Probability Reasoning

(Stuart Russell and Peter Norvig, 2003,
Artificial Intelligence: A Modern Approach, Chap 14)

This book chapter introduces the definition of the Bayesian Network, the approaches to construct the Bayesian Network, and the approaches to reason in the Bayesian Network.

The basic content which is been introduced is:

1. The definition of the Bayesian Network.
   A Bayesian Network is a data structure to represent the dependencies among variables and to give a concise and compact representation of any joint probability distribution.
2. The syntax of the Bayesian Network.
   a. The topology of the network- the set of nodes and links- specifies the conditional independence relationships.
   b. Each node X has a conditional probability distribution $P(X_i \mid Parents(X_i))$ that quantifies the effect of the parent's nodes on the node.
3. The general and local semantics of the Bayesian Network.
   a. General semantics – Bayesian Network represents the joint distribution of all variables. $P(x_1, ..., x_n) = \prod_{i=1}^{n} P(x_i \mid parents(x_i))$;
   b. Local semantics – Bayesian Network represents the conditional independence relation among the nodes.

      Note: "Descendant" and "Markov blanket" are being used to describe the conditional independence relation among the nodes in the Bayesian Network.
      While using "Descendants", a node is conditional independent of its non-descendants, given its parents.
      While using "Markov blanket", a node is conditionally independent of all other nodes in the network, given its Markov blanket (parents, children, and children's parents).
4. Compactness of Bayesian Network to represent the join probability.
   A Bayesian Network can often be more compact then the joint probability distribution. For example, for *n* Boolean variables, the joint distribution contains 2^n numbers. In the case of Bayesian Networks, if we assume that each random variable is directly influenced by at most *k* others (it is normally reasonable to assume in this way), the complete network can be specified by $n2^k$ numbers.
5. Several approaches to simplify the representation of Conditional Probability Table.

a. Canonical distribution – the complete table can be simplified by identifying known patterns, *e.g.* deterministic nodes, noisy-or, continuous variables with known form of the distribution…

b. Hybrid Bayesian Network…

   (1) Continuous variable with continuous variable as the parents, e.g. Linear Gaussian.

   (2) Continue variable with discrete variable as the parents, e.g. conditional Gaussian.

   (3) Discrete variable with continuous variable as the parents, e.g. probit distribution, which uses the Gaussian function, or logit function, which uses the sigmoid function.

6. Approach to construct Bayesian Network.

   a. Joint probability distribution can be written in term of conditional probability distribution. $P(x_1,...,x_n)=\prod_{i=1}^{n}P(x_i \mid x_{i-1},...,x_1)$

   b. Label the node in any order to make $P(X_i \mid X_{i-1},...,X_1)=P(X_i \mid Parent(X_i))$ with constraint $Parents(X_i)\subseteq\{X_{i-1},...,X_1\}$ true.

   Note: Depending on the ordering of the variables, the constructed Bayesian Network is not unique. To get the simplest Bayesian Network, we have to add the "root" causes first, then the variables they influence. In another word, we build a causal model instead of a diagnostic model.

7. Inference in Bayesian Network.

   a. Inference in Bayesian Network means to compute the posterior probability distribution *P(X|e)* for a set of query *X*, given some observed assignment of values *e* to a set of evidence variable *E*.

   b. Exact Inference in Bayesian Network.

      (1) Enumeration

      $$P(X \mid e)=\alpha\, P(X,e)=\alpha \sum_y P(X,e,y)$$

      (2) Variable elimination.

      The idea is simply to store the intermediate results to simplify the computation.

      The time and space complexity of variable elimination is linear with the size of the network in polytrees. For the multiply connected networks, it has the exponential time and space complexity in the worst case.

   c. Approximate inference in Bayesian Network.

      (1) Approximate inference in Bayesian Network can be achieved by:

         i. Using Monte Carlo approaches to generate the events (samples) from the network.

ii. Estimate $\hat{P}(X = x \mid e)$ by counting how often X=x occurs in the events.
Note: The Monte Carlo approaches need to use consistent sampling, which guarantee that the frequency of the counted samples yields the true probability in the large-sample limit.

(2) Several randomized sampling (Monte Carlo) approaches

i. *Direct sampling*.
Direct sampling generates events from a Bayesian Network by sampling each variable in topological order without consideration of the evidence. The probability distribution from which the value is sampled is conditional on the values already assigned to the variable's parents.

ii. *Rejection sampling*.
It uses the same procedure as direct sampling to produce samples. Then it rejects all those samples that do not match the evidence. The drawback of this approach is that it rejects too many samples.

iii. *Likelihood weighting*.
This approach fixes the values for the evidence variable *E* and samples only the query variables *X* and hidden variables *Y*.

iv. *Markov Chain Monte Carlo* (MCMC).
MCMC generates each event by randomly sampling a value for one of the non-evidence variable X_i, conditioned on the previously generated values of the variables in the Markov blanket of X_i.

Question for discussion:

1. By giving the variables' conditional independence and causal information, could we build a unique Bayesian Network? Is this Bayesian Network topologically the simplest Bayesian Network? Could we find a counterexample?

2. In some cases, we can construct a unique Bayesian Network only from the variables' conditional independence, in another words, we can learn the Bayesian Network structure from the data. Under what condition is this true?