Assignment 2: Bayesian Networks

Machine Learning

Fall 2019

Learning Objectives

- The concept of independence and conditional independence
- The basic components of a Bayesian Networks (BN)
- The rules of d-separation to compute conditional independence relationships in a BN
- The Compas recidivism risk algorithm controversy

1 *COMPAS Model of Recidivism*

Notice

We are going to think about race and criminal justice in the United States. Before we dive into this, we want to acknowledge:

- This is a complex and intricate issue that involves policy, society, technology, individual beliefs/values, and history. This topic directly (but not equally) impacts the lives of many people.
- We all have our own lenses through which we view the world.
- In this class, we will scratch the surface of the way the US justice system works. Your instructors are not criminal justice experts, but they do care about this topic. We are also continuing to learn more.
- This topic will likely be uncomfortable to grapple with regardless of your background and identity. It may resonate differently with each of us. We (Sam/Carrie) are available in person and via email, if you would like to discuss how we can best support you in class. We are planning to have some group discussion in class. One method of support could be pairing you with a partner of your choosing for this discussion. Another could be including your ideas about how class discussion can be informative and challenging without creating unnecessary pain. Please reach out to us if you have any concerns or want to discuss this more.

A few basics about the US criminal criminal justice system.

A police officer can place a person under arrest. However, an arrest does not necessarily mean that person committed a crime (both in fact and in a legal sense). Legally, someone is considered innocent until proven guilty in court. However, arrested people are often held in jail for months before trail (this is called pretrial detention). To get out of jail before trial, the arrested person can post bail. Bail is a considerable amount of money ("money bond") that is given to the court to ensure the person shows up to trial. As you might guess, bail represents a way people with money are treated differently by the system than people without money. (Optional: For more on bail, listen to Episode 62 of the podcast Ear Hustle).

Legally, a person is considered guilty if they are convicted in court. Practically, innocent people are sometimes convicted. People with a lot of money can hire many lawyers who will work many days, weeks, or years, fighting the case.

People without funds for a lawyer will be assigned a public defender. Public defenders are often overwhelmed, and might have just a minute to look over the details of a case, right before trail.

If you are interested in this topic and want to hear stories from incarcerated people, we recommend listening to the podcast Ear Hustle. (This is optional)

Exercise 1 (20 minutes)

(a) (Read the Report of The Sentencing Project to the United Nations Special Rapporteur on Contemporary Forms of Racism, Racial Discrimination, Xenophobia, and Related Intolerance.

This article is intended to provide some background information on criminal justice and race in the US.

Exercise 2 (90 minutes)

In this module, we'll be spending time talking about the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) algorithm, produced by the company Northpointe, Inc. COMPAS was intended to assess the risk of recidivism. This is a well-known algorithm in machine learning communities.

Below we will provide you a list of readings on this topic. As you read, please prepare to reflect on the following themes:

- Social justice in a non-ideal world.
- The roles of machine learning engineers and the roles of other professional roles in the application of machine learning in our society.
- The major things that you would want to consider in this type of undertaking.
- Framing things mathematically versus from a social justice standpoint.
- The level of technical debate in model choices that is brought up in this discussion (especially in the technical responses).

(a) First, read the ProPublica article.

(b) Next, you'll read the technical details of the ProPublica analysis.

We would like to be clear about when the COMPAS metric is applied. It is applied after someone is arrested, and the prediction COMPAS gives is if that person will be arrested again. A person who was arrested twice could be legally and actually innocent. It is important to be clear in our language that these are arrested people, and not convicted criminals. The term recidivism, which is generally defined as a criminal who commits a second crime, is used in the readings. ProPublica actually redefines recidivism as "a criminal offense that resulted in a jail booking and took place after the crime for which the person was COMPAS scored." Here, ProPublica conflates an arrest for a crime with a conviction for that crime. Note that a jail booking is pre-trial, and different than prison, which is post-trial and conviction. This language is also used by the Northpointe rebuttal. Carrie would argue that since these are not people who have been yet convicted of crimes, they are not necessarily recidivists, and the language in the readings should be corrected.

Please read How We Analyzed the COMPAS Recidivism Algorithm.

(c) Please read the Northpointe rebuttal. This is a long reading. It has a lot of jargon, and some acronyms are not defined. We suggest limiting yourself to 60 minutes for this reading. You may consider working with a classmate on this reading, so you can both discuss what you think the author is saying. We're including this whole reading here because we would like you to engage with the real-world material.

Exercise 3 (30 minutes)

(a) Please summarize what you see as the key parts of the ProPublica and Northpointe cases. You can use words, diagrams, concept maps, or another method that works for you.

(b) Reflect on what you've just read. We think the themes brought up above will provide good fodder for your response, but please feel free to take it in any direction. Aim for around two paragraphs in your response.

2 Motivation and Context

In the last assignment we learned the basic definition of a probability and acquired some very powerful rules for working with probabilities. In this assignment you'll be taking these ideas and extending them in the following significant ways.

- You'll learn a graphical way to represent the relationships between probabilities that will make it much easier to work with large probabilistic models.
- You'll be taking the ideas of probability using them to derive a whole new way of approaching the classification problem in machine learning.

3 Product Rule and Marginalization for Random Variables

Recall: Product Rule and Marginalization for Events

Last assignment we learned about two very powerful techniques for computing the probability of events.

- We learned the product rule (or conjunction rule), which states that for any two events $\mathcal{A}$ and $\mathcal{B}$,

$$\begin{aligned} p(\mathcal{A},\mathcal{B}) &= p(\mathcal{A})p(\mathcal{B}|\mathcal{A}) \\ &= p(\mathcal{B})p(\mathcal{A}|\mathcal{B}) \ . \end{aligned} \tag{1}$$

- We learned the rule of marginalization, which states that for any two events $\mathcal{A}$ and $\mathcal{B}$,

$$p(\mathcal{A}) = p(\mathcal{A},\mathcal{B}) + p(\mathcal{A},\neg\mathcal{B}) \ . \tag{2}$$

It turns out that these rules can modified slightly to apply to random variables as well (instead of just events).

