Section 10: Bayes' Nets

CS 182 - Artificial Intelligence

Bayes' nets: are a technique for describing large complex joint distributions using simpler local conditional distributions and a model that relates those distributions. This is more generally called graphical models and the big ideas is that large joint distributions can be implicitly defined by a graph and a set of local conditional probability tables. This saves space as a joint distribution over N variables is size $O(2^N)$ whereas an N node Bayes net with at most k parents per node is size $O(N \cdot 2^{k+1})$.

Components:

- **Nodes**: one per variable (with domains) and can be assigned (observed) or unassigned (unobserved)
- **Bayes' Net**: A directed, acyclic graph connecting nodes which define all conditional interactions
- **Conditional Probability Table** for each node given all of the parents (see image on left below)

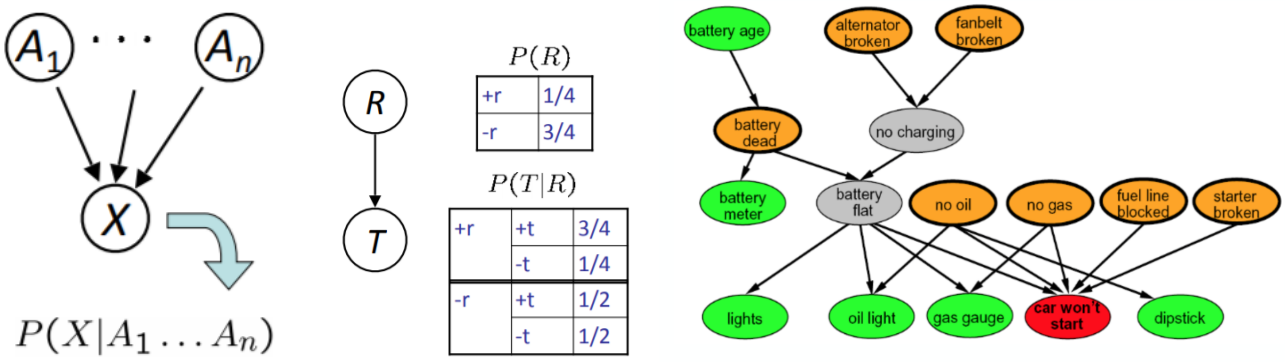


Figure 1: Conditional implications example (left), example conditional probability tables (center), and example bayes net for car repair (right)

Bayes' Nets encode conditional independence. Just like how the Markov assumption allowed us to simplify equations based on conditional independence across time, the topology of Bayes' Nets can allow us to also assume conditional independence given various pieces of evidence (observed variables). **D-Separation** is one powerful and simple procedure for determining if X_i and X_j are conditionally independent given other variables:

- Check all undirected paths between X_i and X_j
- If one or more paths active, then independence not guaranteed (a path is active if each triple is active). Otherwise (i.e. if all paths are inactive), then independence is guaranteed. See Figure 2 for the list of active and inactive triples to analyze on each path.

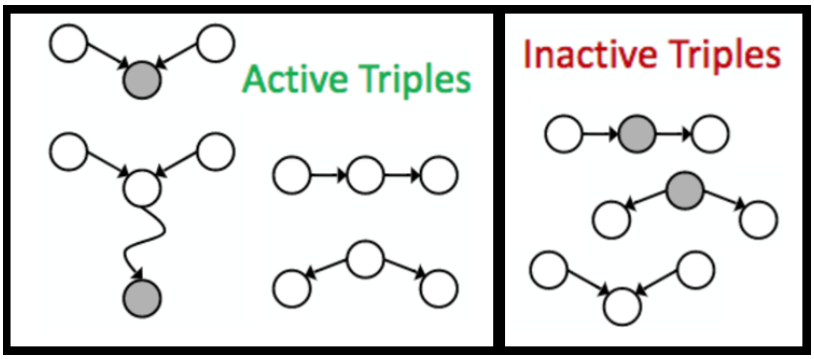


Figure 2: Active and inactive triples for D-Separation. Nodes in grey are assigned. Canonical names for active triples are common effect (left), common cause (bottom right), and causal chain (top right).

Inference is the process of calculating some resulting quantity from a joint distribution (such as a posterior probability $P(X|E_1 = e_1, E_2 = e_2)$, or most likely explanation $argmax_x\ P(X = x|E_1 = e_1, E_2 = e_2)$).

