Discussion 2

CSE 150A/250A
Fall 2025

Agenda

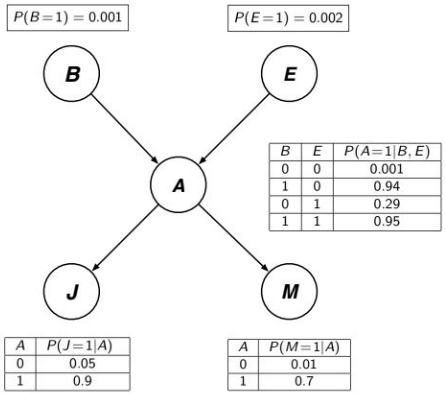
- Review
- HW 2 Written
- HW 2 Coding

Definition

A **belief network** (BN) are a family of probability distributions described by a directed acyclic graph (DAG) in which:

1. Nodes represent random variables.
2. Edges represent direct conditional dependencies.
3. If random variables are discrete, then family of probability distributions are represented as conditional probability tables (CPTs)

In this class, you can assume all variables will be discrete!

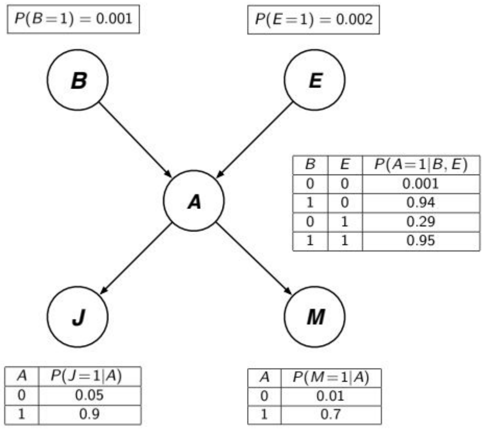


BN = DAG + CPTs

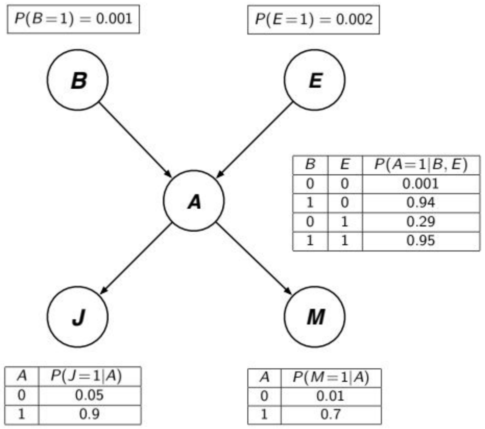
Motivation

Suppose you have **n** binary random variables:

- The joint distribution is: $P(x_1, x_2, ..., x_n)$
- The joint distribution space complexity is: $O(2^n)$



Motivation



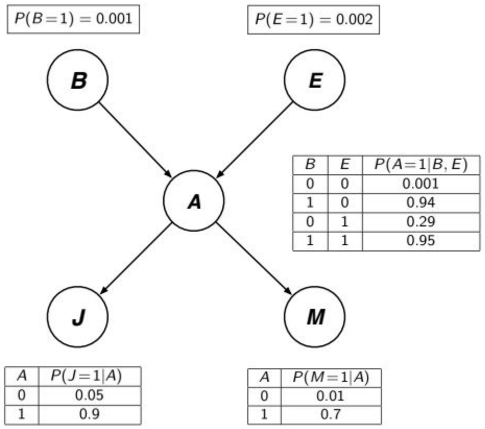
Suppose you have **n** binary random variables:

- The joint distribution is: $P(x_1, x_2, ..., x_n)$
- The joint distribution space complexity is: $O(2^n)$

Recall from chain rule: $p(x_1, x_2, ..., x_n) = p(x_1)p(x_2|x_1)...p(x_n|x_1, x_2, ..., x_{n-1})$

- Using BN's graph, we can remove independent variables
- Where for each variable: $p(x_i|x_1, x_2, ..., x_{i-1}) = p(x_i|\text{pa}(x_i))$
- The space complexity becomes: $O(n2^{k+1})$

Motivation



Suppose you have **n** binary random variables:

- The joint distribution is: $P(x_1, x_2, ..., x_n)$
- The joint distribution space complexity is: $O(2^n)$

Recall from chain rule: $p(x_1, x_2, ..., x_n) = p(x_1)p(x_2|x_1)...p(x_n|x_1, x_2, ..., x_{n-1})$

- Using BN's graph, we can remove independent variables
- Where for each variable: $p(x_i|x_1, x_2, ..., x_{i-1}) = p(x_i|\text{pa}(x_i))$
- The space complexity becomes: $O(n2^{k+1})$

BN's exploit conditional independence relations to reduce space requirements

Exact Inference in BNs

Given a BN, we are interested in using it to answer useful questions:

- $P(X)$
- $P(X|Y)$
- $P(Y|X)$
- **$P(X=x|E)$**

We make the following assumptions:

- The graph structure and CPTs are already defined and learned
- All variables can be observed

Exact Inference in BNs

Given a BN, we are interested in using it to answer useful questions:

- P(X)
- P(X|Y)
- P(Y|X)
- **P(X=x|E)**

We make the following assumptions:

- The graph structure and CPTs are already defined and learned
- All variables can be observed

Goal: Express desired probability expressions in the form of a product of CPTs

Exact Inference in BNs

Two popular approaches (covered in coming lectures):

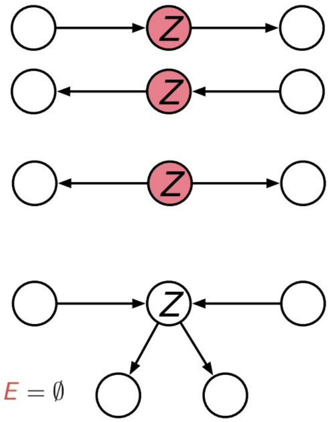
- Variable elimination: Splitting inference into factors and eliminating variables at each factor
- Polytree Inference (Pearl 1986): Start from query node and recursively pass messages up and down polytree, exploiting DAG structure

d-separation

- Theorem

$P(X, Y|E) = P(X|E)\,P(Y|E)$ if and only if every *path* from a node in X to a node in Y is *blocked* by E.

A path π is **blocked** by a set of nodes E if there exists a node $Z \in \pi$ for which one of the three following conditions hold.



(1) $Z \in E$ — edges **align**

(2) $Z \in E$ — edges **diverge**

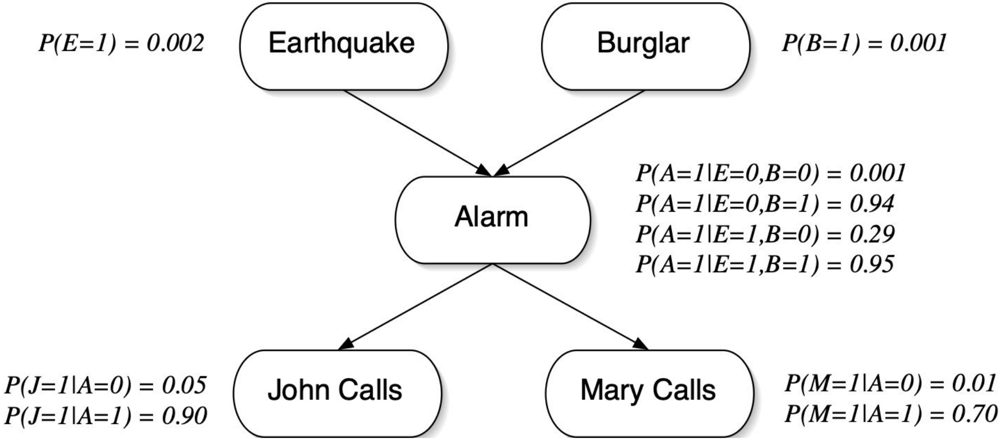
(3) $Z \notin E$ — edges **converge**

descendants$(Z) \cap E = \emptyset$

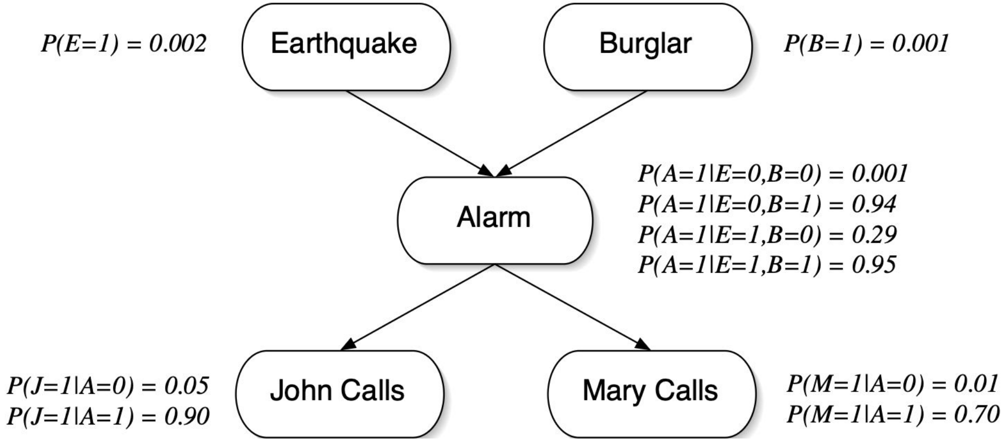
2.1 Probabilistic Inference

You are given the following:

- P(E)
- P(B)
- P(A|B,E)
- P(J|A)
- P(M|A)



2.1 Probabilistic Inference

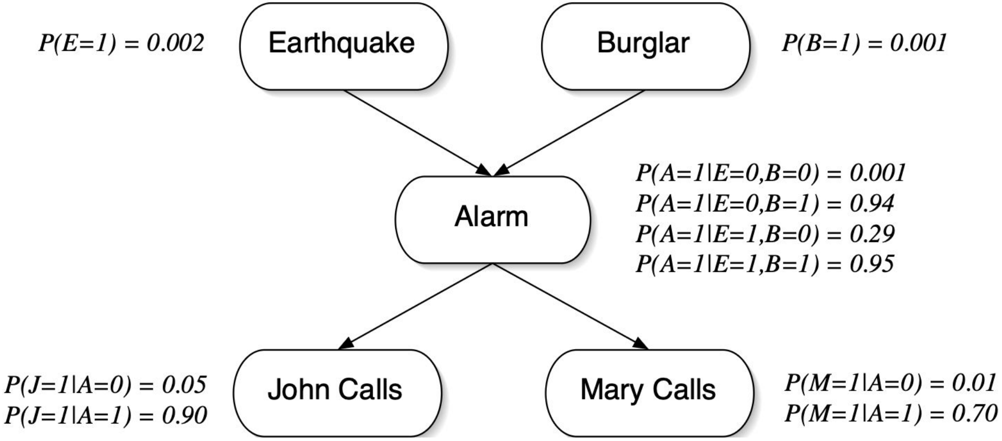


You are given the following:

- P(E)
- P(B)
- P(A|B,E)
- P(J|A)
- P(M|A)

Goal: Manipulate unknown probability expressions into a form containing the given conditional probability tables (CPTs)

2.1 Probabilistic Inference



You are given the following:

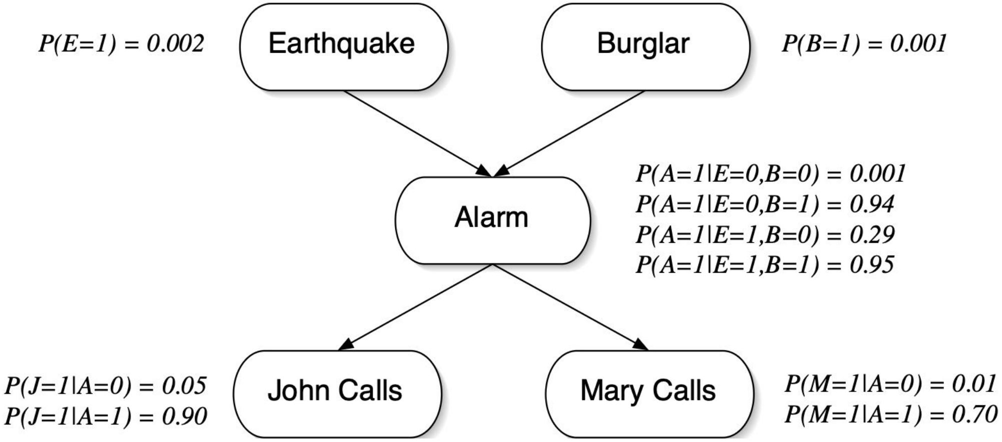
- P(E)
- P(B)
- P(A|B,E)
- P(J|A)
- P(M|A)

Goal: Manipulate unknown probability expressions into a form containing the given conditional probability tables (CPTs)

Solution: Use basic probability theory + exploiting the graph structure to get independence relations

