Harvard CS 182: Artificial Intelligence October 27, 2022

Section 8: Bayesian Networks and Hidden Markov Models

Lecturer: Ariel Procaccia *Authors: Mujin Kwun and Christopher Lee*

Bayesian Networks

Bayesian networks are graphs which reveal probabilistic relationships between events. We can then use *Bayesian inference* to connect observable events to other unobservable events. Bayesian inference is a method by which we initially have a belief called a *prior* which, upon noticing some signal, we update to become our *posterior* belief. Formally, this process of updating involves using conditional probabilities, which we generally calculate using *Bayes' rule.*

Conditional Probabilities in Bayes Nets

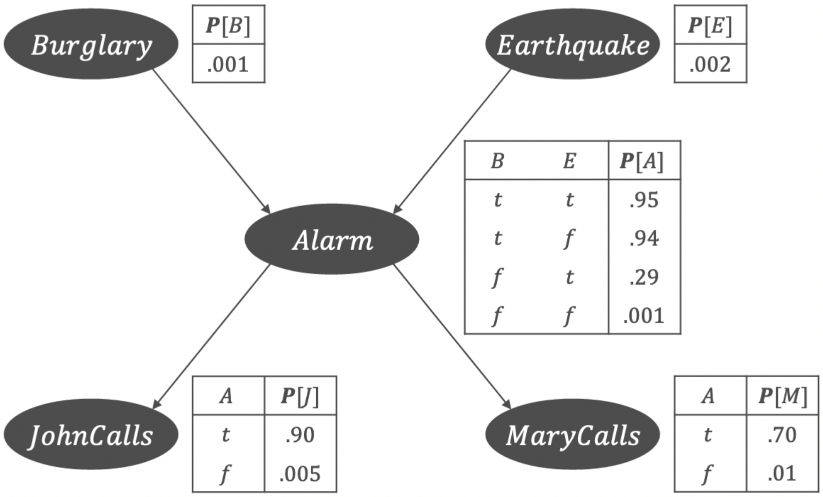
A Bayesian network shows the conditional probability linkages between events. Importantly, each random variable is conditionally independent of its predecessors given its parents; that is, if we have random variables $\{X_i\}_{i=1}^n$ and let x_i be shorthand for the event $X_i = x_i$ for all $i \in \{1, \cdots, n\}$, then

$$\Pr[x_i | x_{i-1}, \cdots, x_1] = \Pr[x_i | parents(X_i)].$$

In other words, if we observe the values of the parents of a node X_i, then we gain no extra information by observing other nodes in the network. Thus, the *joint distribution* of random variables $X_1, \cdots, X_n$ in a Bayesian net is given by

$$\Pr[x_1, \cdots, x_n] = \prod_{i=1}^{n} \Pr[x_i | parents(X_i)].$$

Problem 1 *Consider the Bayesian network from class, shown below, depicting the relationship between two adverse events (burglary and earthquake), an alarm, and two callers (John and Mary). The unconditional probabilities of a burglary and an earthquake are independent, and the probabilities of John and Mary calling conditional on whether or not the alarm has sounded are also independent. Suppose we can only observe whether John calls. If he calls, what is the probability that there was both a burglary and an earthquake?*



P[B]
.001

P[E]
.002

B	E	**P**[A]
t	t	.95
t	f	.94
f	t	.29
f	f	.001

A	**P**[J]
t	.90
f	.005

A	**P**[M]
t	.70
f	.01

We want to calculate $\Pr[B \cap E | J]$, which by definition of conditional probability equals

$$\Pr[B \cap E | J] = \frac{\Pr[B \cap E \cap J]}{\Pr[J]}.$$

Intuitively, conditioning on John having called, there are $2^3 = 8$ possible states of the world (since each of B, E, and A can be either true or false, and we don't care about whether Mary called), and two of these states are in the numerator (where $B = E = J = t$ and A is either t or f). Furthermore, note that in general, we can use the chain rule to calculate

$$\begin{aligned}\Pr[J \cap A \cap B \cap E] &= \Pr[J | A \cap B \cap E] \Pr[A | B \cap E] \Pr[B \cap E] \\ &= \Pr[J | A] \Pr[A | B \cap E] \Pr[B] \Pr[E].\end{aligned}$$