- **Enumeration** is the process of creating the full joint distribution over all of the variables and summing out any unneeded variables. For example if you wanted to determine the probability of the variable Q given evidence $E_1 = e_1 \dots E_k = e_k$ but that also related to other variables $H_1 \dots H_r$ (canonically called "hidden variables") you could:

  1. Multiply all factors to construct the full joint distribution:

  $$P(Q, H_1, \dots H_r, E_1, \dots, E_k)$$

  2. Sum out all H_i to get joint of query and evidence (selecting rows consistent with the evidence):

  $$P(Q, e_1 \cdots e_k) = \sum_{h_1 \cdots h_r} P(Q, h_1 \cdots h_r, e_1 \cdots e_k)$$

  3. Normalize across all Q:

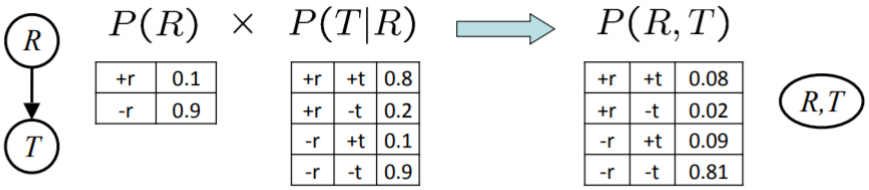
  $$P(Q = q_k|e_1 \cdots e_k) = \frac{P(Q = q_k, e_1 \cdots e_k)}{\sum_{q_i} P(Q = q_i, e_1 \cdots e_k)}$$

  This proves to be slow as you have to join up the whole joint distribution over all $Q, H, E = e$ before you sum out the hidden variables.

- **Variable Elimination** is a (usually) faster approach (although still NP-Hard) which interleaves joining and marginalizing over the hidden variables. It relies on computing **factors** which are simply CPTs where some of the variables are observed (selected, known). The process starts with the initial factors, aka the CPTs selecting the known rows just like in enumeration. However instead of forming the full joint distribution we do (see Figure 3 for a high level schematic difference):

  1. Pick a hidden variable H_k

2. Join all factors mentioning H_k by forming joint distributions:
   - Example: Join on R



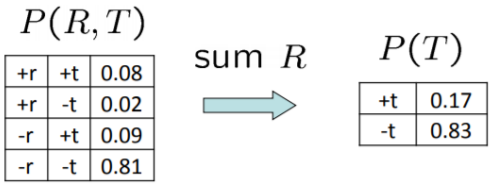
+r	0.1
-r	0.9

+r	+t	0.8
+r	-t	0.2
-r	+t	0.1
-r	-t	0.9

+r	+t	0.08
+r	-t	0.02
-r	+t	0.09
-r	-t	0.81

   - Computation for each entry: pointwise products $\forall r, t: \quad P(r,t) = P(r) \cdot P(t|r)$

3. Eliminate (sum out) H_k by computing the marginal distribution:
   - Example:



+r	+t	0.08
+r	-t	0.02
-r	+t	0.09
-r	-t	0.81

+t	0.17
-t	0.83

4. Repeat until all hidden factors are removed. Then do a final join with anything that is left over and normalize!

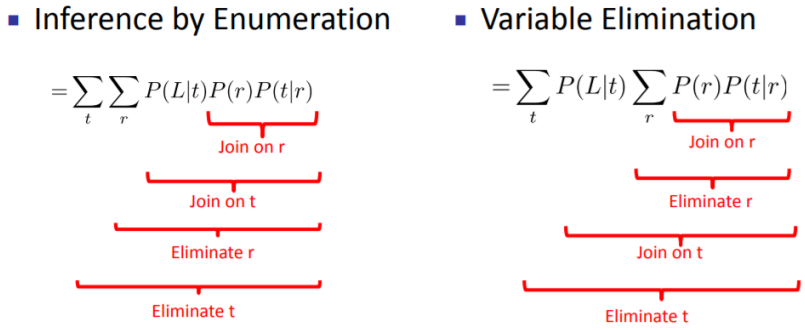


Figure 3: Summary of enumeration vs variable elimination.

