Probabilistic Graphical Models

William Wong

October 19, 2017

1 Probabilistic Graphical Models 1: Representation

1.1 Week 1. Introduction and Overview

Please see the student example in Fig. ??.

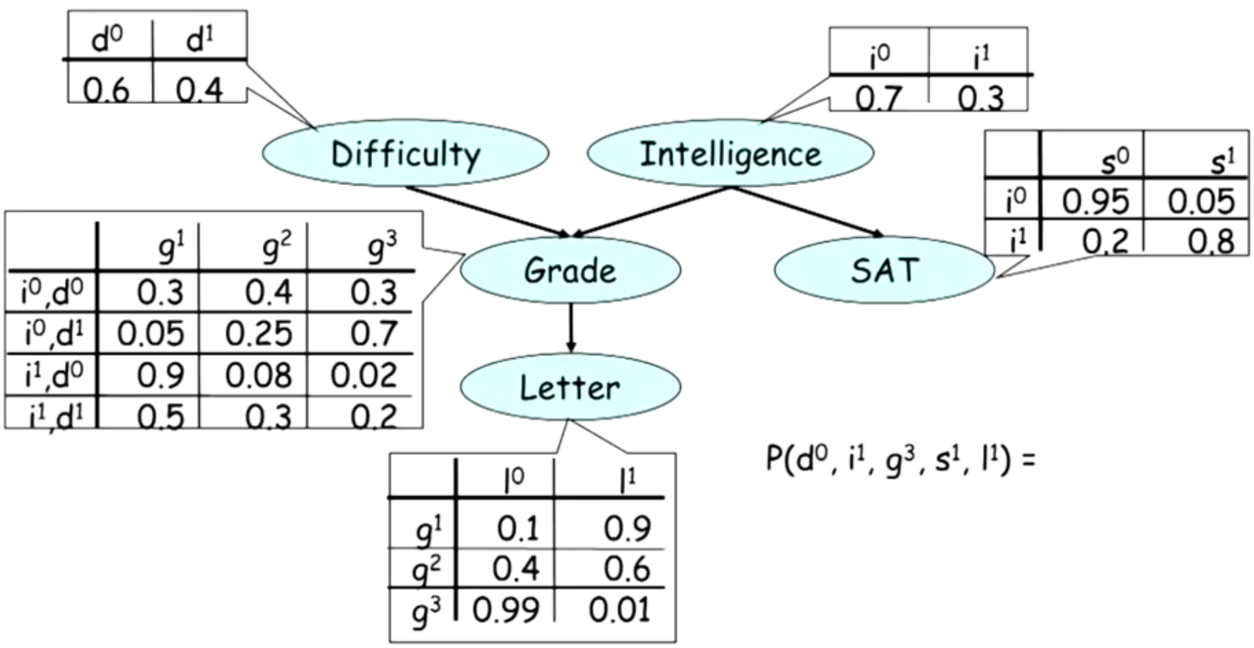

Figure 1: Student Example. Each node is annotated with a conditional probability distribution (CPD).

- Grade (G) — $g^1(A)$, $g^2(B)$, and $g^3(C)$.
- Course Difficulty (D) — d^0(easy), d^1(hard).
- Student Intelligence (I) — i^0(low) and i^1(high).
- SAT score (S).
- Reference Letter (L).

1.1.1 Joint Distribution

Consider $\Pr(I, D, G)$.

1.1.2 Conditioning

For example, we can condition on $G = g^1$. We look at $\Pr(I, D, g^1)$. We need to renormalize. That is

$$\Pr(I, D|g^1) = \frac{\Pr(I, D, g^1)}{\sum_{I=i^j} \sum_{D=d^k} \Pr(I, D, g^1)}.$$

The above equation comes from the definition of conditional probability $\Pr(A|B) = \Pr(A, B)/\Pr(B)$.

1.1.3 Marginalization

For example, we marginalize I.

$$\Pr(D) = \sum_{I=i^j} \Pr(I, D)$$

1.1.4 Factors

A factor ϕ maps random variables to a real number. That is

$$\phi : (X_1, \ldots, X_k) \to \Re$$

A joint distribution $\Pr(I, D, G)$ is a factor.

We compute a factor product using

$$\phi(A = a^i, B = b^j, C = c^k) = \phi_1(A = a^i, B = b^j)\phi_2(B = b^j, C = c^k).$$

We can marginalize a factor too.

$$\phi(A = a^i, C = c^k) = \sum_j \phi(A = a^i, B = b^j, C = c^k).$$

1.2 Week 1. Bayesian Network (Directed Models)

A Bayesian network is a directed acyclic graph (DAG) whose nodes represent the random variables $X_1, \ldots, X_n$.

1.2.1 Factorization

Chain rule:

$$\Pr(D, I, G, S, L) = \Pr(D) \Pr(I) \Pr(G|I, D) \Pr(S|I) \Pr(L|G) \tag{1}$$

$$\Pr(X_1, ..., X_n) = \prod_{i=1}^{n} \Pr(X_i|\mathrm{Par}_{\mathbb{G}}(X_i)), \tag{2}$$

where $\mathrm{Par}_{\mathbb{G}}(X_i)$ are the parents of X_i over the graph $\mathbb{G}$.

1.2.2 Reasoning Patterns

Causal reasoning $\Pr(L = l^1 | I = i^0) = 0.39$.

Given an easier class, $\Pr(L = l^1 | I = i^0, D = d^0) = 0.51$.

Evidential reasoning (from the bottom up) $\Pr(D = d^1 | G = g^3) = 0.63$.

Pay attention to the details in the calculations. We must use Bayes' theorem.

$$\begin{aligned}
\Pr(d^1|g^3) &= \frac{\Pr(d^1, g^3)}{\Pr(g^3)} \\
&= \frac{\Pr(d^1, g^3, i^0) + \Pr(d^1, g^3, i^1)}{\Pr(g^3|i^0, d^0)\Pr(i^0)\Pr(d^0) + \Pr(g^3|i^0, d^1)\Pr(i^0)\Pr(d^1) + \ldots} \\
&= \frac{\Pr(g^3|i^0, d^1)\Pr(i^0)\Pr(d^1) + \Pr(g^3|i^1, d^1)\Pr(i^1)\Pr(d^1)}{\Pr(g^3|i^0, d^0)\Pr(i^0)\Pr(d^0) + \Pr(g^3|i^0, d^1)\Pr(i^0)\Pr(d^1) + \ldots}.
\end{aligned}$$

Intercausal reasoning The probability that the student is highly intelligent is $\Pr(I = i^1) = 0.3$.

Given that the student got a B, $\Pr(I = i^1 | G = g^2) = 0.175$.

Given that he/she got a B and the class was difficult, $\Pr(I = i^1 | G = g^2, D = d^1) = 0.34$.

As another example, let $Y = (X_1 \text{or} X_2)$.

$\Pr(X_2 = 1 | Y = 1) = 2/3$, but $\Pr(X_2 = 1 | Y = 1, X_1 = 1) = 1/2$.

1.2.3 Flow of Probabilistic Influence

When can X influence Y? In other words, does conditioning on X change our beliefs about Y?

- $X \to Y$. X is the parent of Y.
- $X \leftarrow Y$. X is the child of Y.
- $X \to W \to Y$.
- $X \leftarrow W \leftarrow Y$.
- $X \leftarrow W \to Y$. W is the common parent.
- $\cancel{X \to W \leftarrow Y}$. W is the common child.

When can X influence Y given evidence about Z?

Scenario	Can X influence Y given evidence about Z?	
	$W \notin Z$	$W \in Z$
$X \to Y$	Yes	
$X \leftarrow Y$	Yes	
$X \to W \to Y$	Yes	No
$X \leftarrow W \leftarrow Y$	Yes	No
$X \leftarrow W \to Y$	Yes	No
$X \to W \leftarrow Y$	No if W and all its descendants $\notin Z$	Yes if W or one of its descendants $\in Z$

Active Trails A trail $X_1 - \ldots - X_n$ is active given Z if

1. for any v-structure $X_{i-1} \to X_i \leftarrow X_{i+1}$ we have X_i or one of its descendants in Z, and
2. no other X_i is in Z.

1.2.4 Independence

For events α and β, $\Pr \vdash \alpha \perp \beta$ if

- $\Pr(\alpha, \beta) = \Pr(\alpha)\Pr(\beta)$, or
- $\Pr(\alpha|\beta) = \Pr(\alpha)$, or
- $\Pr(\beta|\alpha) = \Pr(\beta)$

From Wikipedia, $P \vdash Q$ means "from P, I know that Q".

1.2.5 Conditional independence

For sets of random variables X, Y, and Z, $\Pr \vdash (X \perp Y|Z)$ if

- $\Pr(X, Y|Z) = \Pr(X|Z)\Pr(Y|Z)$, or
- $\Pr(X|Y, Z) = \Pr(X|Z)$, or
- $\Pr(Y|X, Z) = \Pr(Y|Z)$, or
- $\Pr(X, Y, Z) \propto \phi_1(X, Z)\phi_2(Y, Z)$

Example 1 Tossing a coin twice, which might be fair or not. The two outcomes (from the two tosses) are X_1 and X_2.

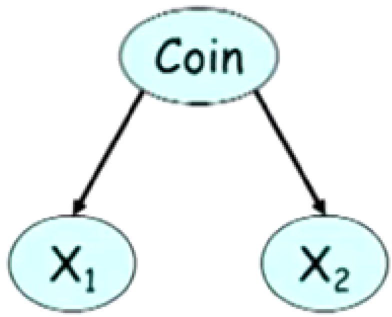

Figure 2: A fair or biased coin is tossed twice. X_1 and X_2 are the two outcomes, respectively.

$\Pr(X_2 = \text{head}|X_1 = \text{head}) > 0.5$ because the coin might not be fair. However, $\Pr(X_2 = \text{head}|C = \text{fair coin})$ is independent of X_1!

Also $\Pr(X_2 = \text{head}|C = \text{biased coin})$ is independent of X_1. That is, once we know what the coin is, the two outcomes are no longer correlated.

In her notation,

- Pr does NOT satisfy $X_1 \perp X_2$
- Pr satisfies $(X_1 \perp X_2 | C)$.

Conditioning can remove independence.

Example 2 We have,

$$I \to G, S$$

$$\Pr(S, G|I = i^0) = \Pr(S|I = i^0) \Pr(G|I = i^0) \, .$$

Example 3 However, conditioning can also introduce independence. We have,

$$I, D \to G$$

Look at $\Pr(I, D|G = 1)$. It couples I and D.

1.2.6 Independencies in Bayesian Networks

X and Y are d-separated if $\mathbb{G}$ given Z, d-sep$_{\mathbb{G}}(X, Y|Z)$, if there is no active trail in $\mathbb{G}$ between X and Y given Z.

If Pr factorizes over $\mathbb{G}$, and $d-\text{sep}_{\mathbb{G}}(X, Y|Z)$, then Pr satisfies $(X \perp Y|Z)$. In other words,

$$\text{factorization} \to \text{independence}$$

Any node is d-separated from its non-descendants given its parents. In other words, if Pr factorizes over $\mathbb{G}$, then in Pr, any variable is independent of its non-descendants given its parents.

I-map (Independency Map)

$$I(\mathbb{G}) = \{(X \perp Y | Z) : d - \text{sep}_{\mathbb{G}}(X, Y | Z)\}$$

If Pr satisfies $I(\mathbb{G})$, $\mathbb{G}$ is an I-map of Pr.
Pr factorizes over $\mathbb{G} \leftrightarrow \mathbb{G}$ is an I-map for Pr.

Quiz I-maps can also be defined directly on graphs as follows. Let $I(\mathbb{G})$ be the set of independencies encoded by a graph $\mathbb{G}$. Then $\mathbb{G}_1$ is an I-map for $\mathbb{G}_2$ if $I(\mathbb{G}_1) \subseteq I(\mathbb{G}_2)$.

A graph $\mathbb{K}$ is an I-map for a graph $\mathbb{G}$ if and only if all of the independencies encoded by $\mathbb{K}$ are also encoded by $\mathbb{G}$.

1.2.7 Naive Bayes

Class $C \to$ Features $X_1, \ldots, X_n$.
Assume that $(X_i \perp X_j | C)$ for all X_i, X_j.

$$\Pr(C, X_1, \ldots, X_n) = \Pr(C) \sum_{i=1}^{n} \Pr(X_i | C)$$

1.3 Week 2. Template Models for Bayesian Networks

1.3.1 Temporal Models

Using the chain rule for probability,

$$\Pr(X^{(0:T)}) = \Pr(X^{(0)}) \prod_{t=0}^{T-1} \Pr\left(X^{(t+1)} \middle| X^{(0:t)}\right) .$$

If we make the Markov assumption where the system has no memory, $\left(X^{(t+1)} \perp X^{(0:t-1)} \middle| X^{(t)}\right)$. We have

$$\Pr(X^{(0:T)}) = \Pr(X^{(0)}) \prod_{t=0}^{T-1} \Pr\left(X^{(t+1)} \middle| X^{(t)}\right) .$$

We can further assume that the probability model is time-invariant.

$$\Pr\left(X^{(t+1)} \middle| X^{(t)}\right) = \Pr\left(X' | X\right) .$$

In the traffic model shown in Fig. 3,

$$\Pr(W', V', L', F', O' | W, V, L, F) = \Pr(W'|W) \Pr(V'|W, V) \Pr(L'|L, V) \Pr(F'|F, W) \Pr(O'|L', F') .$$

Note that O is not referenced as it does not affect anything later in time.

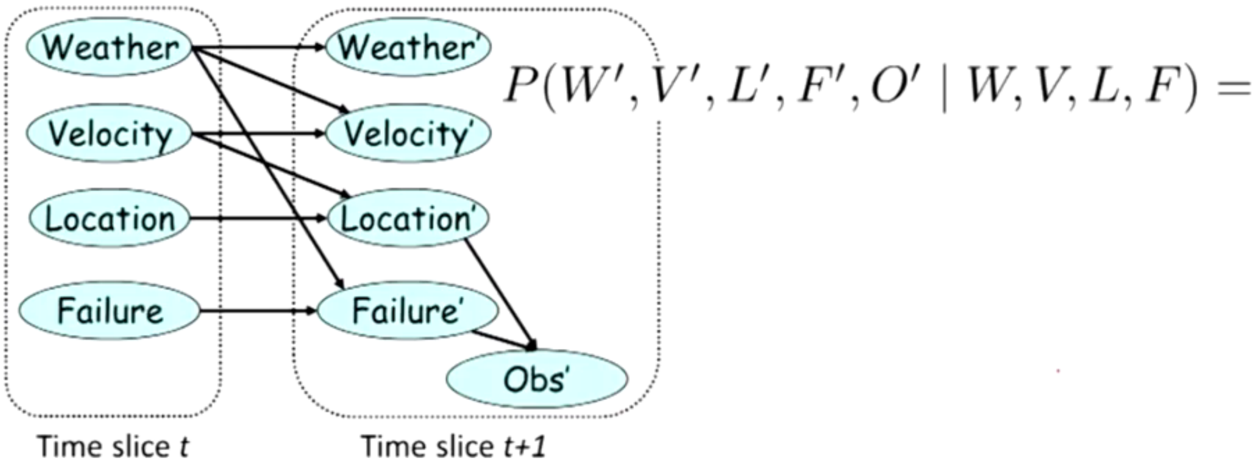

Figure 3: A Traffic Model.

1.4 Week 2. Structured CPDs for Bayesian Networks

1.4.1 Context-Specific Independence

$$\Pr \vdash (X \perp_c Y | Z, c)\,,$$

where c is an assignment of some random variable C.

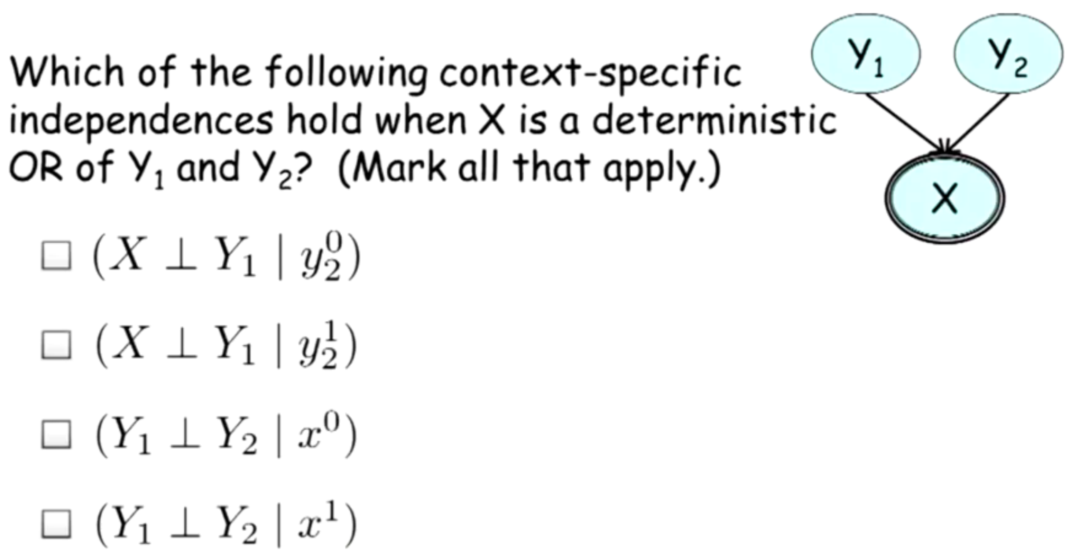

Figure 4: A Sample Question on Context-Specific Independence. The answers are False, True, True, and False.

1.4.2 Tree-Structured CPDs

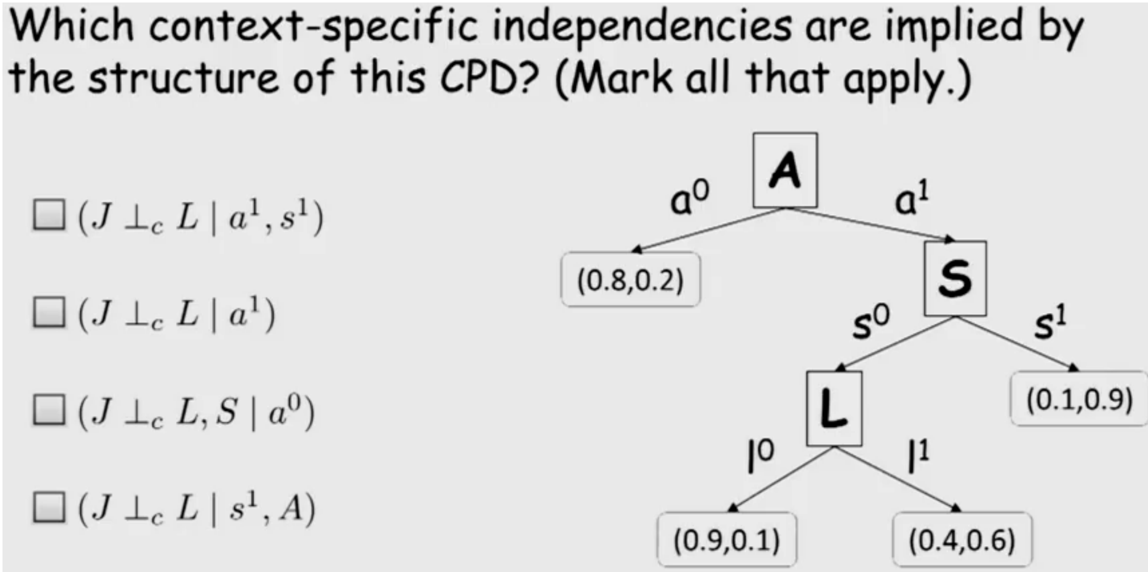

Figure 5: A Sample Question on Tree-Structured CPDs. The answers are True, False, True, and True.

Multiplexer CPD In a multiplexer CPD, A is a selector that decides which of the Z_i's will be copied to Y.

$$\Pr(Y|A, Z_1, \ldots, Z_k) = \begin{cases} 1 & \text{if } Y = Z_A \\ 0 & \text{otherwise} \end{cases}$$

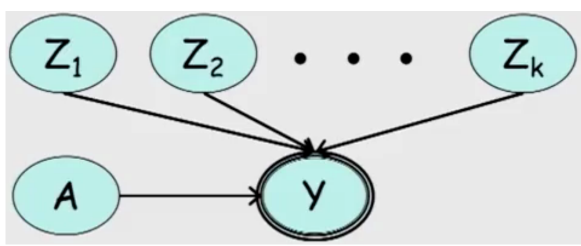

Figure 6: The Multiplexer CPD.