Thus, the numerator becomes

$$\begin{aligned}\Pr[B \cap E \cap J] &= \Pr[J \cap A \cap B \cap E] + \Pr[J \cap A^C \cap B \cap E] \\ &= \Pr[J|A] \Pr[A|B \cap E] \Pr[B] \Pr[E] + \Pr[J|A^C] \Pr[A^C|B \cap E] \Pr[B] \Pr[E] \\ &= 0.90 \cdot 0.95 \cdot 0.001 \cdot 0.002 + 0.005 \cdot 0.05 \cdot 0.001 \cdot 0.002 \\ &= 1.71 \cdot 10^{-6}.\end{aligned}$$

Similarly, the denominator equals

$$\begin{aligned}
\Pr[J] = {} & \Pr[J \cap A \cap B \cap E] + \Pr[J \cap A^C \cap B \cap E] \\
& \quad + \Pr[J \cap A \cap B \cap E^C] + \Pr[J \cap A^C \cap B \cap E^C] \\
& \quad + \Pr[J \cap A \cap B^C \cap E] + \Pr[J \cap A^C \cap B^C \cap E] \\
& \quad + \Pr[J \cap A \cap B^C \cap E^C] + \Pr[J \cap A^C \cap B^C \cap E^C] \\
= {} & \Pr[J|A] \Pr[A|B \cap E] \Pr[B] \Pr[E] + \Pr[J|A^C] \Pr[A^C|B \cap E] \Pr[B] \Pr[E] \\
& \quad \Pr[J|A] \Pr[A|B \cap E^C] \Pr[B] \Pr[E^C] + \Pr[J|A^C] \Pr[A^C|B \cap E^C] \Pr[B] \Pr[E^C] \\
& \quad \Pr[J|A] \Pr[A|B^C \cap E] \Pr[B^C] \Pr[E] + \Pr[J|A^C] \Pr[A^C|B^C \cap E] \Pr[B^C] \Pr[E] \\
& \quad \Pr[J|A] \Pr[A|B^C \cap E^{CC}] \Pr[B^C] \Pr[E^C] + \Pr[J|A^C] \Pr[A^C|B^C \cap E^C] \Pr[B^C] \Pr[E^C] \\
= {} & 7.25 \cdot 10^{-3},
\end{aligned}$$

and thus the final desired probability is

$$\Pr[B \cap E|J] = \frac{\Pr[B \cap E \cap J]}{\Pr[J]} = \frac{1.71 \cdot 10^{-6}}{7.25 \cdot 10^{-3}} = 2.36 \cdot 10^{-4}.$$

Sampling Methods and Likelihood Weighting

Even in relatively simple Bayes nets, the statistical inference calculations can get somewhat lengthy and complicated. Thus, it can be useful to estimate probabilistic relationships using sampling methods. For example, in problem 2, we could estimated $\Pr[B \cap E|J]$ by first sampling B and E then sampling A and J from the given conditional distributions, with our final estimate of being

$$\Pr[B \cap E|J] = \frac{\Pr[J \cap B \cap E]}{\Pr[J]} \approx \frac{\text{\# samples where J, B, and E all occur}}{\text{\# samples where J occurs}}.$$

This process is called *direct sampling*.

However, the accuracy of our estimates relies on having a large enough sample size in both the numerator in the denominator; when the probability of one of the events of interest occurring is very low, satisfying this criterion can require a prohibitively large number of samples. We can resolve this issue by fixing the relevant random variables to the values we want to consider.

Problem 2 *Consider Algorithm 1, the algorithm for calculating likelihood weights presented in class and answer the following questions:*

1. *Mathematically, what does the output w represent?*
2. *How do we use the Likelihood Weighting algorithm to perform statistical inference (i.e. to estimate conditional probabilities)?*

Algorithm 1 Likelihood Weighting

```
Require: $bn, \mathbf{e}$                                  ▷ Input: Bayes net, evidence variables and values
  $w \leftarrow 1$; $\mathbf{x} \leftarrow$ initialized from $\mathbf{e}$          ▷ Initialize: weight to 1, evidence variables
  for $X_i \in \{X_1, \cdots, X_n\}$ do
    if $X_i$ is an evidence variable with value $x_i$ in $\mathbf{e}$ then
      $w \leftarrow w \cdot \Pr[X_i = x_i | parents(X_i)]$      ▷ Weight adjusted by Pr[observing evidence]
    else
      $x_i \leftarrow$ random sample conditioned on its parents
    end if
  end for
  return $\mathbf{x}, w$                      ▷ Return the assigned variable values as well as the weight $w$
```

2.