3.1 Product Rule for Random Variables

Suppose we have two random variables X and Y. If we want to know the probability of random variable X taking on value x (it is common to use a lower case letter to refer to a particular value of a random variable) and random variable

Y simultaneously taking on value y, we can decompose the joint probability (the probability of both of these things occurring simultaneously) using the product rule.

$$\begin{aligned} p(X = x, Y = y) &= p(X = x)p(Y = y|X = x) \qquad\qquad \text{or equivalently,} \\ &= p(Y = y)p(X = x|Y = y) \end{aligned} \tag{3}$$

Notice that this looks pretty much identical to Equation **??** except that instead of referencing whether an event happens, we are now referencing a random variable taking on a particular value.

⚠ Notice

It's very common to use the shorthand $p(x, y)$ to refer to $p(X = x, Y = y)$. The motivation for this shorthand is that it is obvious from the context that $p(x, y)$ really means the probability of random variable X taking on value x and random variable Y taking on value y. In this assignment we're going to avoid using this shorthand, but we will start using the shorthand in future assignments (we'll warn you when we start using it). Also, you may see this notation used in external resources, so it helps to know about it.

3.2 Marginalization for Random Variables

Again, suppose we have two random variables X and Y. We are interested in computing $p(X = x)$, but suppose it is difficult to compute this probability directly. Just as we did for events in the last assignment, we can compute $p(X = x)$ by marginalizing out the random variable Y. For simplicity, let's assume that Y can only take on integer values from 1 to k. We can write the marginal distribution $p(X = x)$ in the following way.

$$p(X = x) = \sum_{y=1}^{k} p(X = x, Y = y) \tag{4}$$

You should notice that this equation is very similar to Equation **??** except that instead of summing over the probability for the two possible outcomes with respect to the event $\mathcal{B}$ (i.e., $\mathcal{B}$ either happens or it does not), we are now summing over the k possible values that Y could take. Random variables don't necessarily have to take on values from 1 to k. In general, if the random variable Y can take on any value from some discrete set of values $\mathcal{Y}$ (we are using the calligraphic font because we are referring to a set), then the marginal distribution of X can be written as:

$$p(X = x) = \sum_{y \in \mathcal{Y}} p(X = x, Y = y)\ \ . \tag{5}$$

Notice that Equation **??** is a special case of Equation **??** where $\mathcal{Y} = \{1, 2, \ldots, k\}$.

Exercise 4 (15 minutes)

This exercise is from the Wikipedia article on marginal distribution.

Notice

Strictly speaking a random variable only takes on real numbers as its value (i.e., a random variable cannot take on the value of “blue”). That said, it’s common to see folks bend this rule a bit by allowing the random variable to take on values that are not real numbers (e.g., you will see that in the Wikipedia article linked above). Allowing for random variables to have non-numeric values doesn’t really change the math except it makes it meaningless to compute quantities such as the mean. Since in many cases allowing random variables to take on non-numeric values will make things clearer, we will allow this in our materials.

Suppose you want to compute the probability that a pedestrian will be hit by a car, while crossing the road at a pedestrian crossing, without paying attention to the traffic light (a bit morbid, we know). Let H be a discrete random variable taking on the value “hit” if the pedestrian is struck and “not hit” if the pedestrian makes it safely across. Let L (for traffic light) be a discrete random variable taking on the value “red” when the light is red, “yellow” when the light is yellow, and “green” when the light is green.
The model that governs the prior probability of the light (L) is as follows.

$$\begin{aligned} p(L = \text{red}) &= 0.2 \\ p(L = \text{yellow}) &= 0.1 \\ p(L = \text{green}) &= 0.7 \end{aligned} \tag{6}$$

The model that governs the conditional probability of H given L is as follows.

$$\begin{aligned} p(H = \text{hit}|L = \text{red}) &= 0.01 \\ p(H = \text{not hit}|L = \text{red}) &= 0.99 \qquad \text{Note: the probability of “not hit” is always 1 - probability of hit} \\ p(H = \text{hit}|L = \text{yellow}) &= 0.1 \\ p(H = \text{not hit}|L = \text{yellow}) &= 0.9 \\ p(H = \text{hit}|L = \text{green}) &= 0.8 \\ p(H = \text{not hit}|L = \text{green}) &= 0.2 \end{aligned}$$

What is $p(H = \text{hit})$?

Solution

$$\begin{aligned} p(H = \text{hit}) &= p(H = \text{hit}, L = \text{red}) + p(H = \text{hit}, L = \text{yellow}) + p(H = \text{hit}, L = \text{green}) && \text{marginalization} \\ &= p(L = \text{red})p(H = \text{hit}|L = \text{red}) + p(L = \text{yellow})p(H = \text{hit}|L = \text{yellow}) && \text{product rule} \\ &\qquad + p(L = \text{green})p(H = \text{hit}|L = \text{green}) \\ &= (0.2 \times 0.01) + (0.1 \times 0.1) + (0.7 \times 0.8) \\ &= 0.572 \end{aligned}$$

4 Some Twists on Bayes' Rule

By now hopefully you are starting to feel comfortable with the vanilla form of Bayes' rule. There are a few quite useful variants that we'd like to point out. There are no exercises for you to do here, just add these to your bag of tricks (you'll be leveraging them later in this assignment, so you'll have a chance to solidify them then).

$$p(\mathcal{A}|\mathcal{B}) = \frac{p(\mathcal{B}|\mathcal{A})p(\mathcal{A})}{p(\mathcal{B})} \qquad \text{as a reminder, here is vanilla Bayes' rule} \tag{7}$$

$$p(\mathcal{A},\mathcal{B}|\mathcal{C}) = \frac{p(\mathcal{C}|\mathcal{A},\mathcal{B})p(\mathcal{A},\mathcal{B})}{p(\mathcal{C})} \qquad \text{you can bring over multiple events} \tag{8}$$

$$p(\mathcal{A}|\mathcal{B},\mathcal{C}) = \frac{p(\mathcal{B}|\mathcal{A},\mathcal{C})p(\mathcal{A}|\mathcal{C})}{p(\mathcal{B}|\mathcal{C})} \qquad \text{you can leave an event to the right of the conditioning bar} \tag{9}$$

5 Independence and Conditional Independence

Two of the most important concepts in probability theory are independence and the closely related concept of conditional independence. These ideas are important because they let you analyze probabilistic quantities in isolation. For instance, if you know that two events that you are interested in predicting are independent of each other, then you can make a model of each event in isolation. Modeling events independently can make your life much easier since you don't have to consider how the two events interact. Next, we'll make this high-level idea precise.