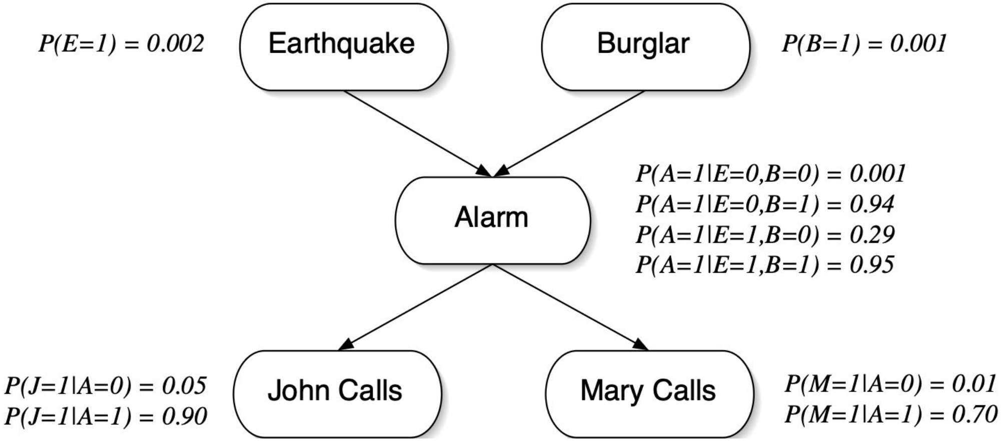
2.1 Probabilistic Inference

Suppose we want to compute: **P(E|J)**



$$P(E|J) =$$

2.1 Probabilistic Inference

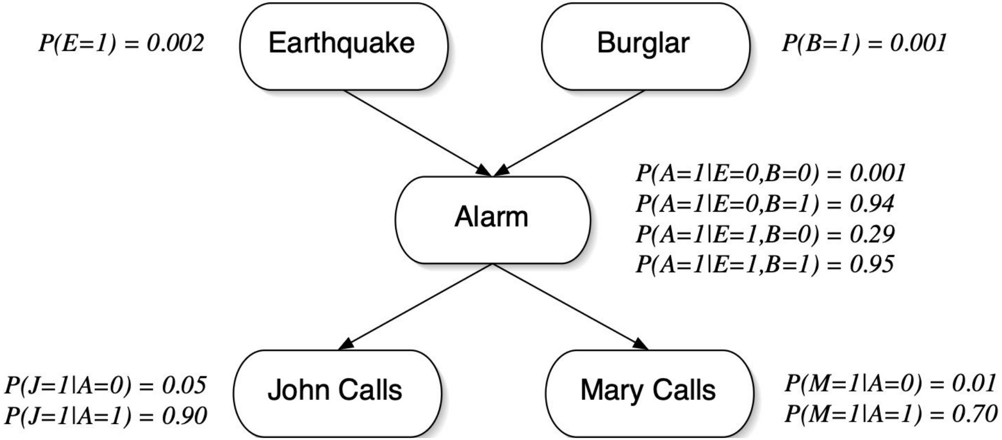


Suppose we want to compute: **P(E|J)**

$$P(E|J) = \frac{P(J|E)P(E)}{P(J)}$$

Bayes rule

2.1 Probabilistic Inference



Suppose we want to compute: **P(E|J)**

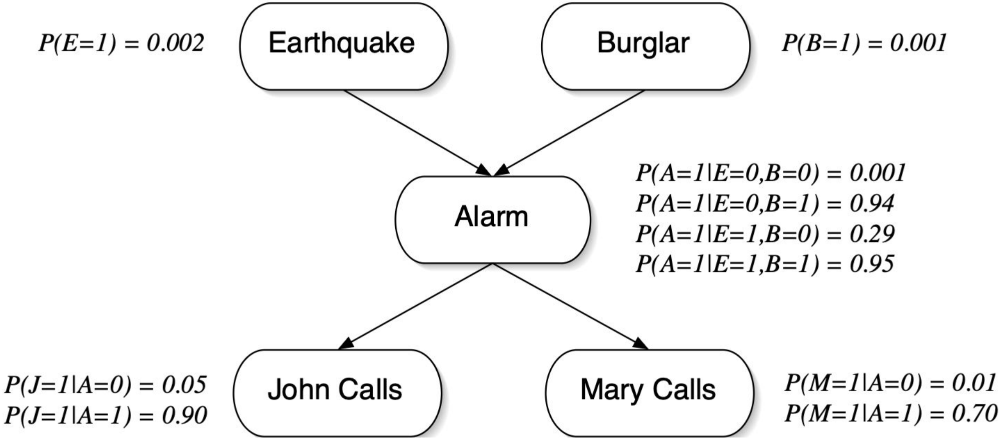
$$P(E|J) = \frac{P(J|E)P(E)}{P(J)}$$

Bayes rule

$$= \frac{\sum_{a,b} P(A=a, B=b, J|E)P(E)}{P(J)}$$

Marginalization

2.1 Probabilistic Inference



Suppose we want to compute: **P(E|J)**

$$P(E|J) = \frac{P(J|E)P(E)}{P(J)}$$

Bayes rule

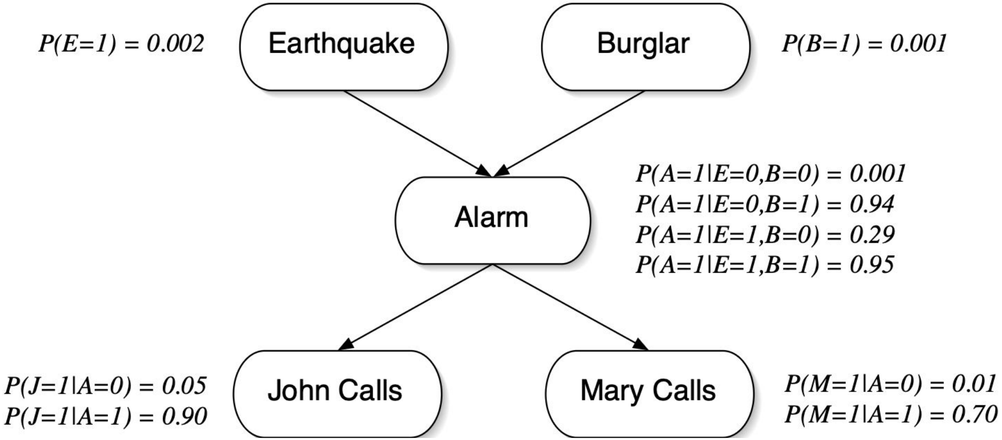
$$= \frac{\sum_{a,b} P(A=a, B=b, J|E)P(E)}{P(J)}$$

Marginalization

$$= \frac{\sum_{a,b} P(B=b|E)P(A=a|B=b,E)P(J|A=a,B=b,E)P(E)}{P(J)}$$

Product rule

2.1 Probabilistic Inference



Suppose we want to compute: **P(E|J)**

$$P(E|J) = \frac{P(J|E)P(E)}{P(J)}$$ **Bayes rule**

$$= \frac{\sum_{a,b} P(A=a, B=b, J|E)P(E)}{P(J)}$$ **Marginalization**

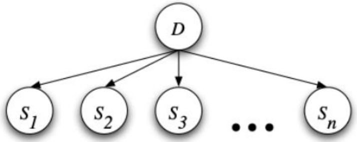
$$= \frac{\sum_{a,b} P(B=b|E)P(A=a|B=b,E)P(J|A=a,B=b,E)P(E)}{P(J)}$$ **Product rule**

$$= \frac{\sum_{a,b} P(A=a|B,E)P(B=b)P(J|A=a)P(E)}{P(J)}$$ **Conditional independence**

2.2 Probabilistic Reasoning

- Part A:

A patient is known to have contracted a rare disease which comes in two forms, represented by the values of a binary random variable $D \in \{0, 1\}$. Symptoms of the disease are represented by the binary random variables $S_k \in \{0, 1\}$, and knowledge of the disease is summarized by the belief network:



The conditional probability tables (CPTs) for this belief network are as follows. In the absence of evidence, both forms of the disease are equally likely, with prior probabilities:

$$P(D=0) = P(D=1) = \frac{1}{2}.$$

In one form of the disease ($D=0$), the first symptom occurs with probability one,

$$P(S_1=1|D=0) = 1,$$

while the k^{th} symptom (with $k \geq 2$) occurs with probability

$$P(S_k=1|D=0) = \frac{f(k-1)}{f(k)},$$

where the function $f(k)$ is defined by

$$f(k) = 2^k + (-1)^k.$$

By contrast, in the other form of the disease ($D=1$), all the symptoms are uniformly likely to be observed, with

$$P(S_k=1|D=1) = \frac{1}{2}$$

for all k. Suppose that on the k^{th} day of the month, a test is done to determine whether the patient is exhibiting the k^{th} symptom, and that each such test returns a positive result. Thus, on the k^{th} day, the doctor observes the patient with symptoms $\{S_1=1, S_2=1, \ldots, S_k=1\}$. Based on the cumulative evidence, the doctor makes a new diagnosis each day by computing the ratio:

$$r_k = \frac{P(D=0|S_1=1, S_2=1, \ldots, S_k=1)}{P(D=1|S_1=1, S_2=1, \ldots, S_k=1)}.$$

If this ratio is greater than 1, the doctor diagnoses the patient with the $D=0$ form of the disease; otherwise, with the $D=1$ form.

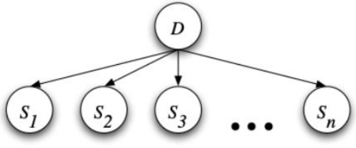
(a) Compute the ratio r_k as a function of k. How does the doctor's diagnosis depend on the day of the month? Show your work.

(b) Does the diagnosis become more or less certain as more symptoms are observed? Explain.

2.2 Probabilistic Reasoning

- Part A: Represent each term in the ratio in terms of the given probabilities:
  - $P(D), P(S_i = 1|D)$
- Similar approach to 2.1

A patient is known to have contracted a rare disease which comes in two forms, represented by the values of a binary random variable $D \in \{0, 1\}$. Symptoms of the disease are represented by the binary random variables $S_k \in \{0, 1\}$, and knowledge of the disease is summarized by the belief network:



The conditional probability tables (CPTs) for this belief network are as follows. In the absence of evidence, both forms of the disease are equally likely, with prior probabilities:

$$P(D=0) = P(D=1) = \frac{1}{2}.$$

In one form of the disease ($D=0$), the first symptom occurs with probability one,

$$P(S_1=1|D=0) = 1,$$

while the k^{th} symptom (with $k \geq 2$) occurs with probability

$$P(S_k=1|D=0) = \frac{f(k-1)}{f(k)},$$

where the function $f(k)$ is defined by

$$f(k) = 2^k + (-1)^k.$$

By contrast, in the other form of the disease ($D=1$), all the symptoms are uniformly likely to be observed, with

$$P(S_k=1|D=1) = \frac{1}{2}$$

for all k. Suppose that on the k^{th} day of the month, a test is done to determine whether the patient is exhibiting the k^{th} symptom, and that each such test returns a positive result. Thus, on the k^{th} day, the doctor observes the patient with symptoms $\{S_1=1, S_2=1, \ldots, S_k=1\}$. Based on the cumulative evidence, the doctor makes a new diagnosis each day by computing the ratio:

$$r_k = \frac{P(D=0|S_1=1, S_2=1, \ldots, S_k=1)}{P(D=1|S_1=1, S_2=1, \ldots, S_k=1)}.$$

If this ratio is greater than 1, the doctor diagnoses the patient with the $D=0$ form of the disease; otherwise, with the $D=1$ form.

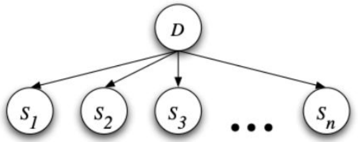
(a) Compute the ratio r_k as a function of k. How does the doctor's diagnosis depend on the day of the month? Show your work.

(b) Does the diagnosis become more or less certain as more symptoms are observed? Explain.

2.2 Probabilistic Reasoning

- Part A: Represent each term in the ratio in terms of the given probabilities:
  - $P(D), P(S_i = 1|D)$
- Similar approach to 2.1
- Part B: take the limit of the ratio as k goes to infinity. Does this limit converge or diverge?

A patient is known to have contracted a rare disease which comes in two forms, represented by the values of a binary random variable $D \in \{0, 1\}$. Symptoms of the disease are represented by the binary random variables $S_k \in \{0, 1\}$, and knowledge of the disease is summarized by the belief network:



The conditional probability tables (CPTs) for this belief network are as follows. In the absence of evidence, both forms of the disease are equally likely, with prior probabilities:

$$P(D=0) = P(D=1) = \frac{1}{2}.$$

In one form of the disease ($D=0$), the first symptom occurs with probability one,

$$P(S_1=1|D=0) = 1,$$

while the k^{th} symptom (with $k \geq 2$) occurs with probability

$$P(S_k=1|D=0) = \frac{f(k-1)}{f(k)},$$

where the function $f(k)$ is defined by

$$f(k) = 2^k + (-1)^k.$$

By contrast, in the other form of the disease ($D=1$), all the symptoms are uniformly likely to be observed, with

$$P(S_k=1|D=1) = \frac{1}{2}$$

for all k. Suppose that on the k^{th} day of the month, a test is done to determine whether the patient is exhibiting the k^{th} symptom, and that each such test returns a positive result. Thus, on the k^{th} day, the doctor observes the patient with symptoms $\{S_1=1, S_2=1, \ldots, S_k=1\}$. Based on the cumulative evidence, the doctor makes a new diagnosis each day by computing the ratio:

$$r_k = \frac{P(D=0|S_1=1, S_2=1, \ldots, S_k=1)}{P(D=1|S_1=1, S_2=1, \ldots, S_k=1)}.$$

If this ratio is greater than 1, the doctor diagnoses the patient with the $D=0$ form of the disease; otherwise, with the $D=1$ form.

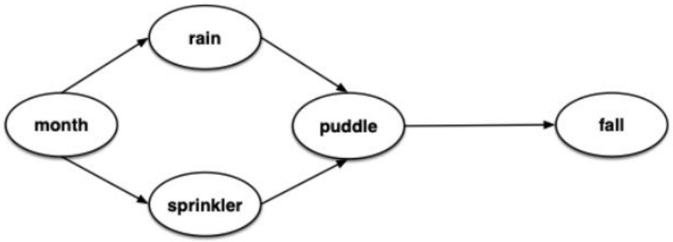
(a) Compute the ratio r_k as a function of k. How does the doctor's diagnosis depend on the day of the month? Show your work.

(b) Does the diagnosis become more or less certain as more symptoms are observed? Explain.

2.3 Conditional Independence

- Main idea: Find all conditional independence relations (using d-separation)
- Note: X and Y are individual nodes while E is a set of nodes
- Hint: Only a subset of nodes can possibly be independent.

Consider the DAG shown below, describing the following domain. Given the `month` of the year, there is some probability of `rain`, and also some probability that the `sprinkler` is turned on. Either of these events leads to some probability that a `puddle` forms on the sidewalk, which in turn leads to some probability that someone has a `fall`.



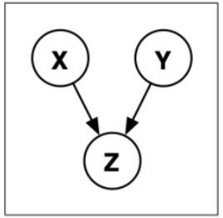
List all the conditional independence relations that must hold in any probability distribution represented by this DAG. More specifically, list all tuples $\{X, Y, E\}$ such that $P(X, Y|E) = P(X|E)P(Y|E)$, where

$$
\begin{aligned}
X, Y &\in \{\texttt{month}, \texttt{rain}, \texttt{sprinkler}, \texttt{puddle}, \texttt{fall}\}, \\
E &\subseteq \{\texttt{month}, \texttt{rain}, \texttt{sprinkler}, \texttt{puddle}, \texttt{fall}\}, \\
X &\neq Y, \\
X, Y &\notin E.
\end{aligned}
$$

Hint: There are sixteen such tuples, not counting those that are equivalent up to exchange of X and Y. Do any of the tuples contain the case $E = \emptyset$?

2.4 Noisy-OR

- Noisy-OR: OR but not guaranteed to turn on even X and/or Y are on.
- Approaches
  - Intuition (see L2)
  - Mathematical derivation (2.1-2.2)
- Not required to show work, just fill in the boxes with the corresponding operator: =,<, or >



Nodes: $X \in \{0,1\}, Y \in \{0,1\}, Z \in \{0,1\}$

Noisy-OR CPT: $P(Z=1|X,Y) = 1 - (1-p_x)^X (1-p_y)^Y$

Parameters: $p_x \in [0,1], p_y \in [0,1], p_x < p_y$

Suppose that the nodes in this network represent binary random variables and that the CPT for $P(Z|X,Y)$ is parameterized by a noisy-OR model, as shown above. Suppose also that

$$0 < P(X=1) < 1,$$
$$0 < P(Y=1) < 1,$$

while the parameters of the noisy-OR model satisfy:

$$0 < p_x < p_y < 1.$$

Consider the following pairs of probabilities. In each case, indicate whether the probability on the left is equal (=), greater than (>), or less than (<) the probability on the right. The first one has been filled in for you as an example. (You should use your intuition for these problems; you are **not** required to show work.)