1. Let the evidence variables be $E = \{E_1, \cdots, E_{n_E}\}$ and the randomly sampled variables be $Z = \{Z_1, \cdots, Z_{n_Z}\}$, and let the values assigned to each variable given by output $\mathbf{x}$ be $\mathbf{e}$ and $\mathbf{z}$, respectively. We see that w is initialized to 1 and is multiplied by $\Pr[E_i = e_i | parents(E_i)]$ for each evidence variable E_i, so the final value of w is

$$w = \prod_{i=1}^{n_E} \Pr[E_i = e_i | parents(E_i)].$$

   Since this is a Bayesian network, the distribution of each decision variable given its parents is independent of all the other nodes in the network, and thus we can simplify this equation to

$$w = \Pr[E = \mathbf{e} | \{X_1, \cdots, X_n\} = \mathbf{x}] = \Pr[E = \mathbf{e} | Z = \mathbf{z}],$$

   where the second step comes from removing the evidence variables E_i from the condition. Thus, w is the probability of observing evidence $\mathbf{e}$ given randomly sampled values $\mathbf{z}$.

2. Once we have sufficiently many random samples, we can estimate the probability distributions of each randomly sampled variable Z_i given the evidence variables E using w as the likelihood of observing evidence variables E given each particular random sample Z. In particular, suppose we run the likelihood weighting function m times on the same Bayesian network using the same evidence variables E and values $\mathbf{e}$ and get outputs $\{(\mathbf{x}^1, w^1), \cdots, (\mathbf{x}^m, w^m)\}$. Let $\mathbf{z}^k$ be the values of the randomly sampled variables in the kth sample. Then we can estimate

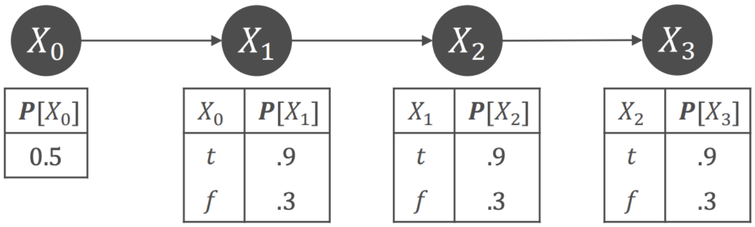
$$\begin{aligned}\Pr[Z_i = z_i | E = \mathbf{e}] &= \frac{\Pr[Z_i = z_i \cap E = \mathbf{e}]}{\Pr[E = \mathbf{e}]} \\ &\approx \frac{\sum_{k \in \{k | Z_i^k = z_i\}} w^k}{\sum_{k=1}^m w^k}.\end{aligned}$$

Then, for example, suppose we want to find $\Pr[Z_i = z_i | Z_j = z_j, E = \mathbf{e}]$ for some $i, j \in \{1, \cdots, n_Z\}$. Mathematically, we can expand this as

$$\Pr[Z_i = z_i | Z_j = z_j, E = \mathbf{e}] = \frac{\Pr[Z_i = z_i, Z_j = z_j, E = \mathbf{e}]}{\Pr[Z_j = z_j, E = \mathbf{e}]},$$

which we can approximate using the previous estimation.

Hidden Markov Models (HMMs)



$\boldsymbol{P}[X_0]$
0.5

X_0	$\boldsymbol{P}[X_1]$
t	.9
f	.3

X_1	$\boldsymbol{P}[X_2]$
t	.9
f	.3

X_2	$\boldsymbol{P}[X_3]$
t	.9
f	.3

Consider the simple Bayes Net shown above. We are now concerned with infinite processes defined by random variables $X_0, X_1, \ldots$. Our bayes net satisfies the following assumptions:

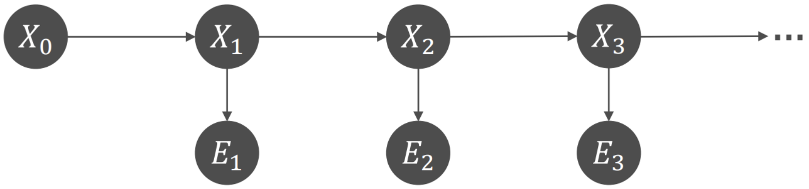
- **Markov Assumption:**

$$\mathbf{P}[X_t | \mathbf{X}_{0:t-1}] = \mathbf{P}[X_t | X_{t-1}]$$

  where $\mathbf{X}_{0:t-1}$ is $X_0 \ldots X_{t-1}$

- **Stationarity Assumption:**

$$\mathbf{P}[X_t | X_{t-1}] = \mathbf{P}[X'_t | X'_{t-1}], \forall t, t'$$



We now add an additional element, which is that rather than observing the state X itself, we can only observe evidence E. Our broad goal is to learn about X while only observing our evidence. In addition to the assumptions we had before for the Bayes Net, we have the **Markov Sensor Assumption:**

$$\mathbf{P}[E_t|\mathbf{X}_{0:t}, \mathbf{E}_{0:t-1}] = \mathbf{P}[E_t|X_t]$$

What are some of our goals with the Markov model?

- **Filtering:** Find $\mathbf{P}[X_{t+1}|\mathbf{e}_{1:t+1}]$ using e_{t+1} and our previous calculation $\mathbf{P}[X_t|\mathbf{e}_{1:t}]$
  - In class we showed:

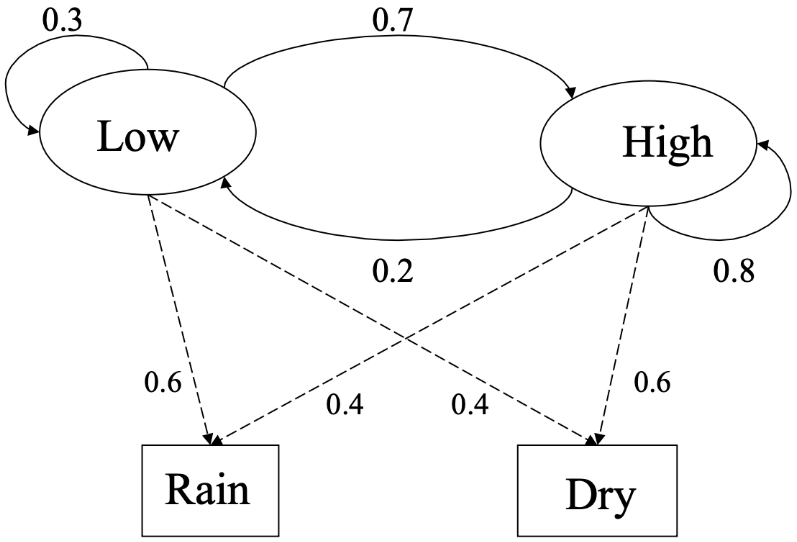
$$\mathbf{P}[X_{t+1}|\mathbf{e}_{1:t+1}] \propto \mathbf{P}[e_{t+1}|X_{t+1}] * \sum_{x_t} Pr[X_t|\mathbf{e}_{1:t}] * \mathbf{P}[X_{t+1}|x_t]$$

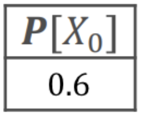
- **Prediction:** $\mathbf{P}[X_{t+k}|\mathbf{e}_{1:t}]$. Predicting state of the world sometime in the future
- **Smoothing:** $\mathbf{P}[X_k|\mathbf{e}_{1:t}]$ where k ¡ t. Using additional evidence to improve our original estimate of the state at an earlier time
- **Max Likelihood:** Interested in finding the most likely sequence of states up to time t
  - In class we showed:

$$max_{x_{0:t}}\mathbf{P}[\mathbf{x}_{0:t}, X_{t+1}|\mathbf{e}_{1:t+1}] \propto \mathbf{P}[e_{t+1}|X_{t+1}]*max_{x_t}\mathbf{P}[X_{t+1}|x_t]*max_{x_{0:t-1}}Pr[\mathbf{x}_{0:t}|\mathbf{e}_{1:t}]$$

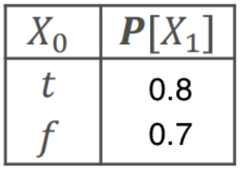
  We can find this using the Viterbi algorithm shown in class

