Notes: Introduction to Bayesian Networks

Zheng Rui

November 4, 2014

Lecture 1: Probability

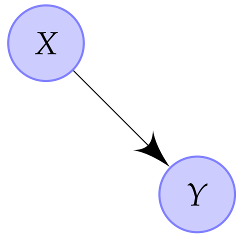
Lecture 2: Concepts of BN

- To specify a joint probability $P(X_1, X_2, ..., X_n)$, it needs at least $2^n - 1$ numbers. Exponential model size.
- Chain rule:
$$P(X_1, X_2, ..., X_n) = \prod_{i=1}^{n} P(X_i | X_1, ..., X_{i-1})$$
from this perspective, the number of parameters required for the knowledge of $P(X_1, X_2, ..., X_n)$ is also
$$1 + ... + 2^{n-1} = 2^n - 1$$
why?
$$P(\overline{X_i} | X_1, ..., X_{i-1}) = 1 - P(X_i | X_1, ..., X_{i-1})$$
when $X_i, ..., X_{i-1}$ are fixed, and there are 2^{i-1} possible combination of them.
- Define $pa(X_i)$ as the X_i relevant subset $pa(X_i) \subseteq \{X_1, ..., X_{i-1}\}$ such that
$$P(X_i | X_1, ..., X_{i-1}) = P(X_i | pa(X_i))$$
then
$$P(X_1, X_2, ..., X_n) = \prod_{i=1}^{n} P(X_i | pa(X_i))$$
in this way the number of parameters might be substantially reduced.
- Bayesian network: DAG, each node represents a random variable, and is associated with the conditional probability of the node given its parents, arcs represent direct probabilistic dependence.A BN represents a factorization of a joint distribution. CPT means conditional probability table, multiplying them together gives a joint distribution.
- Causal Markov Assumption: a variable is independent of all its non-effects (non-descendants) given its direct causes (i.e. parents).
- Causal independence and Context specific independence.

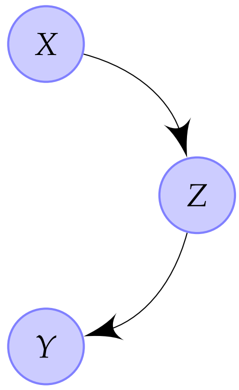
Lecture 3: D Separation

- Cases:

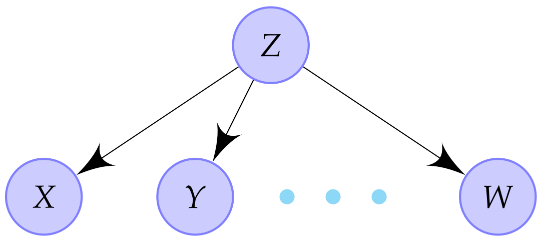
1. Direction connection: if X and Y are connected by an edge, then X and Y are dependent.



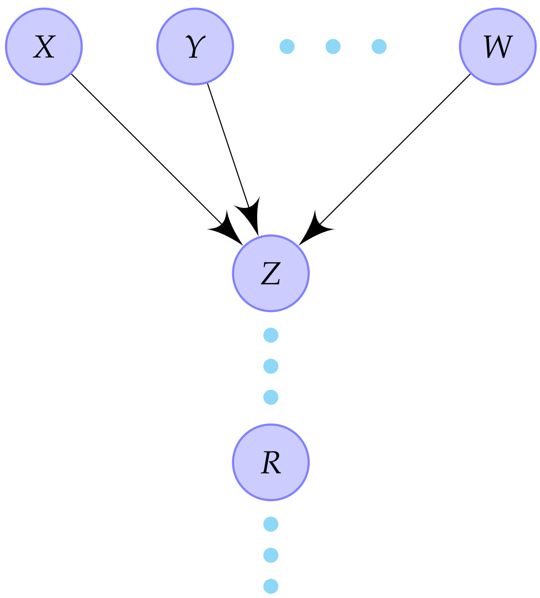
2. Serial connection: Z observation makes X and Y become **independent** from **dependent**.



