q' \mid E=e, E'=e') = \frac{\sum_{i=1}^{n} I(q, q_i)\, I(q', q_i')\, P(E=e \mid pa(E))\, P(E'=e' \mid pa_i(E'))}{\sum_{i=1}^{n} P(E=e \mid pa_i(E))\, P(E'=e' \mid pa_i(E'))}$$

Converges faster then rejection sampling. But still slow in certain rare evidences.

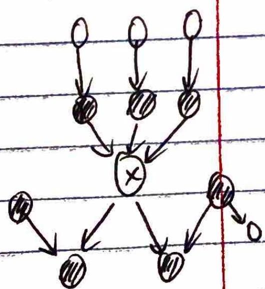


Well-suited for LW



Ill-suited for LW

Remind HW2



Def: Markov Blanket B_X of node X consists of parent/children/spouses of X.

Thm: Nodes outside of B_X are conditionally independent from X.

MCMC Simulation

Query Node Q, Q'
Evidence Node E, E'
Estimate $P(Q=q, Q'=q' \mid E=e, E'=e')$?

Procedure : - Fix evidence nodes to observed values, e, e'.
- Initialize non-evidence nodes to random values.
- Repeat N times :
  - Pick non-evidence node U at random ($U \notin E$)
  - Use Bayes Rule to compute $P(U \mid \text{all other nodes})$
    $= P(U \mid B_U)$
    where B_U is fixed to current values
  - resample U from $P(U \mid B_U)$
  - record values of nodes in BN.

* count # times $N(q, q')$ where $Q = q, Q' = q'$
* Estimate $P(Q = q, Q' = q' \mid E = e, E' = e') = \dfrac{N(q', q)}{N}$
* Converges to true value as $N \to \infty$

* Key difference between MCMC and LW.

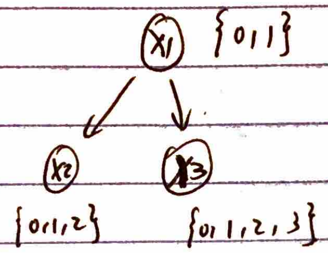
$$\left.\begin{array}{l} LW \\ MCMC \end{array}\right\} \text{sample non-evidence nodes} \left\{\begin{array}{l} \overset{\text{CPT}}{P(X \mid pa(X))} \\ \underset{\text{Bayes Rule.}}{P(X \mid B_X)} \end{array}\right.$$

$\boxed{\text{Learning}}$: * BN = DAG + CPTs not always available from expert.
How to learn from examples?
* Maximum likelihood (ML) estimation
  - simplest form of learning in BNs.
  - choose (estimate) model (DAG + CPTs) to maximize $P(\underbrace{\text{observed data} \mid \text{DAG} + \text{CPTs}}_{\text{"likelihood.}})$

Case I: known DAG structure, lookup tables for CPTs, complete data

- DAG fixed over some known finite set of discrete variables $\{X_1, X_2, \cdots X_n\}$.
- CPTs enumerate $P(X_i = x \mid Pa_i(X_i) = \pi)$ as lookup tables.
- Data is T complete instantiations of nodes in BN

Ex:



example #	X_1	X_2	X_3
1	0	1	2
⋮	⋮	⋮	⋮

(Not a CPT!).

Jargons : complete data = fully observed = no hidden nodes

More generally, denote data as: $\{x_1^{(t)}, x_2^{(t)}, \cdots x_n^{(t)}\}^T$

(dimension n, T is # of rows).

* IID assumption :
Examples are identically independently distributed from joint distribution $P(X_1, \cdots X_n)$.

* Probability of IID data set.

$$P(\text{data}) = \prod_{t=1}^{T} P(x_1^{(t)}, x_2^{(t)} \cdots x_n^{(t)}) \longrightarrow \text{IID}$$

* Probability of t^{th} example :

$$P(X_1 = x_1^{(t)}, X_2 = x_2^{(t)}, \cdots X_n = x_n^{(t)}) = \prod_{i=1} P(X_i = x_i^{(t)} \mid X_1 = x_1^{(t)}, \cdots X_{i-1} = x_{i-1}^{(t)})$$

(product rule).

$$= \prod_{i=1}^{T} P(X_i = x_i^{(t)} \mid Pa(X_i) = pa(x_i)^{(t)})$$ — and indep from DAG.

* Log-likelihood

$$\mathcal{L} = \log P(\text{data})$$

$$= \log \prod_{t=1}^{T} P(x_1^{(t)}, x_2^{(t)}, \ldots x_n^{(t)}) \quad \text{— IID}$$

$$= \log \prod_{t=1}^{T} \prod_{i=1}^{n} P(x_i^{(t)} \mid pa_i^{(t)}) \quad \text{— product rule \& CI}$$

$$= \sum_{t=1}^{T} \sum_{i=1}^{n} \log P(x_i^{(t)} \mid pa_i^{(t)}).$$

switching sum order.

$$\mathcal{L} = \sum_{i=1}^{n} \sum_{t=1}^{T} \log P(x_i^{(t)} \mid pa_i^{(t)}).$$

Let count($X_i = x_i$, $pa_i = \pi_i$) denote the # of examples where $X_i = x_i$, $pa_i = \pi_i$.

Now:
$$\mathcal{L} = \sum_{i=1}^{n} \underbrace{\sum_{x}}_{\text{possible values of } x_i} \sum_{\pi} \underbrace{\log P(X_i = x \mid pa_i = \pi)}_{\text{unknown CPTs to be optimized.}} \cdot \underbrace{\text{count}(X_i = x, pa_i = \pi)}_{\text{constants of data.}}$$

How to optimize? (Assert Solution):
$$P_{ML}(X_i = x \mid pa_i = \pi) = \cancel{\frac{\text{count}(X_i = x, pa_i = \pi)}{\text{count}(pa_i = \pi)}}$$

$$= \frac{\text{count}(X_i = x, pa_i = \pi)}{\sum_x \text{count}(X_i = x \mid pa_i = \pi)} \quad \text{(empirical frequency).}$$

Nodes w/ parents: $P_{ML}(X_i = x \mid pa_i = \pi) = \dfrac{\text{count}(X_i = x, pa_i = \pi)}{\text{count}(pa_i = \pi)}$

Root nodes: $P_{ML}(X_i = x) = \dfrac{\text{count}(X_i = x)}{T}$