**208891 Probabilistic Graphical Models** **Spring 2019**

# Lecture 3: December 27

*Lecturer: Donlapark Pornnopparath*

**Example 3.1.** Suppose that a company want to hire a recent college graduate. Before any major decision, the company wants to measure the student's intelligence based on their performance on an exam. So they come up with a Bayesian network shown in Figure 3.1 which consists of five variables related to the performance of a student on an exam: the difficult of the class $D$, the student's intelligence $I$, their letter grade $G$, SAT score $S$ and the strength of the letter of recommendation $L$. For simplicity, we assume that there are only three grades: $A$, $B$ and $C$, represented by $g^1$ $g^2$ and $g^3$, respectively, and the rest of the variables are binary. For instance, a student can have either high intelligence $i^1$ or low intelligence $i^0$.

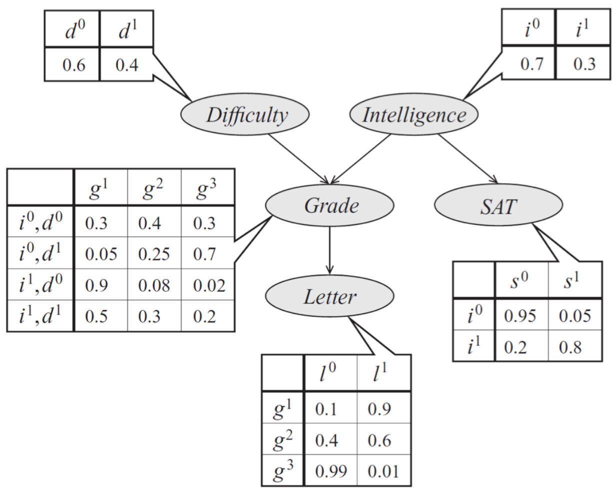


Figure 3.1: A Bayesian network of a student's performance on an exam.

**Factorization** The factorization of the joint probability distribution follows the top-down approach:

$$\mathbb{P}(I, D, G, S, L) = \mathbb{P}(I)\mathbb{P}(D)\mathbb{P}(G|I, D)\mathbb{P}(S|I)\mathbb{P}(L|G).$$

For example, the probability that a smart student with a high SAT score and a strong letter of recommendation got a $B$ in an easy class is

$$\mathbb{P}(i^1, d^0, g^2, s^1, l^1) = 0.3 \times 0.6 \times 0.08 \times 0.8 \times 0.6 = 0.006912. \tag{3.1}$$

Note that there is a higher chance that this student got an $A$ in this class, as can be seen from replacing 0.08 by $\mathbb{P}(g^1|i^1, d^0) = 0.9$.

**Independence** As in the definition of a Bayesian network, the graph imposes the following conditional independence:

- $I \perp\!\!\!\perp D$.
  It is reasonable to assume that $I$ is independent of $D$ but not other variables since they are all descendants of $I$.

- $D \perp\!\!\!\perp I, S$.
  Since $S$ is not a descendant of $D$, it is also independent of $D$. Certainly, there should be no relationship between the difficulty of the exam and the SAT score.

- $S \perp\!\!\!\perp D, G, L | I$.
  Since the SAT score only depends only on the student's intelligence and does not have any descendant, it must be independent of any other variables.

- $G \perp\!\!\!\perp S | I, D$.
  The strength of the letter of recommendation can be explained directly from the grade, so $G$ is not independent of $L$. On the other hand, given the student's intelligence, the SAT score provides no additional information about the student's grade and vice versa.

- $L \perp\!\!\!\perp I, S, D | G$
  Lastly, since the strength of the letter of recommendation can be fully explained by the letter grade, we will not gain any additional information about $L$ from any other variables.

◇

## 3.1 Independencies from a Bayesian network

Note that dependency condition $X \not\perp\!\!\!\perp Y | Z$ in a little V graph $X \longrightarrow Z \longleftarrow$ also holds when $Z$ is an ancestor of an observed variable. To see this, we look at the subgraph **??** from Example 3.1. The graph tells us that a student who had a good recommendation letter was more likely to receive a good grade. As a consequence, if the exam was difficult, there is a high chance that the student had high intelligence.

We can extend this and the other two structures to obtain independencies between any two sets of variables $\boldsymbol{X}$ and $\boldsymbol{Y}$ given a group of observed variables $\boldsymbol{O}$. First, we use dependencies in these graphs to characterize a path that carries along the dependency of its source node, then introduce a notion of $d$-separated ($d$ stands for directed) between two sets of nodes that are not connected by an active path.

**Definition 3.2.** An undirected path in $G$ is an *active path* given a set of observed variables $\boldsymbol{O}$ if for any consecutive triplet of variables $X, Y, Z$, one of the following conditions holds:

- $X \longrightarrow Y \longrightarrow Z$ and $Y \notin \boldsymbol{O}$
- $X \longleftarrow Y \longleftarrow Z$ and $Y \notin \boldsymbol{O}$
- $X \longleftarrow Y \longrightarrow Z$ and $Y \notin \boldsymbol{O}$
- $X \longrightarrow Y \longleftarrow Z$ and $Y$ or any of its descendants is in $\boldsymbol{O}$.