3. Diverging connection: Z observation makes all its children $X, Y, \cdots, W$ become **independent** from **dependent**.

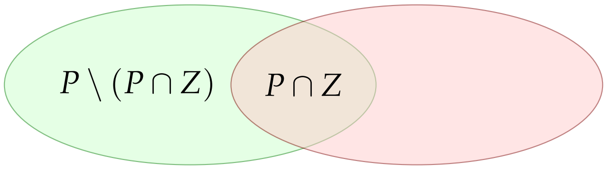


4. Converging connection: observation of Z or any of its descendants R makes $X, Y, \cdots, W$ become **dependent** from **independent**.



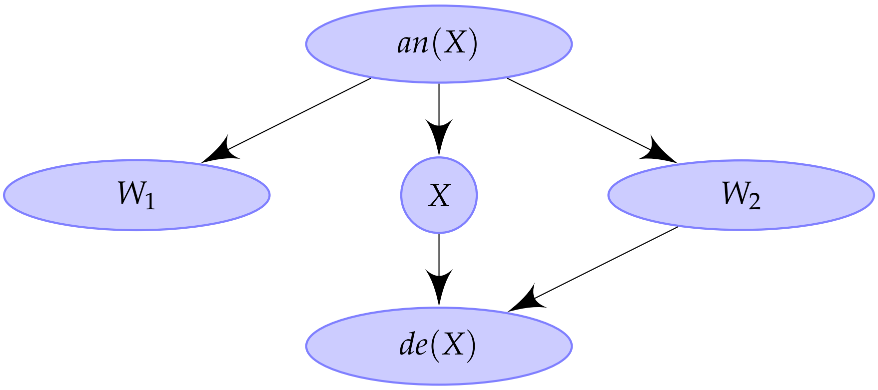
- Rules: Hard evidence **blocks** info-path for serial and diverging connection; Soft evidence **opens** info-path for converging connection.
- A path between X and Y is blocked by a nodes set Z if: Either one node in Z is in the path and the connection of that node is serial or diverging case. Or the path contains a converging node *s.j.t* this node and all its descendants are not in set Z.

  So when asking if a path P is blocked by a nodes set Z?

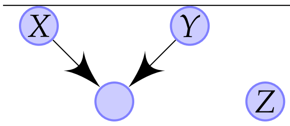
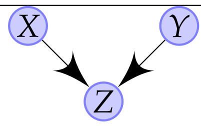
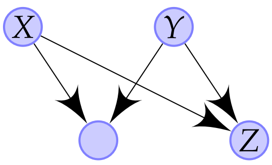
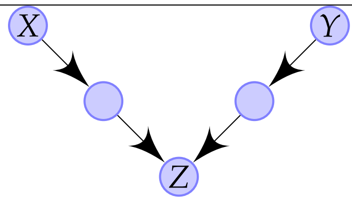
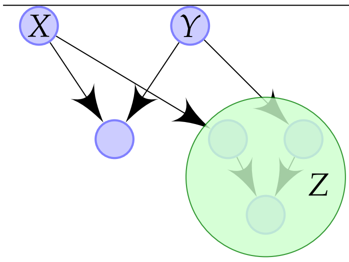
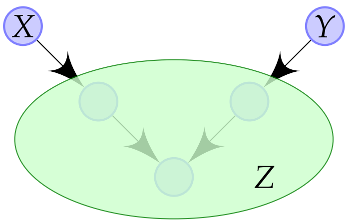
  First check if there are serial or diverging node in $P \cap Z$, if not then check if there are converging node in $P \setminus (P \cap Z)$ and none of the converging node's descendants is in Z.

  Bascially for a DAG, the other nodes with respect to node X fall into 4 groups:

- D-separation: Two nodes X and Y are d-separated by a set Z if all paths between X and Y are blocked by Z, $X \perp Y|Z$.