Approximate Inference via Sampling: both Enumeration and Variable Elimination are methods of exact inference (aka you compute the exact answer), but that can be hard as mentioned above. Sampling can be computationally cheaper and provide a very good approximation in many cases (with theoretical guarantees as the number of samples goes to ∞). There are a variety of ways to sample:

- **Prior Sampling** is the simplest approach where one simply samples from the CPTs to get a list of samples. Then one can simply use this list to compute approximate probabilities via counting. This is very fast but does not take into account the inference problem we are trying to solve and is simply trying to sample the full joint distribution.
- **Rejection Sampling** modifies Prior Sampling by rejecting any samples that are irrelevant for our inference problem. This allows us to only have to store in memory and compute over the relevent samples. For example if we condition on $E_1 = 0$ (e.g., we are trying to compute $P(Q|E_1 = 0)$) then all samples with $E_1 \neq 0$ can be rejected. Unfortunately, this may reject many many samples.

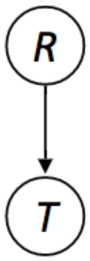
- **Likelihood Weighting** reduces the number of wasted samples by fixing the evidence variables and then sampling the rest of the distribution. Importantly (to remain consistent) it weights these samples by their likelihood of occurring. Note: evidence only effects "downstream" variables from it.

$$w(Q, E_1 = e_1) = \prod P(e_1|\text{Parents}(E))$$

- **Gibbs Sampling** improves Likelihood Weighting by constructing weights that take into account both upstream and downstream evidence through a process known as Markov Chain Monte Carlo (MCMC). You don't need to know this in detail.

Exercises:

1. Given the following Bayes' Net and CPT, calculate the joint distribution table:



$P(R)$

+r	1/4
-r	3/4

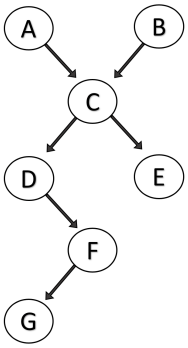
$P(T|R)$

+r	+t	3/4
	-t	1/4
-r	+t	1/2
	-t	1/2

$P(T, R)$

+r	+t	3/16
+r	-t	1/16
-r	+t	6/16
-r	-t	6/16

2. Given this Bayes' Net, answer the following questions:



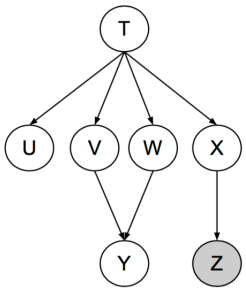
a) Are A and B conditionally independent, given D and F?
No due to common effect (long one)

b) Are A and B conditionally independent, given C?
No due to common effect (short one)

c) Are D and E conditionally independent, given C?
Yes

d) Are D and E marginally independent (aka with nothing conditioned on)?
No due to common cause

3. For this Bayes net, we are given the query $P(Y|+z)$. All variables have binary domains.



We can solve this via enumeration by computing:

$$P(Y|+z) = \sum_t \sum_u \sum_v \sum_w \sum_x P(T)P(U|T)P(V|T)P(W|T)P(X|T)P(Y|V,W)P(+z|X)$$

However, as we know this is often very slow so instead we will use variable elimination to with the following variable elimination ordering: X, T, U, V, W.

a) After inserting evidence, what are the initial factors?
$P(T), P(U|T), P(V|T), P(W|T), P(X|T), P(Y|V,W), P(+z|X)$

b) Eliminating X generates what new factor f_1?
$f_1(T, +z) = \sum_x P(x|T)P(+z|x)$

c) This leaves us with what factors?
$P(T), P(U|T), P(V|T), P(W|T), P(Y|V,W), f_1(T, +z)$

d) Eliminating T generates what new factor f_2?
$f_2(U, V, W, +z) = \sum_t P(t)P(U|t)P(V|t)P(W|t)f_1(t, +z)$

e) This leaves us with what factors?
$P(Y|V,W), f_2(U, V, W, +z)$

f) Eliminating U generates what new factor f_3?
$f_3(V, W, +z) = \sum_u f_2(u, V, W, +z)$

g) This leaves us with what factors?
$P(Y|V,W), f_3(V, W, +z)$

h) Eliminating V generates what new factor f_4?
$f_4(W, Y, +z) = \sum_v f_3(v, W, +z)P(Y|v, W)$

i) This leaves us with what factors?
$f_4(W, Y, +z)$

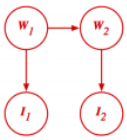
j) Finally eliminating Y generates what new factor f_5?
$f_5(Y, +z) = \sum_w f_4(w, Y, +z)$

k) How would you obtain $P(Y|+z)$ from the remaining factors?
Well we only have one factor left so simply renormalize $f_5(Y, +z)$ to obtain $P(Y|+z)$.

$$P(y|+z) = \frac{f_5(y, +z)}{\sum_{y'} f_5(y', +z)}$$

l) What is the size of the largest factor that gets generated during the above process? Does there exist a better elimination ordering (one which generates smaller largest factors)?
$f_2(U, V, W, +z)$, which has 3 binary variables, has size $2^3 = 8$. Yes, elimination ordering of X, U, T, V, W generates only factors of up to size $2^2 = 4$.