Problem 3 *Consider the Markov Model below where we have two states: low (f) atmospheric pressure and high (t) atmospheric pressure. Unfortunately, we don't have a barometer and can only observe our evidence: whether or not it rained with rain being true and dry being false.*

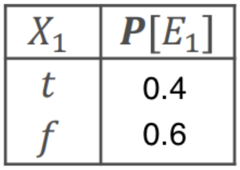




$\boldsymbol{P}[X_0]$
0.6

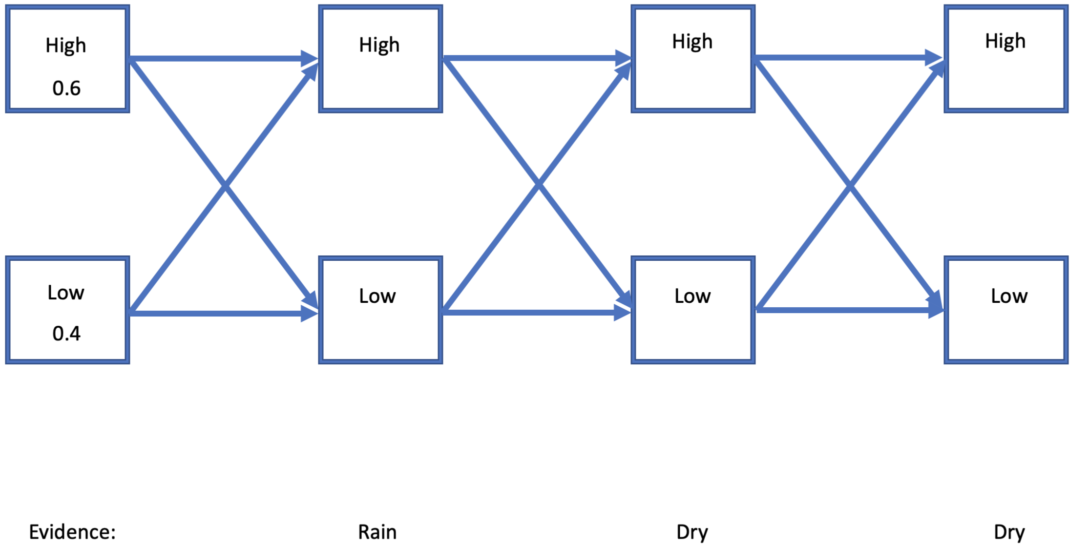


X_0	$\boldsymbol{P}[X_1]$
t	0.8
f	0.7



X_1	$\boldsymbol{P}[E_1]$
t	0.4
f	0.6

1. *What is* $\boldsymbol{P}[X_1 = t|E_1 = t]$*?*
2. *What is* $\boldsymbol{P}[X_1 = f|E_1 = t]$*?*
3. *Using the Viterbi Algorithm presented in lecture, find the missing edge weights in the graph below, given that the evidence we observe is Rain, Dry, Dry for days 1, 2, and 3*



4. *Using the edge weights you found (or by referring to problem 3 solution,) find the probability that we observe each state. What is the most likely sequence of states given the evidence we observe?*

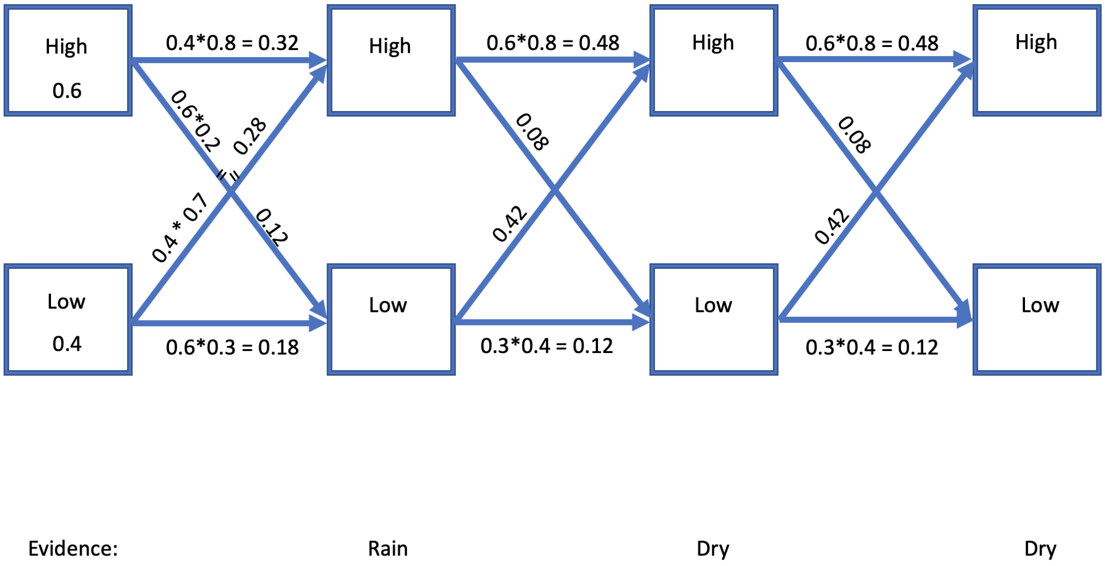
3.