5.1 Independence

The product rule of probability can be simplified when two events, $\mathcal{A}$ and $\mathcal{B}$ are independent. As an example, suppose $\mathcal{A}$ represents the event that the first flip of a coin comes up heads and event $\mathcal{B}$ is the event that the second flip of the same coin comes up heads. Since whether or not $\mathcal{A}$ occurs tells us nothing about whether $\mathcal{B}$ would occur, we say that $\mathcal{A}$ and $\mathcal{B}$ are independent events (we use the notation $\mathcal{A} \perp\!\!\!\perp \mathcal{B}$ to indicate that $\mathcal{A}$ is independent of $\mathcal{B}$). An event $\mathcal{A}$ is independent of another event $\mathcal{B}$ if and only if the following condition holds.

$$p(\mathcal{A},\mathcal{B}) = p(\mathcal{A})p(\mathcal{B}) \tag{10}$$

A direct consequence of Equation **??** is that if $\mathcal{A} \perp\!\!\!\perp \mathcal{B}$, then

$$\begin{aligned} p(\mathcal{A}|\mathcal{B}) &= p(\mathcal{A}) && \text{and} \\ p(\mathcal{B}|\mathcal{A}) &= p(\mathcal{B}) \ . \end{aligned}$$

A very similar equation to Equation **??** can be defined for random variables. Two random variables X and Y are independent if and only if the following condition holds for any values x and y.

$$p(X = x, Y = y) = P(X = x)p(Y = y) \tag{11}$$

Similar to the rule for events, $p(X = x|Y = y) = P(X = x)$ if $X \perp\!\!\!\perp Y$.

Exercise 5 (10 minutes)

(a) Provide at least 3 examples of events or random variables that are independent of each other.

☆ Solution

- The event that a coin comes up heads on the first throw and the event that the coin comes up heads on the second throw.
- A random variable that represents that last digit on a car's license plate and a random variable that represents the last digit on another car's license plate.
- The event that captures whether or not it rains tomorrow in Boston and a random variable that represents the number of people who attend a rock concert in California tomorrow night (at least it seems that these things are unrelated).

(b) Provide at least 3 examples of events or random variables that are not independent of each other.

☆ Solution

Here are some ideas.

- The daily increase in the Dow Jones Industrial average and the daily increase in the NASDAQ.
- The event that the American League wins the World Series in 2019 and the event that the National League wins the World Series in 2019.
- Testing positive for a disease and having that disease.

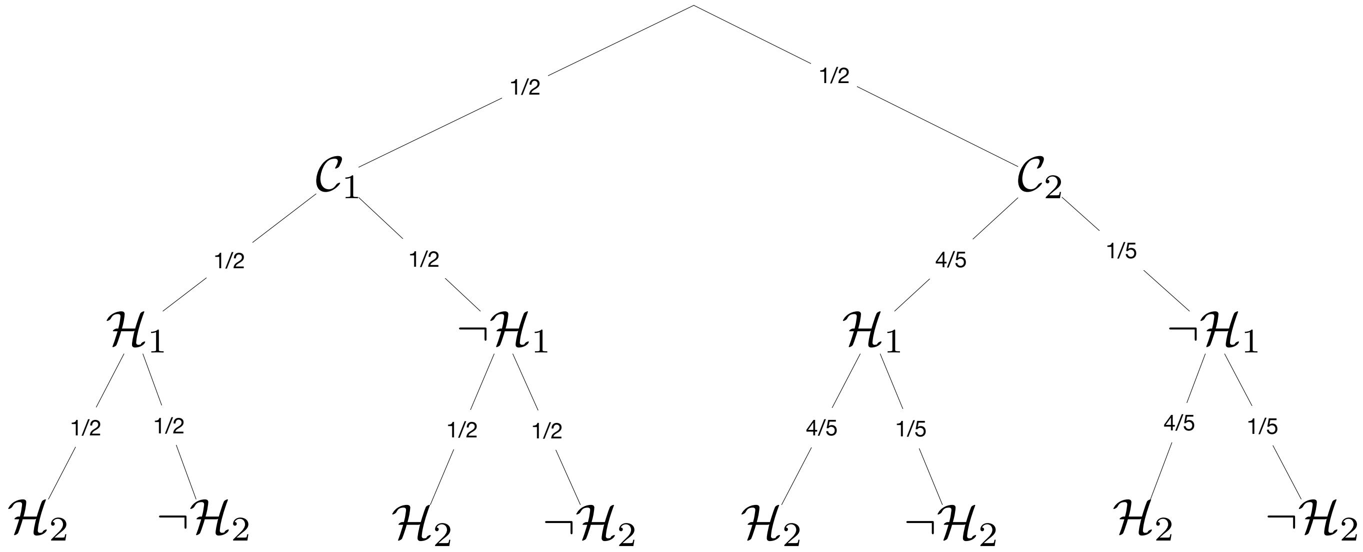
5.2 Conditional Independence

Sometimes two events (or two random variables) that are not independent might become independent when conditioned on another event.

Exercise 6 (20 minutes)

As a motivating example for the concept of conditional independence, consider a variant of the coin problem we saw last assignment. A bag contains two coins. Suppose we choose one of the two coins with equal probability. Let $\mathcal{C}_1$ represents the event that we choose coin 1 and $\mathcal{C}_2$ represent the event that we choose coin 2. Coin 1 is fair $p(\mathcal{H}|\mathcal{C}_1) = \frac{1}{2}$. Coin 2 is not fair ($p(\mathcal{H}|\mathcal{C}_2) = \frac{4}{5}$). We then flip the coin twice (we don't pick a new coin for the second flip). Let $\mathcal{H}_1$ represent the event that the first flip comes up heads and $\mathcal{H}_2$ represent the event that the second flip comes up heads. Are $\mathcal{H}_1$ and $\mathcal{H}_2$ independent (i.e., is $\mathcal{H}_1 \perp\!\!\!\perp \mathcal{H}_2$)?

To help you get started, here is a tree diagram illustrating the problem.



Given the tree diagram above, is $\mathcal{H}_1 \perp\!\!\!\perp \mathcal{H}_2$?