4. Suppose you are given the following Bayes net and conditional probability distributions: This is



$W_1 = \mathrm{S}$	$W_1 = \mathrm{R}$
.6	.4

(a) $P(W_1)$

	$W_2 = \mathrm{S}$	$W_2 = \mathrm{R}$
$W_1 = \mathrm{S}$	.7	.3
$W_1 = \mathrm{R}$	.5	.5

(b) $P(W_2|W_1)$

	$I = \mathrm{T}$	$I = \mathrm{F}$
$W = \mathrm{S}$	.9	.1
$W = \mathrm{R}$	.2	.8

(c) $P(I|W)$

sufficient to specify a joint probability distribution for the four variables. We want to do approximate inference through sampling. We sample and produce the following samples for (W_1, I_1, W_2, I_2):

R, F, R, F R, F, R, F S, F, S, T S, T, S, T S, T, R, F

R, F, R, T S, T, S, T S, T, S, T S, T, R, F R, F, S, T

a) What is $\hat{P}(W_2 = R)$, the probability that sampling assigns to the event $W_2 = R$?
$W_2 = R$ in 5 of the 10 samples, therefore the probability is 0.5.

b) Cross of samples rejected by rejection sampling if we are computing $P(W_2|I_1 = T, I_2 = F)$

~~R, F, R, F~~ ~~R, F, R, F~~ ~~S, F, S, T~~ ~~S, T, S, T~~ S, T, R, F

~~R, F, R, T~~ ~~S, T, S, T~~ ~~S, T, S, T~~ S, T, R, F ~~R, F, S, T~~

Rejection sampling seems to be wasting a lot of effort, so we decide to use likelihood weighting instead. Assume we generate the following six samples given the evidence $I_1 = T$ and $I_2 = F$:

$$(W_1, I_1, W_2, I_2) = \Big\{\langle \mathrm{S}, \mathrm{T}, \mathrm{R}, \mathrm{F}\rangle, \langle \mathrm{R}, \mathrm{T}, \mathrm{R}, \mathrm{F}\rangle, \langle \mathrm{S}, \mathrm{T}, \mathrm{R}, \mathrm{F}\rangle, \langle \mathrm{S}, \mathrm{T}, \mathrm{S}, \mathrm{F}\rangle, \langle \mathrm{S}, \mathrm{T}, \mathrm{S}, \mathrm{F}\rangle, \langle \mathrm{R}, \mathrm{T}, \mathrm{S}, \mathrm{F}\rangle\Big\}$$

Recall that in likelihood weighting, we give weight to a sample according to:

$$\prod_{\text{Evidence variables } e} P(e|\text{Parents}(e))$$

c) What is the weight of the first sample (S, T, R, F)?
The evidence is $I_1 = T, I_2 = F$. The weight of the first sample is therefore:

$$w = P(I_1 = T|W_1 = S) \cdot P(I_2 = F|W_2 = R) = 0.9 \cdot 0.8 = 0.72$$

d) Use likelihood weighting to estimate $\hat{P}(W_2|I_1 = T, I_2 = F$
The sample weights are shown in in the following table:

(W_1, I_1, W_2, I_2)	w	(W_1, I_1, W_2, I_2)	w
S, T, R, F	0.72	S, T, S, F	0.09
R, T, R, F	0.16	S, T, S, F	0.09
S, T, R, F	0.72	R, T, S, F	0.02

To compute the probabilities, we normalize the weights and find that:

$$\hat{P}(W_2 = R|I_1 = T, I_2 = F) = \frac{0.72 + 0.16 + 0.72}{0.72 + 0.16 + 0.72 + 0.09 + 0.09 + 0.02} = 0.889$$
$$\hat{P}(W_2 = S|I_1 = T, I_2 = F) = 1 - 0.889 = 0.111$$