Two sets of variables $\boldsymbol{X}$ and $\boldsymbol{Y}$ are $d$-separated given $\boldsymbol{O}$ if they are not connected by an active path give $\boldsymbol{O}$.

This leads us to the *Bayes ball algorithm*, whose term was coined from an action of rolling a ball until it is *blocked* under a certain condition. We start at a node in $X$ and mark all nodes in $\boldsymbol{O}$ and their ancestors. Then we travel along a path until it is blocked by one of the two following conditions:

- The next node is in the middle of a little V structure which is not marked.
- The next node is in the middle of one of the other structures which is in $\boldsymbol{O}$.

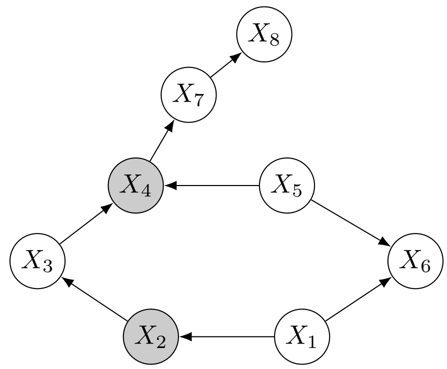


Figure 3.2: In this graph, $(X_3, X_4, X_5, X_6)$ is an active path.

**Example 3.3.** In the graph in Figure 3.2, we have the following

$$\begin{aligned} &X_3 \perp\!\!\!\perp X_1 | X_2, X_4 \\ &X_3 \not\perp\!\!\!\perp X_6 | X_2, X_4 \\ &X_3 \not\perp\!\!\!\perp X_6 | X_8. \end{aligned}$$

The second dependency holds because the path $X_3 \longrightarrow X_4 \longrightarrow X_5 \longrightarrow X_6$ is active, and the third dependency follows similarly since $X_4$, which is a middle node of a V structure, is an ancestor of $X_8$. ◇

## 3.2 Representation of a Bayesian network

Since we have been trying to reconstruct a probability distribution using a graphical model, this raises a concern of whether *any* distributions can be modeled using a Bayesian network. Unfortunately, this is not the case, as the following example shows:

**Example 3.4.** Consider three random variables $X, Y$ and $Z$ where $X, Y \sim \text{Ber}(0.5)$ are independent and $Z = X$ xor $Y$ i.e. $Z = 1$ if $X = Y$ and $Z = 0$ otherwise. Then by symmetry we have

$$\begin{aligned} \mathbb{P}(Z) &= \mathbb{P}(Z|X=0)\mathbb{P}(X=0) + \mathbb{P}(Z|X=1)\mathbb{P}(X=1) \\ &= \frac{1}{2}\mathbb{P}(Z|X=0) + \frac{1}{2}\mathbb{P}(Z|X=1) \\ &= \mathbb{P}(Z|X=0), \end{aligned}$$

which implies $\mathbb{P}(Z) = \mathbb{P}(Z|X=1)$, and so $X \perp\!\!\!\perp Z$. Similarly, $Y \perp\!\!\!\perp Z$. However, $X, Y \not\perp\!\!\!\perp Z$ since

$$\begin{aligned} \mathbb{P}(Z) &= \sum_{x,y} \mathbb{P}(Z|X=x, Y=y)\mathbb{P}(X=x, Y=y) \\ &= \mathbb{P}(Z|X=0, Y=0)\mathbb{P}(X=0, Y=0) + \mathbb{P}(Z|X=1, Y=1)\mathbb{P}(X=1, Y=1) \\ &= \frac{1}{2}, \end{aligned}$$

and $\mathbb{P}(Z|X,Y)$ is either 0 or 1. We can check and see that there is no Bayesian network that satisfies these conditions at the same time. $\diamond$

Thus we might want to relax the task of finding a *perfect map* for a distribution $p$ to finding a map (i.e. a graph) all of whose independencies hold in $p$. Suppose that there are $n$ variables, then we can start with the complete graph $K_n$. Then we might try to remove edges while preserving independencies in $p$ until the graph is "minimal", meaning that further edge removal would result in independencies that do not hold in $p$. This can be done precisely by first sorting the node in order:

**Definition 3.5.** A topological ordering of a DAG $G$ is an order of nodes $X_1, X_2, \ldots, X_n$ such that for any directed edge $X_i \longrightarrow X_j$ we have $i < j$.

Note that a topological ordering always exists for any DAG. Then a minimal map can be made as follows: for each step $i = 1, 2, \ldots, n$, we do

- For any $i$, find the smallest $U_i \subseteq \{X_1, \ldots, X_{i-1}\}$ such that $X_i \perp\!\!\!\perp \{X_1, \ldots, X_{i-1}\} - U_i | U_i$
- Let all nodes in $U_i$ be the parents of $X_i$.