☆ Solution

In order to test $\mathcal{H}_1 \perp\!\!\!\perp \mathcal{H}_2$ we need to check the following condition:

$$p(\mathcal{H}_1, \mathcal{H}_2) \overset{?}{=} p(\mathcal{H}_1)p(\mathcal{H}_2) \tag{12}$$

We can compute each of the terms in the preceding equation using the tree diagram. In total there are 8 possible paths through the tree. Recall that we can find the probability of a path by multiplying the numbers on the arrows. To find the probability of a particular event, say $p(\mathcal{H}_1)$ we just add up the probability of all of the paths that include $\mathcal{H}_1$. We can apply this technique to each of the events we care about.

$$\begin{aligned}
p(\mathcal{H}_1) &= p(\mathcal{C}_1, \mathcal{H}_1, \mathcal{H}_2) + p(\mathcal{C}_1, \mathcal{H}_1, \neg\mathcal{H}_2) + p(\mathcal{C}_2, \mathcal{H}_1, \mathcal{H}_2) + p(\mathcal{C}_2, \mathcal{H}_1, \neg\mathcal{H}_2) \\
&= \left(\frac{1}{2} \times \frac{1}{2} \times \frac{1}{2}\right) + \left(\frac{1}{2} \times \frac{1}{2} \times \frac{1}{2}\right) + \left(\frac{1}{2} \times \frac{4}{5} \times \frac{4}{5}\right) + \left(\frac{1}{2} \times \frac{4}{5} \times \frac{1}{5}\right) \\
&= \frac{13}{20} \\
p(\mathcal{H}_2) &= p(\mathcal{C}_1, \mathcal{H}_1, \mathcal{H}_2) + p(\mathcal{C}_1, \neg\mathcal{H}_1, \mathcal{H}_2) + p(\mathcal{C}_2, \mathcal{H}_1, \mathcal{H}_2) + p(\mathcal{C}_2, \neg\mathcal{H}_1, \mathcal{H}_2) \\
&= \left(\frac{1}{2} \times \frac{1}{2} \times \frac{1}{2}\right) + \left(\frac{1}{2} \times \frac{1}{2} \times \frac{1}{2}\right) + \left(\frac{1}{2} \times \frac{4}{5} \times \frac{4}{5}\right) + \left(\frac{1}{2} \times \frac{1}{5} \times \frac{4}{5}\right) \\
&= \frac{13}{20} \\
p(\mathcal{H}_1, \mathcal{H}_2) &= p(\mathcal{C}_1, \mathcal{H}_1, \mathcal{H}_2) + p(\mathcal{C}_2, \mathcal{H}_1, \mathcal{H}_2) \\
&= \left(\frac{1}{2} \times \frac{1}{2} \times \frac{1}{2}\right) + \left(\frac{1}{2} \times \frac{4}{5} \times \frac{4}{5}\right) \\
&= \frac{89}{200} \\
p(\mathcal{H}_1)p(\mathcal{H}_2) &= \frac{13}{20} \times \frac{13}{20} = \frac{169}{400} \neq \frac{89}{200} = p(\mathcal{H}_1, \mathcal{H}_2)
\end{aligned}$$

Since $p(\mathcal{H}_1, \mathcal{H}_2) \neq p(\mathcal{H}_1)p(\mathcal{H}_2)$, $\mathcal{H}_1$ is not independent of $\mathcal{H}_2$.

It turns out that even though $\mathcal{H}_1$ and $\mathcal{H}_2$ are not independent, they are what's called *conditionally independent* given $\mathcal{C}_1$ (or $\mathcal{C}_2$). Formally, events $\mathcal{A}$ and $\mathcal{B}$ are considered conditionally independent given $\mathcal{C}$ (written $\mathcal{A} \perp\!\!\!\perp \mathcal{B} \mid \mathcal{C}$) if and only if

$$p(\mathcal{A}, \mathcal{B}|\mathcal{C}) = p(\mathcal{A}|\mathcal{C})p(\mathcal{B}|\mathcal{C})$$

Exercise 7 (20 minutes)

(a) Show that $\mathcal{H}_1 \perp\!\!\!\perp \mathcal{H}_2 \mid \mathcal{C}_1$

☆ Solution

We need to show that $p(\mathcal{H}_1, \mathcal{H}_2|\mathcal{C}_1) = p(\mathcal{H}_1|\mathcal{C}_1)p(\mathcal{H}_2|\mathcal{C}_1)$. We can use the tree diagram to compute these conditional probabilities by starting our multiplication after the branch that we are conditioning on.

$$\begin{aligned}
p(\mathcal{H}_1, \mathcal{H}_2|\mathcal{C}_1) &= \frac{1}{2} \times \frac{1}{2} = \frac{1}{4} \\
p(\mathcal{H}_1|\mathcal{C}_1) &= p(\mathcal{H}_1, \mathcal{H}_2|\mathcal{C}_1) + p(\mathcal{H}_1, \neg\mathcal{H}_2|\mathcal{C}_1) \\
&= \left(\frac{1}{2} \times \frac{1}{2}\right) + \left(\frac{1}{2} \times \frac{1}{2}\right) = \frac{1}{2} \\
p(\mathcal{H}_2|\mathcal{C}_1) &= p(\mathcal{H}_1, \mathcal{H}_2|\mathcal{C}_1) + p(\neg\mathcal{H}_1, \mathcal{H}_2|\mathcal{C}_1) \\
&= \left(\frac{1}{2} \times \frac{1}{2}\right) + \left(\frac{1}{2} \times \frac{1}{2}\right) = \frac{1}{2} \\
p(\mathcal{H}_1, \mathcal{H}_2|\mathcal{C}_1) &= \frac{1}{2} \times \frac{1}{2} \\
&= p(\mathcal{H}_1|\mathcal{C}_1)p(\mathcal{H}_2|\mathcal{C}_1)
\end{aligned}$$

(b) Show that $\mathcal{H}_1 \perp\!\!\!\perp \mathcal{H}_2 \mid \mathcal{C}_2$

☆ Solution

We need to show that $p(\mathcal{H}_1, \mathcal{H}_2|\mathcal{C}_2) = p(\mathcal{H}_1|\mathcal{C}_2)p(\mathcal{H}_2|\mathcal{C}_2)$. We can use the tree diagram to compute these conditional probabilities by starting our multiplication after the branch that we are conditioning on.

$$\begin{aligned}
p(\mathcal{H}_1, \mathcal{H}_2|\mathcal{C}_2) &= \frac{4}{5} \times \frac{4}{5} = \frac{16}{25} \\
p(\mathcal{H}_1|\mathcal{C}_2) &= p(\mathcal{H}_1, \mathcal{H}_2|\mathcal{C}_2) + p(\mathcal{H}_1, \neg\mathcal{H}_2|\mathcal{C}_2) \\
&= \left(\frac{4}{5} \times \frac{4}{5}\right) + \left(\frac{4}{5} \times \frac{1}{5}\right) = \frac{4}{5} \\
p(\mathcal{H}_2|\mathcal{C}_2) &= p(\mathcal{H}_1, \mathcal{H}_2|\mathcal{C}_2) + p(\neg\mathcal{H}_1, \mathcal{H}_2|\mathcal{C}_2) \\
&= \left(\frac{4}{5} \times \frac{4}{5}\right) + \left(\frac{1}{5} \times \frac{4}{5}\right) = \frac{4}{5} \\
p(\mathcal{H}_1, \mathcal{H}_2|\mathcal{C}_1) &= \frac{4}{5} \times \frac{4}{5} \\
&= p(\mathcal{H}_1|\mathcal{C}_2)p(\mathcal{H}_2|\mathcal{C}_2)
\end{aligned}$$

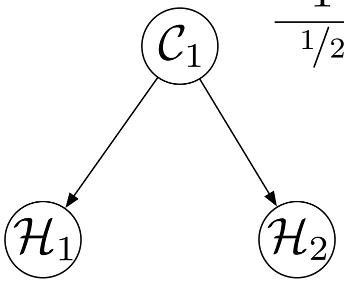
The definition of the conditional independence of events extends to random variables. Random variables X and Y are conditionally independent given random variable Z (i.e., $X \perp\!\!\!\perp Y \mid Z$) if and only if the following holds for all x, y, z.

$$p(X = x, Y = y|Z = z) = p(X = x|Z = z)p(Y = y|Z = z) \tag{13}$$

6 Bayesian Networks

The calculations in the previous section were a bit tedious. It would be great if there was some way to reason about the conditional independence properties of two random variables conditioned on some other random variable. Luckily... drum roll... there is! A Bayesian network (sometimes called a Bayesian belief network or a probabilistic directed acyclic graphical model) represents the conditional independence relationships between random variables through a graphical, causal structure. We'll use BN as shorthand for "Bayesian network." Take for instance, the BN that represents the coin problem that we did in the last section.

$\mathcal{C}_1$	
T	F
$1/2$	$1/2$



	$\mathcal{H}_1$	
$\mathcal{C}_1$	T	F
T	$1/2$	$1/2$
F	$4/5$	$1/5$

	$\mathcal{H}_2$	
$\mathcal{C}_1$	T	F
T	$1/2$	$1/2$
F	$4/5$	$1/5$

The graphical structure (edges and nodes in the graph) tell us everything we need to infer the conditional independence properties in the graph (Note that we haven't told you *how* you can extract the conditional independence properties from the graph; that's coming later in the assignment). The tables by each node provide the probability of the event conditioned on whether or not the node's parents (a "parent" of a node, A, is a node B where there is an edge pointing from A to B) happened (T stands for *True* or that the event happens and F stands for *False* or that the event doesn't happen).

The BN provides us with a way of computing any relevant probability (e.g., marginal, conditional, joint) for the nodes in the network. The condition that must hold for any BN is that if we write the joint distribution of all of the random variables (or events, the relationship is the same for either) in the network, it must factorize in the following way (we'll use $X_1, X_2, \dots X_n$ to represent random variables in the network and we'll define the function $Pa(X_i)$ to return all of the random variables that are parents of X_i).

$$p(X_1, X_2, \dots, X_n) = p(X_1|Pa(X_1)) \times p(X_2|Pa(X_2)) \times \dots p(X_n|Pa(X_n)) \tag{14}$$

Back to our coin BN, this means that we can write the joint distribution like so.

$$p(\mathcal{C}_1, \mathcal{H}_1, \mathcal{H}_2) = p(\mathcal{C}_1)p(\mathcal{H}_1|\mathcal{C}_1)p(\mathcal{H}_2|\mathcal{C}_1) \tag{15}$$

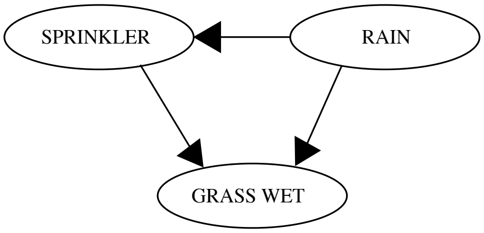
✔ Understanding Check

Make sure you understand how we arrived at Equation **??**. Refer back to Equation **??** and hopefully you will see the connection.

Exercise 8 (20 minutes)

Consider the BN below (source: https://en.wikipedia.org/wiki/Bayesian_network#Example).

	SPRINKLER	
RAIN	T	F
F	0.4	0.6
T	0.01	0.99



RAIN	
T	F
0.2	0.8

		GRASS WET	
SPRINKLER	RAIN	T	F
F	F	0.0	1.0
F	T	0.8	0.2
T	F	0.9	0.1
T	T	0.99	0.01

Compute the following probabilities. For brevity we'll use the first letter of each node to indicate that the corresponding event happens (e.g., we'll use $\mathcal{R}$ to refer to the event "rain").

(a) $p(\mathcal{R}, \mathcal{G}, \neg\mathcal{S})$

☆ Solution

$$\begin{aligned}p(\mathcal{R}, \mathcal{G}, \neg\mathcal{S}) &= p(\mathcal{R})p(\neg\mathcal{S}|\mathcal{R})p(\mathcal{G}|\mathcal{R}, \neg\mathcal{S}) \\ &= 0.2 \times 0.99 \times 0.8 \\ &= 0.1584\end{aligned}$$

(b) $p(\mathcal{R})$

☆ Solution

This one is kind of a trick question. Since $\mathcal{R}$ has no parents, we can just read the probability right off the probability table for $\mathcal{R}$. The answer is 0.2.

