CSE 150A - Homework 1 *(80 pts)*

Due: *Sunday, Jan 26* (by 11:59 PM, Pacific Time, via gradescope)

1.0 General Guidelines

You should submit your homework assignments via Gradescope. Typesetting / LaTeX is preferred, but neatly handwritten solutions are also accepted. Upload a PDF of your answers to Homework 1 on Gradescope, and carefully select the pages corresponding to each question using the submission interface. If you are submitting handwritten answers, please scan them and create a PDF for upload. Here is a primer on submitting PDF homework via Gradescope:

`https://tinyurl.com/gradescope-guide`

If you have not done this before, please allow some extra time to familiarize yourself with this process.

Late Day Policy: There will be no penalty for turning in any of the homework assignments up to 24hrs late. However, we cannot guarantee that late assignments will be graded in a timely fashion. Beyond the 24hr grace period, we will not accept late homework.

Collaboration: We strongly encourage collaboration (**but NOT copying**) on the homework assignments. You may talk to anyone in the course about how to solve the problems, and you may even compare your solutions. However, you must write up your solutions yourself, and you may not copy them.

1.1 Conditioning on background evidence *(3 pts)*

It is often useful to consider the impact of specific events in the context of general background evidence, rather than in the absence of information.

(a) Denoting such evidence by E, prove the conditionalized version of the product rule: *(1 pts)*

$$P(X, Y|E) \;=\; P(X|Y, E)P(Y|E).$$

(b) Also, prove the conditionalized version of Bayes rule: *(1 pts)*

$$P(X|Y, E) \;=\; \frac{P(Y|X, E)P(X|E)}{P(Y|E)}.$$

(c) Also, prove the conditionalized version of marginalization: *(1 pts)*

$$P(X|E) \;=\; \sum_y P(X, Y = y|E).$$

1.2 Conditional independence *(3 pts)*

Show that the following three statements about random variables X, Y, and E are equivalent:

$$\begin{aligned} (1) \quad P(X,Y|E) &= P(X|E)P(Y|E) \\ (2) \quad P(X|Y,E) &= P(X|E) \\ (3) \quad P(Y|X,E) &= P(Y|E) \end{aligned}$$

In other words, show that (1) implies (2) & (3), that (2) implies (1) & (3), and that (3) implies (1) & (2). You should become fluent with all these ways of expressing that X is conditionally independent of Y given E.

1.3 Creative writing *(3 pts)*

This problem does not involve any calculations: simply attach events to the binary random variables X, Y, and Z that are consistent with the following patterns of commonsense reasoning. You may use different events for the different parts of the problem. Also, please be creative: do not use the same events (e.g., burglaries, earthquakes, alarms) that were considered in lecture.

(a) Cumulative evidence: *(1 pts)*

$$P(X=1) < P(X=1|Y=1) < P(X=1|Y=1,Z=1)$$

(b) Explaining away: *(1 pts)*

$$\begin{aligned} P(X=1|Y=1) &> P(X=1), \\ P(X=1|Y=1,Z=1) &< P(X=1|Y=1) \end{aligned}$$

(c) Conditional independence: *(1 pts)*

$$\begin{aligned} P(X=1,Y=1) &\neq P(X=1)P(Y=1) \\ P(X=1,Y=1|Z=1) &= P(X=1|Z=1)P(Y=1|Z=1) \end{aligned}$$

1.4 Bayes Rule *(6 pts)*

Suppose that 1% of competitive cyclists use performance-enhancing drugs and that a particular drug test has a 5% false positive rate and a 10% false negative rate. Let $D \in \{0,1\}$ indicate whether a cyclist is doping, and let $T \in \{0,1\}$ indicate the outcome of the drug test.

(a) Cyclist A tests negative for drug use. What is the probability that Cyclist A is not using drugs? *(3 pts)*

(b) Cyclist B tests positive for drug use. What is the probability that Cyclist B is using drugs? *(3 pts)*

1.5 Kullback-Leibler distance *(9 pts)*

Often it is useful to measure the difference between two probability distributions over the same random variable. For example, as shorthand let

$$p_i = P(X{=}i|E), \qquad q_i = P(X{=}i|E')$$

denote the conditional distributions over the random variable X for different pieces of evidence $E \neq E'$. Note that $\sum_i p_i = \sum_i q_i = 1$. The Kullback-Leibler (KL) distance between these distributions (also known as the relative entropy) is defined as:

$$\mathrm{KL}(p, q) = \sum_i p_i \log(p_i/q_i).$$

(a) Consider the natural logarithm (in base e). By sketching graphs of $\log(x)$ and $x - 1$, verify the inequality: *(2 pts)*

$$\log(x) \;\leq\; x - 1,$$

with equality if and only if $x{=}1$. Confirm this result by differentiation of $\log(x){-}(x - 1)$.

(b) Use the previous result to prove that $\mathrm{KL}(p, q) \;\geq\; 0$, with equality if and only if the two distributions p_i and q_i are equal. *(4 pts)*

(c) Using the inequality in (a), as well as the simple equality $\log x = 2 \log \sqrt{x}$, derive the tighter lower bound: *(2 pts)*

$$\mathrm{KL}(p, q) \;\geq\; \sum_i (\sqrt{p_i} - \sqrt{q_i})^2.$$

(d) Provide a counterexample to show that the KL distance is not a symmetric function of its arguments: *(1 pts)*

$$\mathrm{KL}(p, q) \;\neq\; \mathrm{KL}(q, p).$$

Despite this asymmetry, it is still common to refer to $\mathrm{KL}(p, q)$ as a measure of distance. Many algorithms for machine learning are based on minimizing KL distances between probability distributions.

1.6 Mutual information *(3 pts)*

The mutual information $I(X, Y)$ between two discrete random variables X and Y is defined as

$$I(X, Y) \;=\; \sum_x \sum_y P(x, y) \log \left[\frac{P(x, y)}{P(x)\, P(y)} \right],$$

where the sum is over all possible values of the random variables X and Y. Note how the mutual information is related to the definitions in the previous two problems.

(a) Show that the mutual information is nonnegative. (*Hint:* use the result of the previous problem.) *(2 pts)*

(b) Show that the mutual information $I(X, Y)$ vanishes if and only if X and Y are independent random variables. (Thus, $I(X, Y)$ provides one quantitative measure of dependence between X and Y.) *(1 pts)*

1.7 Hangman *(13 pts)*

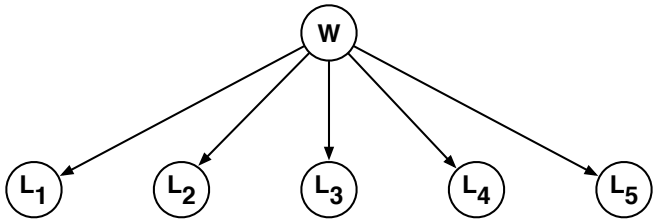
Consider the belief network shown below, where the random variable W stores a five-letter word and the random variable $L_i \in \{\texttt{A}, \texttt{B}, \ldots, \texttt{Z}\}$ reveals only the word's ith letter. Also, suppose that these five-letter words are chosen at random from a large corpus of text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

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

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:

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

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:

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

And the first term we obtain from Bayes rule:

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

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

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:

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

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:

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

Note in particular how the first computation feeds into the second. Your assignment in this problem is implement both of these calculations.

Click on the Open in Colab button in the README of the GitHub repository. This should open a new session in Colab with all the repository files. The starter codebase has the following files:

hw1_word_counts_05.txt: A list of 5-letter words (including names and proper nouns) and their counts from a large corpus of Wall Street Journal articles (roughly three million sentences).

hangman.py: We have provided for you a game engine which does the heavy lifting: opening the word list, parsing the frequencies, and running the game.

cse150a_hangman.ipynb: This notebook is for you to fill in.

(a) Let's first get a feel of how the game works. Feel free to play a few games of hangman on your own by executing the corresponding cell. *(0 pts)*

(b) In the next cell, get a feel for how the system works: implement a random play mode, which returns a random letter guess out of the remaining letters in the alphabet that hasn't been guessed yet. **(2pts)**

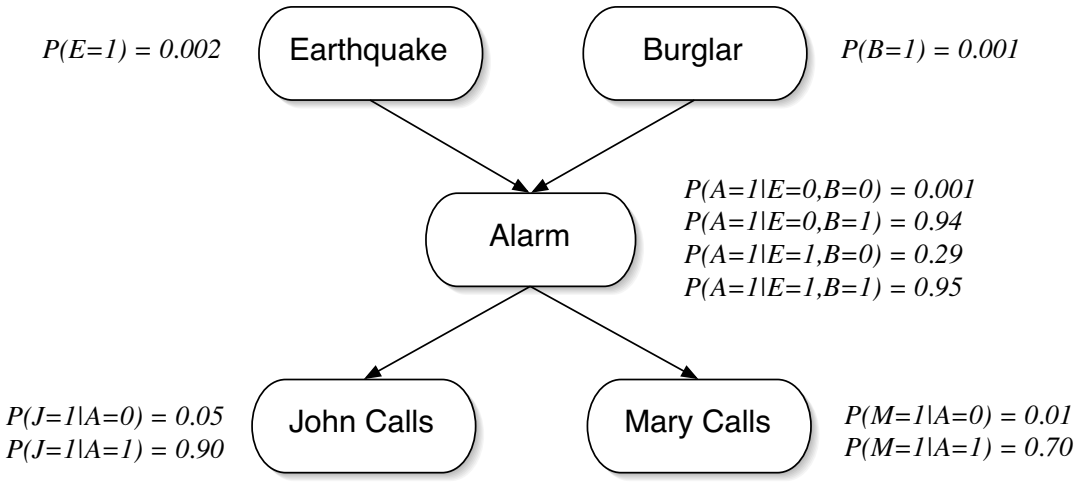
(c) Let's get to the fun part! Implement the bayesian inference function in the next cell, which should predict the next most likely letter based on the word list, with details from the above writeup. The documentation is provided for you in the cell already. *8pts*

(d) Now, we want to measure how well your code does in playing hangman. Design a benchmark function (documentation provided in the cell) which will be able to benchmark both your random and Bayesian inference methods.
Note: Our reference solution on almost all runs at least 94% accuracy over 1000 tries, on average. That means that your benchmark should correctly demonstrate accuracy of at least 93% accuracy on 1000 tries to achieve full credit. *(3pts)*

(e) Export the notebook based on the export instructions, and make sure your output is included. Please upload your homework as a **single** PDF including the writeup (scanned or typed in LaTeX) and the PDF output of the exported notebook. Format errors will have up to 2 points deducted. *(0pts)*

2.1 Probabilistic inference (*12 pts — 2 pts each)*

Recall the alarm belief network described in class. The directed acyclic graph (DAG) and conditional probability tables (CPTs) are shown below:



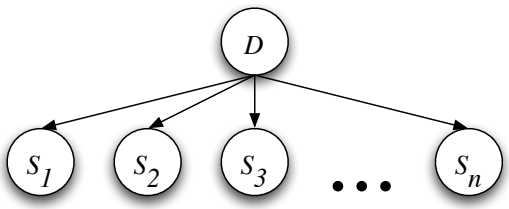
Compute numeric values for the following probabilities, exploiting relations of marginal and conditional independence as much as possible to simplify your calculations. You may re-use numerical results from lecture, but otherwise *show your work*. Be careful not to drop significant digits in your answer.

(a) $P(E=1|A=1)$ (c) $P(A=1|M=1)$ (e) $P(A=1|M=0)$
(b) $P(E=1|A=1,B=0)$ (d) $P(A=1|M=1,J=0)$ (f) $P(A=1|M=0,B=1)$

Consider your results in (b) versus (a), (d) versus (c), and (f) versus (e). Do they seem consistent with commonsense patterns of reasoning?

2.2 Probabilistic reasoning *(4 pts)*

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:

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

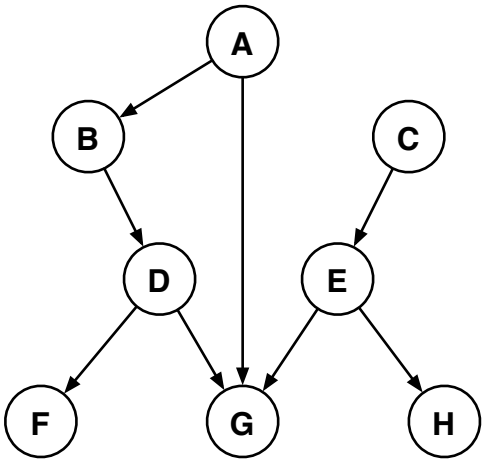
If this ratio is greater than 1, the doctor diagnoses the patient with the $D=0$ form of the disease; otherwise, with the $D=1$ form.

(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. *(3 pts)*

(b) Does the diagnosis become more or less certain as more symptoms are observed? Explain. (1 pts)

2.3 True or false *(10 pts)*

For the belief network shown below, indicate whether the following statements of marginal or conditional independence are **true (T)** or **false (F)**. Your answer will be **only** graded based on correctness. No justifications required.



_____	$P(B\|G,C)$	=	$P(B\|G)$
_____	$P(F,G\|D)$	=	$P(F\|D)\,P(G\|D)$
_____	$P(A,C)$	=	$P(A)\,P(C)$
_____	$P(D\|B,F,G)$	=	$P(D\|B,F,G,A)$
_____	$P(F,H)$	=	$P(F)\,P(H)$
_____	$P(D,E\|F,H)$	=	$P(D\|F)\,P(E\|H)$
_____	$P(F,C\|G)$	=	$P(F\|G)\,P(C\|G)$
_____	$P(D,E,G)$	=	$P(D)\,P(E)\,P(G\|D,E)$
_____	$P(H\|C)$	=	$P(H\|A,B,C,D,F)$
_____	$P(H\|A,C)$	=	$P(H\|A,C,G)$

2.4 More on Belief Networks *(10 pts)*

For the belief network shown below, indicate whether the following statements of marginal or conditional independence are **true (T)** or **false (F)**. Your answer will be **only** graded based on correctness. No justifications required.

(a) _____ $P(F|H) = P(F|C,H)$

(b) _____ $P(E|A,B) = P(E|A,B,F)$

(c) _____ $P(E,F|B,G) = P(E|B,G)\,P(F|B,G)$

(d) _____ $P(F|B,C,G,H) = P(F|B,C,E,G,H)$

(e) _____ $P(A,B|D,E,F) = P(A,B|D,E,F,G,H)$

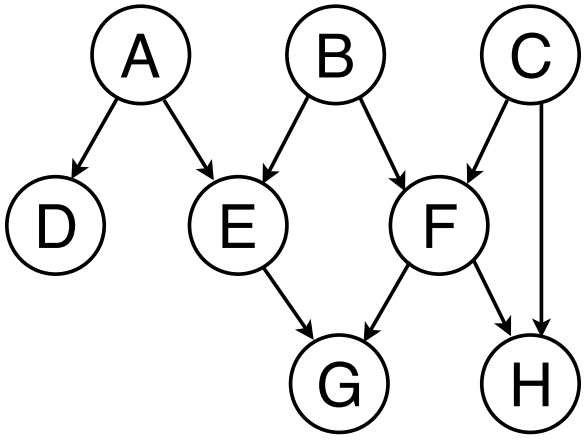
(f) _____ $P(D,E,F) = P(D)\,P(E|D)\,P(F|E)$

(g) _____ $P(A|F) = P(A)$

(h) _____ $P(E,F) = P(E)\,P(F)$

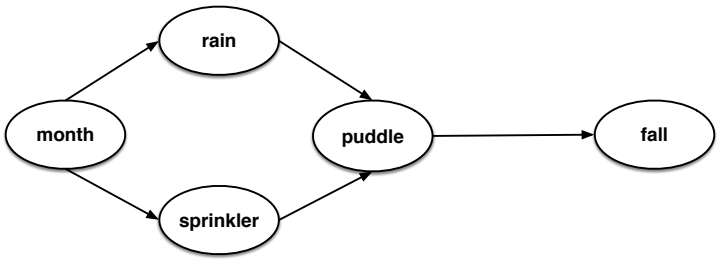
(i) _____ $P(D|A) = P(D|A,E)$

(j) _____ $P(B,C) = P(B)\,P(C)$



2.5 Conditional independence *(8 pts)*

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`.

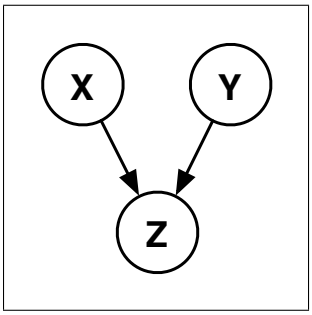


List all the conditional independence relations that must hold in any probability distribution represented by this DAG. More specifically, list all tuples $\{X, Y, E\}$ such that $P(X, Y|E) = P(X|E)P(Y|E)$, where

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

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 $E = \emptyset$?

2.6 Noisy-OR *(7 pts)*



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)	$P(X=1)\,P(Y=1)\,P(Z=1)$		$P(X=1,Y=1,Z=1)$

*Challenge (***optional***):* for each case, prove rigorously the correctness of your answer. You will receive partial credits for your efforts on proofs.