1. $\mathbf{P}[X_1 = t|E_1 = t] \propto 0.4 * (0.6 * 0.8 + 0.4 * 0.7) = 0.304$ $\mathbf{P}[X_1 = f|E_1 = t] \propto 0.6 * (0.6 * 0.2 + 0.4 * 0.3) = 0.144$

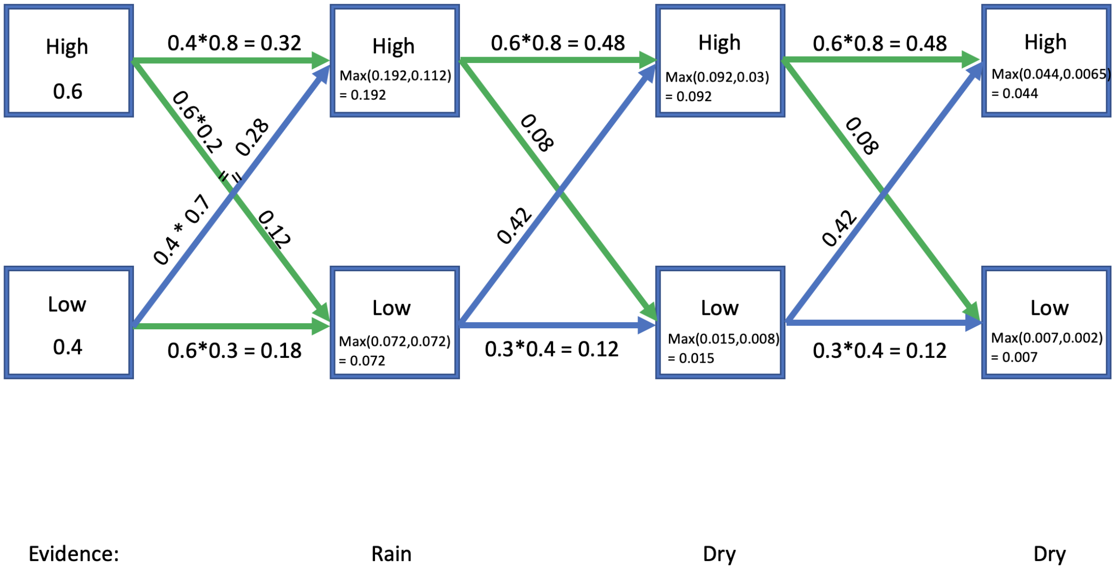
   After normalizing, we get (0.679,0.321), so our answer is 0.679

2. 0.321, see above

3. See graph below



4. After finding these probabilities, we see that Day 1: High, Day 2: High, and Day 3: High is the most likely sequence of states.


High
0.6
Low
0.4
0.4*0.8 = 0.32
0.6*0.2 = 0.12
0.4 * 0.7 = 0.28
0.6*0.3 = 0.18
High
Max(0.192,0.112)
= 0.192
Low
Max(0.072,0.072)
= 0.072
0.6*0.8 = 0.48
0.08
0.42
0.3*0.4 = 0.12
High
Max(0.092,0.03)
= 0.092
Low
Max(0.015,0.008)
= 0.015
0.6*0.8 = 0.48
0.08
0.42
0.3*0.4 = 0.12
High
Max(0.044,0.0065)
= 0.044
Low
Max(0.007,0.002)
= 0.007
Evidence:
Rain
Dry
Dry