(c) $p(\neg\mathcal{G}, \neg\mathcal{S})$ (hint: marginalize over $\mathcal{R}$)

☆ Solution

$$\begin{aligned}
p(\neg\mathcal{G}, \neg\mathcal{S}) &= p(\neg\mathcal{G}, \neg\mathcal{S}, \mathcal{R}) + p(\neg\mathcal{G}, \neg\mathcal{S}, \neg\mathcal{R}) \\
&= p(\mathcal{R})p(\neg\mathcal{S}|\mathcal{R})p(\neg\mathcal{G}|\mathcal{R}, \neg\mathcal{S}) + p(\neg\mathcal{R})p(\neg\mathcal{S}|\neg\mathcal{R})p(\neg\mathcal{G}|\neg\mathcal{R}, \neg\mathcal{S}) \\
&= (0.2 \times 0.99 \times 0.2) + (0.8 \times 0.6 \times 1.0) \\
&= 0.5196
\end{aligned}$$

6.1 D-separation

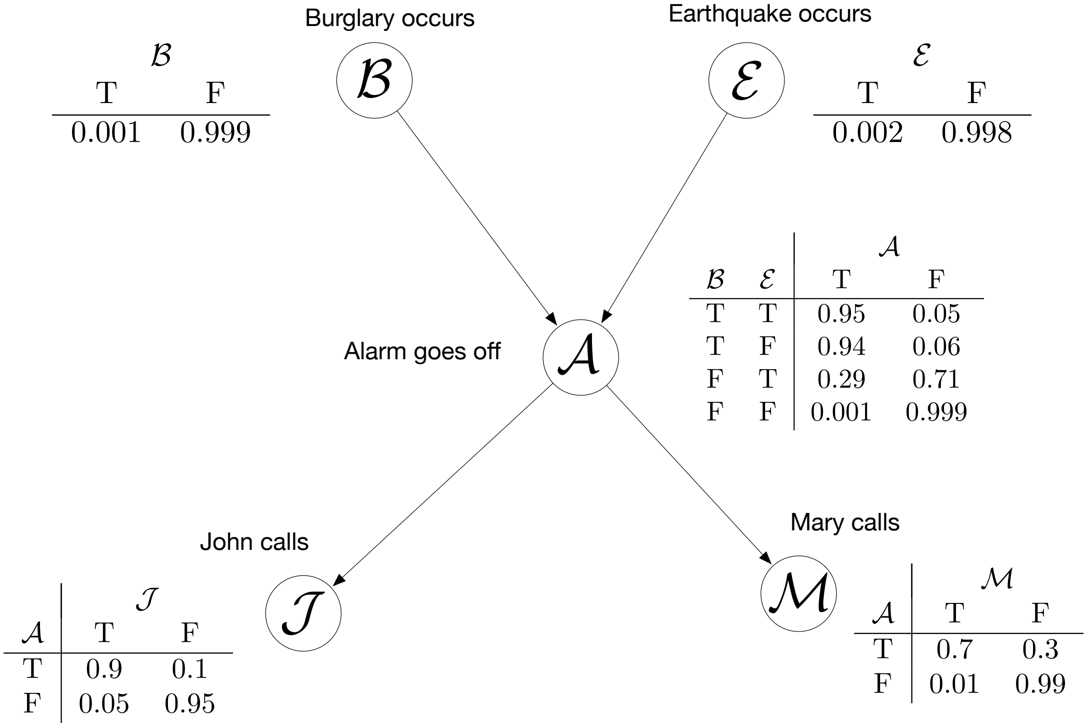
While the graphical structure of the BN is useful for decomposing the joint distribution of the random variables in the graph, it can also be used to reason about the conditional independence relationships in the graph. For instance, it's possible that simply by looking at the graph structure in the BN for the coin problem, we can determine $\mathcal{H}_1 \perp\!\!\!\perp \mathcal{H}_2 \mid \mathcal{C}_1$. In order to figure out conditional independence relationships from a BN, we need to learn about the concept of d-separation.

External Resource(s) (30 minutes)

- Read d-Separation without Tears (don't worry about the third page).
- These videos are pretty good, but the reading seems clearer (let us know what you think on NB) part 1, part 2.

Exercise 9 (15 minutes)

Consider the following BN that describes how two people John and Mary respond to an alarm in their apartment building. In this case the alarm is triggered either by an earthquake, a burglary, or might go off on accident.



$\mathcal{B}$	
T	F
0.001	0.999

$\mathcal{E}$	
T	F
0.002	0.998

		$\mathcal{A}$	
$\mathcal{B}$	$\mathcal{E}$	T	F
T	T	0.95	0.05
T	F	0.94	0.06
F	T	0.29	0.71
F	F	0.001	0.999

	$\mathcal{J}$	
$\mathcal{A}$	T	F
T	0.9	0.1
F	0.05	0.95

	$\mathcal{M}$	
$\mathcal{A}$	T	F
T	0.7	0.3
F	0.01	0.99

For each of the following potential conditional independence relationships, state whether they are true or false (justify your answer). You should use the rules of d-separation to determine your answers. Hint: the specific probability values given in the BN are not relevant for answering this question. The connections between the nodes are all you need to determine conditional independence (we will use the probability tables in the next exercise).

(a) $\mathcal{B} \perp\!\!\!\perp \mathcal{E}$

☆ Solution

True. The only path between these two nodes is blocked by a collider.

(b) $\mathcal{B} \perp\!\!\!\perp \mathcal{M} \mid \mathcal{A}$

☆ Solution

True. The only path between these two nodes is blocked by virtue of the fact we are conditioning on $\mathcal{A}$.

(c) $\mathcal{B} \perp\!\!\!\perp \mathcal{E} \mid \mathcal{J}$

☆ Solution

False. $\mathcal{A}$ no longer acts as a collider since we are conditioning on one of its descendants ($\mathcal{J}$).

(d) $\mathcal{J} \perp\!\!\!\perp \mathcal{M}$

☆ Solution

False. There is a collider-free path between the two nodes (through $\mathcal{A}$).

(e) $\mathcal{J} \perp\!\!\!\perp \mathcal{M} \mid \mathcal{A}$

☆ Solution

True. Conditioning on $\mathcal{A}$ breaks the only path between these nodes.

Exercise 10 (60 minutes)

Consider the following BN from the previous problem that describes how two people John and Mary respond to an alarm in their apartment building. Compute the following probabilities (for some problems you will be able to simplify your calculations by testing for the independence (or conditional independence) using d-separation.