  Examples:

Bayesian Networks	Z separate X and Y ?	Bayesian Networks	Z separate X and Y ?
	$\surd$		$\times$
	$\surd$		$\times$
	$\surd$		$\surd$

Things to note in the examples: *X, Y may be dependent or independent, but $X|Z$, $Y|Z$ can be independent so long as they share descendants that are $\notin Z \cup an(Z)$, that's how Z separates X and Y.*

- ancestral set $an(X)$ of nodes set X, X is ancestral if

$$X = an(X)$$

- $P_N(X) = P_{N'}(X)$ where $N' = N \setminus Y$, Y is a leaf node of N.
- $P_N(X) = P_{N'}(X)$ where $N' = X$, X is ancestral.
- Suppose X, Y, Z are disjoint sets, $X \cup Y \cup Z$ is all the nodes, then Z separates X and Y leads to $X \perp Y|Z$, key is there is no converging node in Z which has parents from both X and Y, otherwise the length-2 path with 1 parent from X and the other from Y is not separated by Z.
- Global Markov property: variables X and Y, $X \perp Y|Z$ if $X \notin Z$, $Y \notin Z$, and Z separates them.

$$\mathcal{S}_{\mathcal{G}}(X, Y, Z) \Rightarrow X \perp_P Y|Z$$

- Markov blanket: (parents + children + parents of children) of node X.
- Local Markov property: given parents, variable X is independent of all its non-descendants.

$$X \perp_P nd_{\mathcal{G}}(X)|pa_{\mathcal{G}}(X)$$

- $\mathcal{G}$ to $P(V)$ is called **I-map**: $\mathcal{S}_{\mathcal{G}}(X, Y, Z) \Rightarrow X \perp_P Y|Z$, **D-map**: $X \perp_P Y|Z \Rightarrow \mathcal{S}_{\mathcal{G}}(X, Y, Z)$, **Perfect map**: $\mathcal{S}_{\mathcal{G}}(X, Y, Z) \Leftrightarrow X \perp_P Y|Z$.

Lecture 4: Inference in BN & VE Algorithm

- Diagnostic inference: effects -> causes; Predictive/Causal inference: causes -> effects; Inter-causal inference (explaining away): between causes of a common effect; Mixed inference: combing two or more of the above.
- A naive inference algorithm is like:

$$\begin{aligned} P(Q, E) &= \sum_{X \notin Q \cup E} P(X) \\ P(E) &= \sum_{Q} P(Q, E) \\ P(Q|E = e) &= \frac{P(Q, E = e)}{P(E = e)} \end{aligned}$$

  exponential complexity in this naive way, not making use of factorization

- A factorization of a joint distribution is a list of functions whose product is the joint distribution, functions on the list are called factors.

$$P(X_1, X_2, ..., X_n) = \prod_{i=1}^{n} P(X_i|pa(X_i))$$

  $P(X_i|pa(X_i))$ are factors.

- Elimination a variable Z_1 from $P(Z_1, Z_2, ..., Z_m)$: suppose $\mathcal{F} = \{f_1, f_2, ..., f_n\}$ is its factorization, and Z_1 appears in and only in factors $f_1, f_2, ..., f_k$, then

$$P(Z_2, ..., Z_m) = [\prod_{i=k+1}^{n} f_i][\sum_{Z_1} \prod_{i=1}^{k} f_i] = [\prod_{i=k+1}^{n} f_i]h$$

  and its new factorization after Elimination of Z_1 is $\{f_{k+1}, ..., f_n, h\}$

- Variable Elimination Algorithm: $VE(\mathcal{F}, Q, E, e, \rho)$, with $\mathcal{F}$ factors, Q query variables, E observered variables and e are their observed values, $\rho \notin Q \cup E$ is the ordering of variables to be eliminated,

  ❶ While ρ is not empty
  - ➥ remove the first variable in ρ
  - ➥ call procedure of eliminating a single variable

  ❷ set $h = \prod_{f \in \mathcal{F}} f$, this is the factorization of joint probability $P(Q, E)$

  ❸ set $E = e$, instantiate E to observered values e

  ❹ re-normalization $P(Q|E = e) = \frac{h(Q)}{\sum_Q h(Q)}$

  a modification is put ❸ in front of ❶, this more efficient version was proposed by Zhang and Poole (1994).

- Structural graph, moral graph and cost of elimination variables: