# Bayesian Networks - m. Hopkins

(1) To understand Bayesian networks, the first thing one needs to understand is the concept on which they are founded: conditional independence.

Let's consider an example involving blood types. Assume that:

- everyone has two genes ("genotype") that determines their blood type; possible values are: AA, AB, AO, BB, BO, OO
- one of these genes is sampled from their mother's genotype; the other is sampled from their father's
- genotype translates to blood type through the mapping

$$\{AA, AO\} \mapsto A \qquad AB \mapsto AB$$
$$\{BB, BO\} \mapsto B \qquad OO \mapsto O$$

Our specific example involves Tim and his parents Rhonda and Sam. Consider the following variables:

- $\bar{R}$ is Rhonda's genotype and $R$ is her blood type
- $\bar{S}$ is Sam's genotype and $S$ is his blood type
- $\bar{T}$ is Tim's genotype and $T$ is his blood type

② Conditional independence is intuitively easy. Here are some true conditional independence statements:

- given we know Rhonda's genotype, knowing her blood type gives us no additional info about Tim's genotype
- given we know Tim's genotype, knowing Sam's genotype gives us no additional info about Tim's blood type

---

③ A more standard way to express these statements:

- $R$ and $\bar{T}$ are conditionally independent given $\bar{R}$ $\left(R \perp\!\!\!\perp \bar{T} \mid \bar{R}\right)$
- $\bar{S}$ and $T$ are conditionally independent given $\bar{T}$ $\left(\bar{S} \perp\!\!\!\perp T \mid \bar{T}\right)$

---

④ How do we concretely define conditional independence?

Two variables $X$ and $Y$ are <u>conditionally independent</u> given a set of variables $Z = \{Z_1, \ldots, Z_k\}$ if:

$$Pr(x \mid y, z) = Pr(x \mid z) \text{ and } Pr(y \mid x, z) = Pr(y \mid z)$$

for all values $x, y, z$ of $X, Y, Z$.

In other words, <u>if we know the value of $Z$</u>, then knowing $Y$ does not change our opinion about the probability of $X$, nor does knowing $X$ change our opinion about the probability of $Y$.

⑤ Here's an example:

| X | Y | Z | P(X, Y, Z) |
|---|---|---|---|
| 0 | 0 | 0 | .04 |
| 0 | 0 | 1 | .10 |
| 0 | 1 | 0 | .06 |
| 0 | 1 | 1 | .10 |
| 1 | 0 | 0 | .16 |
| 1 | 0 | 1 | .15 |
| 1 | 1 | 0 | .24 |
| 1 | 1 | 1 | .15 |

For distribution $P(X, Y, Z)$, try any values:

e.g.
$$P(X=0 \mid Y=1, Z=0) = \frac{P(X=0, Y=1, Z=0)}{P(Y=1, Z=0)} = \frac{.06}{.06+.24} = .2$$
$$P(X=0 \mid Z=0) = \frac{P(X=0, Z=0)}{P(Z=0)} = \frac{.04+.06}{.04+.06+.16+.24} = .2$$

You'll find for all values $x, y, z$, that:

$$P(x \mid y, z) = P(x \mid z) \quad \text{and} \quad P(y \mid x, z) = P(y \mid z)$$

⑥ Note that if the conditioning set $Z$ is empty, we don't typically say:

"X and Y are conditionally independent given $\{\}$."

Rather we say:

"X and Y are marginally independent."

Or just:

"X and Y are independent."

---

⑦ Conditional independence is non-monotonic. For instance:

- Sam's blood type and Rhonda's blood type are marginally independent (since they are not related by blood).
- however Sam and Rhonda's blood types are <u>not</u> conditionally independent given Tim's blood type (Say we know Tim is type B. Then knowing Sam is type O tells us that Rhonda <u>must be</u> either type B or AB.)
- Sam and Rhonda's blood type <u>are</u> conditionally independent given Tim's blood type and Rhonda's genotype (why?)

⑧ A useful application of conditional independence is for expressing joint probability distributions. Consider the full tabular expression of our example:

| $R$ | $S$ | $T$ | $\bar{R} \cup \bar{S}$ | $\bar{T}$ | $Pr(R, S, T, \bar{R}, \bar{S}, \bar{T})$ |
|---|---|---|---|---|---|
| A | A | A | AA | AA | $\theta_{A,A,A,AA,AA,AA}$ |
| A | A | A | AA | AB | $\theta_{A,A,A,AA,AA,AB}$ |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| O | O | O | OO | BO | $\theta_{O,O,O,OO,OO,BO}$ |
| O | O | O | OO | OO | $\theta_{O,O,O,OO,OO,OO}$ |

$4^3 \cdot 6^3 = 13824$ rows

these sum to one

Even with just six variables, it's a lot of work to create this table. We have to estimate and store 13824 parameters (technically we can get away with 13823, because they must sum to one, but still).

(9) But remember the Chain Rule of Probability, which tells us that:

$$\Pr(\bar{r},\bar{s},\bar{t},r,s,t) = \Pr(\bar{r})\cdot\Pr(\bar{s}|\bar{r})\cdot\Pr(\bar{t}|\bar{r},\bar{s})\cdot\Pr(r|\bar{r},\bar{s},\bar{t}) \cdot\Pr(s|\bar{r},\bar{s},\bar{t},r)\cdot\Pr(t|\bar{r},\bar{s},\bar{t},r,s)$$

By itself, this doesn't get us anywhere, until we observe that:

- Sam's genotype is marginally independent of Rhonda's [so $\Pr(\bar{s}|\bar{r}) = \Pr(\bar{s})$ for all values $\bar{r},\bar{s}$]
- Rhonda's blood type is conditionally independent of Sam and Tim's genotypes, given her genotype [so $\Pr(r|\bar{r},\bar{s},\bar{t}) = \Pr(r|\bar{r})$ for all values $\bar{r},\bar{s},\bar{t},r$]
- Sam's blood type is conditionally independent of Rhonda's blood type and Rhonda and Tim's genotype, given Sam's genotype [so $\Pr(s|\bar{r},\bar{s},\bar{t},r) = \Pr(s|\bar{s})$ for all values $\bar{r},\bar{s},\bar{t},r,s$]
- Tim's blood type is conditionally independent of Rhonda and Sam's genotype and blood type, given Tim's genotype [so $\Pr(t|\bar{r},\bar{s},\bar{t},r,s) = \Pr(t|\bar{t})$ for all values $\bar{r},\bar{s},\bar{t},r,s,t$]

This means that:

$$\Pr(\bar{r},\bar{s},\bar{t},r,s,t) = \Pr(\bar{r})\cdot\Pr(\bar{s})\cdot\Pr(\bar{t}|\bar{r},\bar{s})\cdot\Pr(r|\bar{r}) \cdot\Pr(s|\bar{s})\cdot\Pr(t|\bar{t})$$

(10) Now let's take a second look at how many parameters we need to estimate the joint probability:

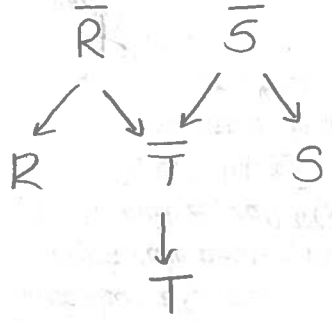
$$\Pr(\bar{r},\bar{s},\bar{t},r,s,t) = \Pr(\bar{r})\Pr(\bar{s})\cdot\Pr(\bar{t}|\bar{r},\bar{s})\cdot\Pr(r|\bar{r})\cdot\Pr(s|\bar{s})\cdot\Pr(t|\bar{t})$$

| $\Pr(\bar{r})$ | $\Pr(\bar{s})$ | $\Pr(\bar{t}\|\bar{r},\bar{s})$ | $\Pr(r\|\bar{r})$ | $\Pr(s\|\bar{s})$ | $\Pr(t\|\bar{t})$ |
|---|---|---|---|---|---|
| 5 | 5 | $6^3-1$ $=215$ | $4\cdot6-1$ $=23$ | $4\cdot6-1$ $=23$ | $4\cdot6-1$ $=23$ |

So we need $5+5+215+23+23+23 = 294$ parameters.

This is way fewer than 13823.

(11) A Bayesian network is simply a graphical depiction of the above factorization:

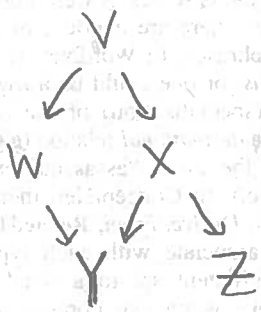



$\equiv$

$$\Pr(\bar{r},\bar{s},\bar{t},r,s,t)$$
$$= \Pr(\bar{r})\cdot\Pr(\bar{s})\cdot\Pr(\bar{t}|\bar{r},\bar{s})$$
$$\cdot\Pr(r|\bar{r})\cdot\Pr(s|\bar{s})\cdot\Pr(t|\bar{t})$$

for all values $\bar{r},\bar{s},\bar{t},r,s,t$

(12) So if we see the following Bayesian network:




Then we can "read" it as making the claim that

$$\Pr(v,w,x,y,z) = \Pr(v)\cdot\Pr(w|v)\cdot\Pr(x|v)\cdot\Pr(y|w,x)\cdot\Pr(z|x)$$

for all values $v, w, x, y, z$.

---

(13) Question: which of these is the "boldest" Bayesian network?

(i) $X \rightarrow Y$ (ii) $X \leftarrow Y$ (iii) $X \quad Y$

Well, let's see what claims they make:

$\Pr(x,y) = \Pr(x)\cdot\Pr(y|x)$ $\Pr(y)\cdot\Pr(x|y)$ $\Pr(x)\,\Pr(y)$

both of these claims are trivially satisfied by all distributions $\Pr(x,y)$ (by the Chain Rule)

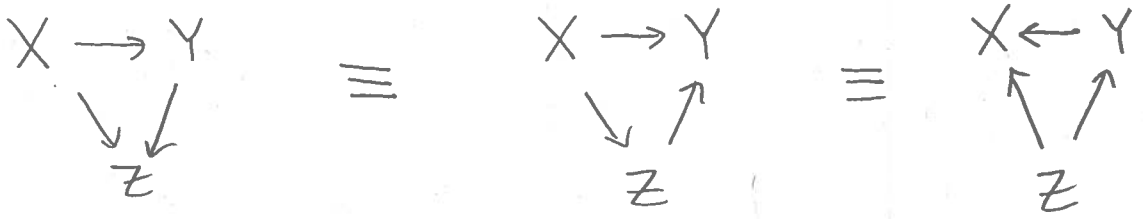
however there are lots of distributions $\Pr(x,y)$ that don't satisfy this claim

So (iii) is the boldest. In fact, (i) and (ii) are equivalent,

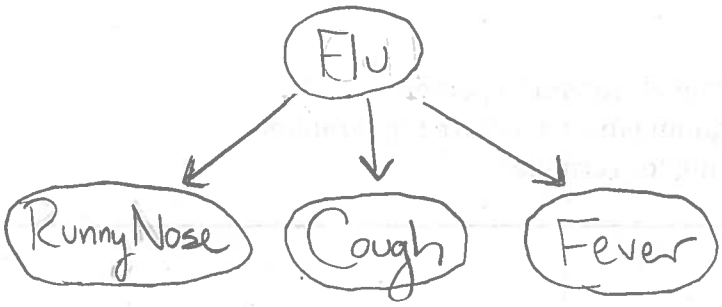
(14) A good way to look at Bayesian networks is that the important part isn't the edges — it's the absent edges.

These are what make the bold claims that certain distributions are not possible, thus allowing us to estimate fewer parameters.

A complete graph always permits every distribution (it is simply a graphical statement of the Chain Rule):



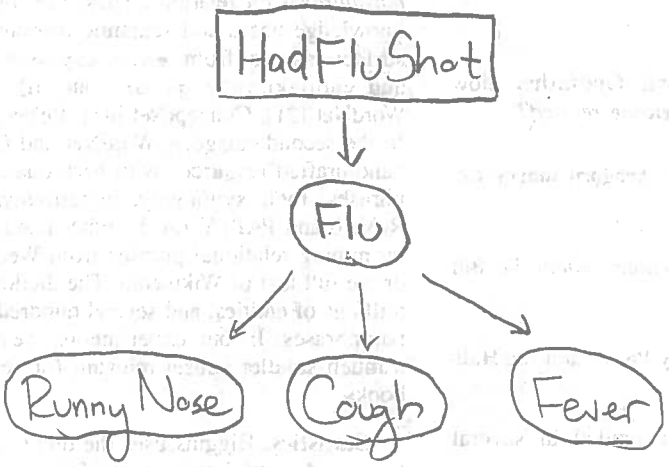
(15) Bayesian networks are a convenient tool for specifying a model, i.e. for communicating the assumptions of your model to someone else. Consider the following example:




I can look at this and immediately see that this model assumes that having a cough and having a fever are conditionally independent, given that I know a person has the flu.

Now perhaps I agree or perhaps I disagree with this assumption, but the utility of the network is that it makes the assumption transparent and explicit.

(16) There are two further notational refinements that help in model communication. First, often one wants to consider relevant variables that one doesn't want to be part of the joint distribution, e.g.

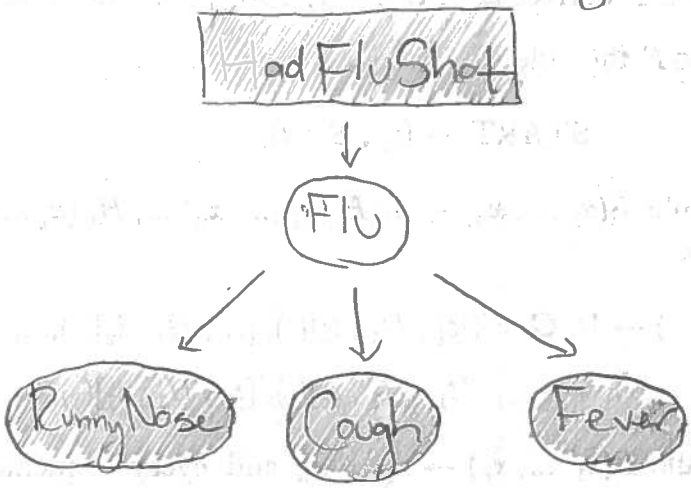



The background variable HadFluShot affects the probability of Flu, but this model declines to go the extra step and model the probability of having a flu shot itself. It is still encoding just the joint distribution:

$$\Pr(\text{Flu}, \text{RunnyNose}, \text{Cough}, \text{Fever})$$

Background variables are assumed to be set outside the scope of the model.

(17) Sometimes we also know which variables are typically observed in practice. The usual notational convention is shading:




This particular shading reflects that we typically can observe symptoms, but not the disease directly. Note that background variables can also be observed or unobserved.