(a) $p(\mathcal{B}, \mathcal{E})$

☆ Solution

As we saw in the previous exercise, $\mathcal{B}$ and $\mathcal{E}$ are d-separated. Therefore, $\mathcal{B} \perp\!\!\!\perp \mathcal{E}$.

$$\begin{aligned} p(\mathcal{B}, \mathcal{E}) &= p(\mathcal{B})p(\mathcal{E}) \\ &= 0.001 \times 0.002 \\ &= 0.000002 \end{aligned}$$

(b) $p(\mathcal{J}, \mathcal{M}|\mathcal{A})$

☆ Solution

As we saw in the previous exercise, $\mathcal{J}$ and $\mathcal{M}$ are d-separated when conditioning on $\mathcal{A}$ (since it breaks the path connecting them). Therefore $\mathcal{J} \perp\!\!\!\perp \mathcal{M} \mid \mathcal{A}$.

$$\begin{aligned} p(\mathcal{J}, \mathcal{M}|\mathcal{A}) &= p(\mathcal{J}|\mathcal{A})p(\mathcal{M}|\mathcal{A}) \\ &= 0.9 \times 0.7 \\ &= 0.63 \end{aligned}$$

(c) $p(\mathcal{B}|\mathcal{A})$ (hint: don't forget about Bayes' rule) (hint 2: don't forget about marginalization)

☆ Solution

First, we apply Bayes' rule.

$$p(\mathcal{B}|\mathcal{A}) = \frac{p(\mathcal{A}|\mathcal{B})p(\mathcal{B})}{p(\mathcal{A})}$$

The only one of those terms that is easy to get is $p(\mathcal{B})$. To compute $p(\mathcal{A}|\mathcal{B})$ we marginalize out $\mathcal{E}$.

$$\begin{aligned}
p(\mathcal{A}|\mathcal{B}) &= p(\mathcal{A},\mathcal{E}|\mathcal{B}) + p(\mathcal{A},\neg\mathcal{E}|\mathcal{B}) && \\
&= p(\mathcal{E}|\mathcal{B})p(\mathcal{A}|\mathcal{E},\mathcal{B}) + p(\neg\mathcal{E}|\mathcal{B})p(\mathcal{A}|\neg\mathcal{E},\mathcal{B}) && \text{product rule} \\
&= p(\mathcal{E})p(\mathcal{A}|\mathcal{E},\mathcal{B}) + p(\neg\mathcal{E})p(\mathcal{A}|\neg\mathcal{E},\mathcal{B}) && \mathcal{E} \perp\!\!\!\perp \mathcal{B} \\
&= 0.002 \times 0.95 + 0.998 \times 0.94 && \\
&= 0.94002 &&
\end{aligned}$$

Next, we concentrate on $p(\mathcal{A})$. To tackle this one, we marginalize over two events ($\mathcal{B}$ and $\mathcal{E}$).

$$\begin{aligned}
p(\mathcal{A}) &= p(\mathcal{B},\mathcal{E},\mathcal{A}) + p(\mathcal{B},\neg\mathcal{E},\mathcal{A}) + p(\neg\mathcal{B},\mathcal{E},\mathcal{A}) + p(\neg\mathcal{B},\neg\mathcal{E},\mathcal{A}) && \\
&= p(\mathcal{B})p(\mathcal{E}|\mathcal{B})p(\mathcal{A}|\mathcal{B},\mathcal{E}) + p(\mathcal{B})p(\neg\mathcal{E}|\mathcal{B})p(\mathcal{A}|\mathcal{B},\neg\mathcal{E}) && \\
&\qquad + p(\neg\mathcal{B})p(\mathcal{E}|\neg\mathcal{B})p(\mathcal{A}|\neg\mathcal{B},\mathcal{E}) + p(\neg\mathcal{B})p(\neg\mathcal{E}|\neg\mathcal{B})p(\mathcal{A}|\neg\mathcal{B},\neg\mathcal{E}) && \text{product rule} \\
&= p(\mathcal{B})p(\mathcal{E})p(\mathcal{A}|\mathcal{B},\mathcal{E}) + p(\mathcal{B})p(\neg\mathcal{E})p(\mathcal{A}|\mathcal{B},\neg\mathcal{E}) && \\
&\qquad + p(\neg\mathcal{B})p(\mathcal{E})p(\mathcal{A}|\neg\mathcal{B},\mathcal{E}) + p(\neg\mathcal{B})p(\neg\mathcal{E})p(\mathcal{A}|\neg\mathcal{B},\neg\mathcal{E}) && \mathcal{E} \perp\!\!\!\perp \mathcal{B} \\
&= 0.001 \times 0.002 \times 0.95 + 0.001 \times 0.998 \times 0.94 && \\
&\qquad + 0.999 \times 0.002 \times 0.29 + 0.999 \times 0.998 \times 0.001 && \\
&= 0.002516 && \\
p(\mathcal{B}|\mathcal{A}) &= \frac{0.94002 \times 0.001}{0.002516} && \text{plugging in our calculatations} \\
&= 0.3736 &&
\end{aligned}$$

(d) $p(\mathcal{B}|\mathcal{A},\mathcal{E})$ (this is known as the phenomenon of *explaining away*). Hint: when you apply Bayes' rule, you can leave some of the events on the right hand side of the conditioning bar (look back at the earlier section "Some Twists on Bayes' Rule". If you need a hint to get you started, try applying the following version of Bayes' rule.

$$p(\mathcal{B}|\mathcal{A},\mathcal{E}) = \frac{p(\mathcal{A}|\mathcal{B},\mathcal{E})p(\mathcal{B}|\mathcal{E})}{p(\mathcal{A}|\mathcal{E})}$$

☆ Solution

Staring with the hint we can simplify $p(\mathcal{B}|\mathcal{E})$ since $\mathcal{B} \perp\!\!\!\perp \mathcal{E}$.

$$p(\mathcal{B}|\mathcal{A},\mathcal{E}) = \frac{p(\mathcal{A}|\mathcal{B},\mathcal{E})p(\mathcal{B})}{p(\mathcal{A}|\mathcal{E})}$$

The two terms in the numerator can be read right from the BN, but the denominator requires a little bit more work. We'll follow the same step we did in part (c), except this time we'll marginalize out $\mathcal{B}$.

$$\begin{aligned}
p(\mathcal{A}|\mathcal{E}) &= p(\mathcal{A},\mathcal{B}|\mathcal{E}) + p(\mathcal{A},\neg\mathcal{B}|\mathcal{E}) \\
&= p(\mathcal{B}|\mathcal{E})p(\mathcal{A}|\mathcal{B},\mathcal{E}) + p(\neg\mathcal{B}|\mathcal{E})p(\mathcal{A}|\neg\mathcal{B},\mathcal{E}) && \text{product rule} \\
&= p(\mathcal{B})p(\mathcal{A}|\mathcal{B},\mathcal{E}) + p(\neg\mathcal{B})p(\mathcal{A}|\neg\mathcal{B},\mathcal{E}) && \mathcal{B} \perp\!\!\!\perp \mathcal{E} \\
&= 0.29066 \\
p(\mathcal{B}|\mathcal{A},\mathcal{E}) &= \frac{0.95 \times 0.001}{0.29066} \\
&= 0.003268
\end{aligned}$$

7 Generative versus Discriminative Models

In this assignment and the previous one we've built up a lot of machinery that allows us to work with probabilities. Next we're going to take this machinery and turn back towards machine learning. Specifically, we'll be looking at the classification problem and using probability theory to see it in a whole new light (who's excited?!?).