1.4.3 Independence of Causal Influence

Noise OR CPD

$$\Pr(Z_i|X_i) = \begin{cases} 0 & \text{if } X_i = 0 \\ \lambda_i & \text{if } X_i = 1 \end{cases}$$

$$\Pr(Y = 0|X_1, \ldots, X_k) = (1 - \lambda_0) \prod_{i:X_i=1} (1 - \lambda_i)\,.$$

$$\Pr(Y = 1|X_1, \ldots, X_k) = 1 - \Pr(Y = 0|X_1, \ldots, X_k)\,.$$

In other words, $\Pr(Z_i = 0 | X_i = 1) = 1 - \lambda_i$.

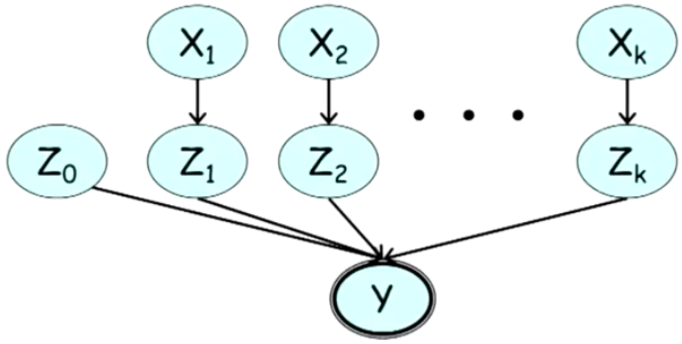

Figure 7: A Noisy OR CPD.

Sigmoid CPD

$$Z = w_0 + \sum_{i=1}^{k} w_i X_i \,.$$
$$\Pr(y^1 | X_1, \ldots, X_k) = \text{sigmoid}(Z)\,.$$

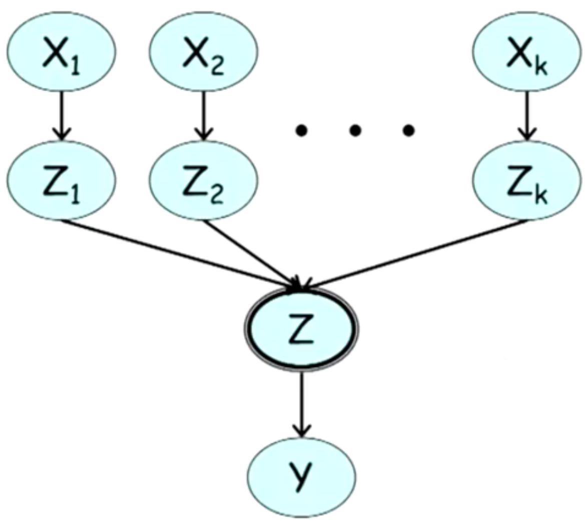

Figure 8: A Sigmoid CPD.

1.4.4 Continuous Variables

Temperature Example

$$S \sim \mathcal{N}(T, \sigma^2)$$
$$T' \sim \begin{cases} \mathcal{N}(\alpha_0 T + (1-\alpha_0)O, \sigma_{0T}^2) & \text{if } D = 0 \\ \mathcal{N}(\alpha_1 T + (1-\alpha_1)O, \sigma_{1T}^2) & \text{if } D = 1 \end{cases}$$

Conditional Linear Gaussian

$$Y \sim \mathcal{N}\left(w_{a0} + \sum_i w_{ai} X_i, \sigma_a^2\right) .$$

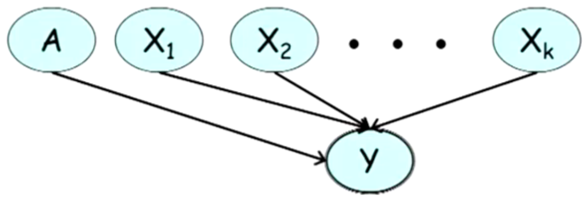

Figure 9: A Conditional Linear Guassian.

1.5 Week 3. Markov Networks (Undirected Models)

1.6 Week 4. Decision Making

1.7 Week 5. Knowledge Engineering & Summary