	$P(X=1)$	=	$P(X=1)$
(a)	$P(Z=1\|X=0,Y=0)$		$P(Z=1\|X=0,Y=1)$
(b)	$P(Z=1\|X=1,Y=0)$		$P(Z=1\|X=0,Y=1)$
(c)	$P(Z=1\|X=1,Y=0)$		$P(Z=1\|X=1,Y=1)$
(d)	$P(X=1)$		$P(X=1\|Z=1)$
(e)	$P(X=1)$		$P(X=1\|Y=1)$
(f)	$P(X=1\|Z=1)$		$P(X=1\|Y=1,Z=1)$
(g)	$P(X=1)\,P(Y=1)\,P(Z=1)$		$P(X=1,Y=1,Z=1)$

2.5 Hangman

You are tasked with coding up a portion of the hangman problem:

- Given a state, find the next letter with the highest probability of being in the word. (predictive probability)
- State is represented by the BN and characters already guessed (correct and incorrect)

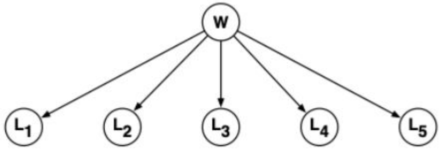
Consider the belief network shown below, where the random variable W stores a five-letter word and the random variable $L_i \in \{\mathtt{A}, \mathtt{B}, \ldots, \mathtt{Z}\}$ reveals only the word's ith letter. Also, suppose that these five-letter words are chosen at random from a large corpus of text according to their frequency:

$$P(W\!=\!w) \;=\; \frac{\text{COUNT}(w)}{\sum_{w'} \text{COUNT}(w')},$$

where $\text{COUNT}(w)$ denotes the number of times that w appears in the corpus and where the denominator is a sum over all five-letter words. Note that in this model the conditional probability tables for the random variables L_i are particularly simple:

$$P(L_i\!=\!\ell|W\!=\!w) \;=\; \begin{cases} 1 & \text{if } \ell \text{ is the } i\text{th letter of } w, \\ 0 & \text{otherwise.} \end{cases}$$

Now imagine a game in which you are asked to guess the word w one letter at a time. The rules of this game are as follows: after each letter (A through Z) that you guess, you'll be told whether the letter appears in the word and also where it appears. Given the *evidence* that you have at any stage in this game, the critical question is what letter to guess next.



Let's work an example. Suppose that after three guesses—the letters D, I, M—you've learned that the letter I does *not* appear, and that the letters D and M appear as follows:

M ___ D ___ M

Now consider your next guess: call it ℓ. In this game the best guess is the letter ℓ that maximizes

$$P\Big(L_2\!=\!\ell \textbf{ or } L_4\!=\!\ell \,\Big|\, L_1\!=\!\mathtt{M}, L_3\!=\!\mathtt{D}, L_5\!=\!\mathtt{M}, L_2\!\notin\!\{\mathtt{D},\mathtt{I},\mathtt{M}\}, L_4\!\notin\!\{\mathtt{D},\mathtt{I},\mathtt{M}\}\Big).$$

In other works, pick the letter ℓ that is most likely to appear in the blank (unguessed) spaces of the word. For any letter ℓ we can compute this probability as follows:

$$\begin{aligned} &P\Big(L_2\!=\!\ell \textbf{ or } L_4\!=\!\ell \,\Big|\, L_1\!=\!\mathtt{M}, L_3\!=\!\mathtt{D}, L_5\!=\!\mathtt{M}, L_2\!\notin\!\{\mathtt{D},\mathtt{I},\mathtt{M}\}, L_4\!\notin\!\{\mathtt{D},\mathtt{I},\mathtt{M}\}\Big) \\ &\quad= \sum_w P\Big(W\!=\!w, L_2\!=\!\ell \textbf{ or } L_4\!=\!\ell \,\Big|\, L_1\!=\!\mathtt{M}, L_3\!=\!\mathtt{D}, L_5\!=\!\mathtt{M}, L_2\!\notin\!\{\mathtt{D},\mathtt{I},\mathtt{M}\}, L_4\!\notin\!\{\mathtt{D},\mathtt{I},\mathtt{M}\}\Big), \quad \boxed{\textbf{marginalization}} \\ &\quad= \sum_w P(W\!=\!w|L_1\!=\!\mathtt{M}, L_3\!=\!\mathtt{D}, L_5\!=\!\mathtt{M}, L_2\!\notin\!\{\mathtt{D},\mathtt{I},\mathtt{M}\}, L_4\!\notin\!\{\mathtt{D},\mathtt{I},\mathtt{M}\}\Big)\, P(L_2\!=\!\ell \textbf{ or } L_4\!=\!\ell|W\!=\!w) \quad \boxed{\textbf{product rule \& CI}} \end{aligned}$$

2.5 Hangman

Your tasks:

- Translate the hangman formulation provided into code
- Provide the following deliverables:
  - Two sets containing the 15 most frequent and 14 least frequent words in the corpus
  - Table containing the most likely letter and probability for the given evidence

where in the third line we have exploited the conditional independence (**CI**) of the letters L_i given the word W. Inside this sum there are two terms, and they are both easy to compute. In particular, the second term is more or less trivial:

$$P(L_2\!=\!\ell \text{ or } L_4\!=\!\ell|W\!=\!w) \;=\; \begin{cases} 1 & \text{if } \ell \text{ is the second or fourth letter of } w \\ 0 & \text{otherwise.} \end{cases}$$

And the first term we obtain from Bayes rule:

$$P\big(W\!=\!w|L_1\!=\!\mathtt{M}, L_3\!=\!\mathtt{D}, L_5\!=\!\mathtt{M}, L_2\!\notin\!\{\mathtt{D},\mathtt{I},\mathtt{M}\}, L_4\!\notin\!\{\mathtt{D},\mathtt{I},\mathtt{M}\}\big)$$

$$= \frac{P\big(L_1\!=\!\mathtt{M}, L_3\!=\!\mathtt{D}, L_5\!=\!\mathtt{M}, L_2\!\notin\!\{\mathtt{D},\mathtt{I},\mathtt{M}\}, L_4\!\notin\!\{\mathtt{D},\mathtt{I},\mathtt{M}\}|W\!=\!w\big)P(W\!=\!w)}{P(L_1\!=\!\mathtt{M}, L_3\!=\!\mathtt{D}, L_5\!=\!\mathtt{M}, L_2\!\notin\!\{\mathtt{D},\mathtt{I},\mathtt{M}\}, L_4\!\notin\!\{\mathtt{D},\mathtt{I},\mathtt{M}\})} \quad \boxed{\textbf{Bayes rule}}$$

In the numerator of Bayes rule are two terms; the left term is equal to zero or one (depending on whether the evidence is compatible with the word w), and the right term is the prior probability $P(W=w)$, as determined by the empirical word frequencies. The denominator of Bayes rule is given by:

$$\begin{aligned} &P(L_1\!=\!\mathtt{M}, L_3\!=\!\mathtt{D}, L_5\!=\!\mathtt{M}, L_2\!\notin\!\{\mathtt{D},\mathtt{I},\mathtt{M}\}, L_4\!\notin\!\{\mathtt{D},\mathtt{I},\mathtt{M}\}) \\ &\quad= \sum_w P(W\!=\!w, L_1\!=\!\mathtt{M}, L_3\!=\!\mathtt{D}, L_5\!=\!\mathtt{M}, L_2\!\notin\!\{\mathtt{D},\mathtt{I},\mathtt{M}\}, L_4\!\notin\!\{\mathtt{D},\mathtt{I},\mathtt{M}\}), \quad \boxed{\textbf{marginalization}} \\ &\quad= \sum_w P(W\!=\!w)P(L_1\!=\!\mathtt{M}, L_3\!=\!\mathtt{D}, L_5\!=\!\mathtt{M}, L_2\!\notin\!\{\mathtt{D},\mathtt{I},\mathtt{M}\}, L_4\!\notin\!\{\mathtt{D},\mathtt{I},\mathtt{M}\}|W\!=\!w), \quad \boxed{\textbf{product rule}} \end{aligned}$$

where again all the right terms inside the sum are equal to zero or one. Note that the denominator merely sums the empirical frequencies of words that are compatible with the observed evidence.

Now let's consider the general problem. Let E denote the evidence at some intermediate round of the game: in general, some letters will have been guessed correctly and their places revealed in the word, while other letters will have been guessed incorrectly and thus revealed to be absent. There are two essential computations. The first is the *posterior* probability, obtained from Bayes rule:

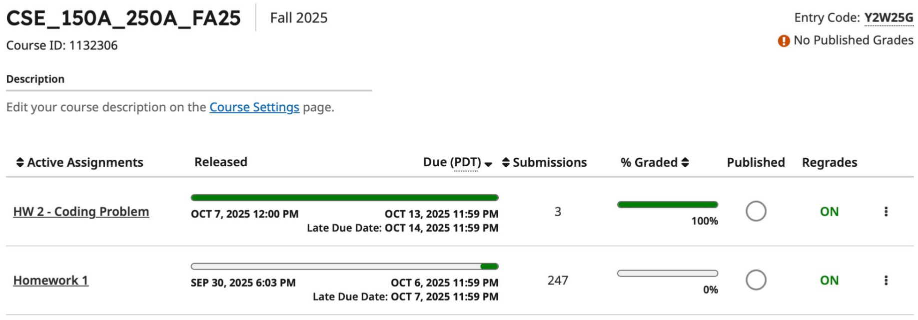
$$P(W\!=\!w|E) \;=\; \frac{P(E|W\!=\!w)\,P(W\!=\!w)}{\sum_{w'} P(E|W\!=\!w')\,P(W\!=\!w')}.$$

The second key computation is the *predictive* probability, based on the evidence, that the letter ℓ appears somewhere in the word:

$$P\Big(L_i\!=\!\ell \text{ for some } i\!\in\!\{1,2,3,4,5\}\Big|E\Big) \;=\; \sum_w P\Big(L_i\!=\!\ell \text{ for some } i\!\in\!\{1,2,3,4,5\}\Big|W\!=\!w\Big)P\Big(W\!=\!w\Big|E\Big).$$

Submission

- Submit just the .py file to HW 2 - Coding problem
- Feel free to modify any function signatures (except those which have do not modify) or add addition functions
- We will run your run() function and compare with our solution code
- **Due Oct 13th, 11:59 PM** (both parts!)



That's all folks!