7.1 Discriminative Models: a Look Back at Logistic Regression (10 minute read)

Let's think back to the logistic regression model for binary classification that we learned about in module 1. Given an input point $\mathbf{x_i}$, the logistic regression algorithm applied a weight vector $\mathbf{w}$ to compute the probability that the corresponding output y_i was 1 via the formula $\sigma(\mathbf{w}^\top\mathbf{x_i}) = \frac{1}{1+e^{-\mathbf{w}^\top\mathbf{x_i}}}$ (recall that σ is known as the sigmoid function and serves to squash its input into a number between 0 and 1, which can serve as a valid probability). While we didn't quite have the vocabulary for it then, what we were really doing was computing a conditional probability. We can think of Y_i as a random variable that represents the output corresponding to the input $\mathbf{x_i}$ (in the case of binary classification Y_i is either 0 or 1). We can also think of the input as a random variable X_i (thinking of the input as a random variable will be helpful later in this section). Framed in this way the logistic regression algorithm computes the following conditional probability:

$$p(Y_i = 1|X_i = \mathbf{x_i}) = \sigma(\mathbf{w}^\top\mathbf{x_i}) \ . \tag{16}$$

We defined a loss function to specify which weights were better or worse given a training set $(\mathbf{x_1}, y_1), (\mathbf{x_2}, y_2), \ldots, (\mathbf{x_n}, y_n)$. The details of how we did this are not important to the point we are trying to make now, so it'll suffice to say that learning in a logistic regression model meant tuning the conditional distribution of the outputs (the Y_i's) given the inputs ($\mathbf{x_i}$'s) to fit the training data the best. This type of model is what is known as a *discriminative model* (the Wikipedia article on discriminative models has more details if you are interested).

✔ Understanding Check

Intuitively, why does being able to accurately estimate $p(Y = 1|X = \mathbf{x})$ allow you to classify unseen test data?

☆ Solution

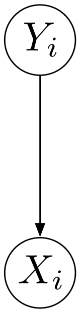
In a classification problem we are interested in predicting the output for a given input, $\mathbf{x}$. If we can accurately calculate $p(Y = 1|X = \mathbf{x})$, then we can simply threshold this value at 0.5 (or whatever makes sense given our application) and predict $Y = 1$ if this condition is met and $Y = 0$ otherwise.

7.2 Generative Models (10 minute read)

The approach outlined above is great, but it's not the only way to approach binary classification (and supervised learning in general). Since we are interested in predicting Y_i given some inputs $\mathbf{x_i}$, it of course makes sense, for example for a binary classification problem, to want to determine $p(Y_i = 1|\mathbf{x_i})$. However, Instead of modeling that distribution directly, we can use Bayes' rule.

$$\begin{aligned} p(Y_i = 1|X_i = \mathbf{x_i}) &= \frac{p(X_i = \mathbf{x_i}|Y_i = 1)p(Y_i = 1)}{p(X_i = \mathbf{x_i})} \\ &= \frac{p(X_i = \mathbf{x_i}|Y_i = 1)p(Y_i = 1)}{p(X_i = \mathbf{x_i}|Y_i = 1)p(Y_i = 1) + p(X_i = \mathbf{x_i}|Y_i = 0)p(Y_i = 0)} \end{aligned} \tag{17}$$

These equations tell us is that if we have a model of the probability of the output being 1 *a priori*, $p(Y_i = 1)$, and a model of the inputs $\mathbf{x_i}$ given the output y_i, $p(X_i = \mathbf{x_i}|Y = y_i)$, then we can compute $p(Y_i = 1|X_i = \mathbf{x}_i)$. This amounts to adopting the perspective that the hidden output Y_i causes the input X_i. We call this sort of model a probabilistic generative model (PGM). The BN corresponding to this model is given below.



The natural question is *why?* Here are some potential advantages of using probabilistic generative models.

- Suppose you found out that $p(Y_i)$ changed for some reason (any thoughts on when this might happen? Post here on NB). Incorporating this change into a probabilistic graphical model would be very straightforward (just modify $p(Y_i = 1)$ in Equation **??**).
- Suppose you found out that $p(X_i = \mathbf{x_i}|Y_i = y_i)$ changed for some reason. For example, if one of the elements of X_i represents a result obtained by running some sort of medical test, the sensitivity of that medical test might change (any other examples on when this might happen? Post here on NB.).

- Suppose that instead of classifying data (i.e., predicting Y_i), you instead wanted to generate samples $\mathbf{x_i}$ conditioned on a particular value of Y_i (e.g., you might want to synthesize samples of hand written digits based on training a probabilistic graphical model). This can be done naturally with a PGM. More modern versions of this idea are generative adversarial networks (GANs), which are behind such work as this person does not exist and better language models and their implications (the second link is the work of a former Oliner!).

✔ **Understanding Check**

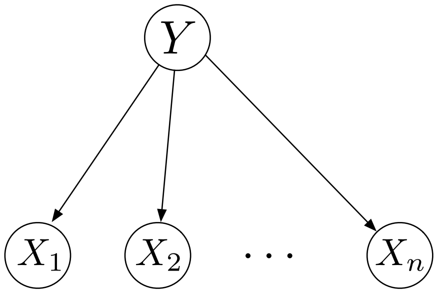
What are the probabilities needed to classify input data in a discriminative model? What are the probabilities needed to classify input data in a generative model? How does Bayes' rule connect these two models?

☆ **Solution**

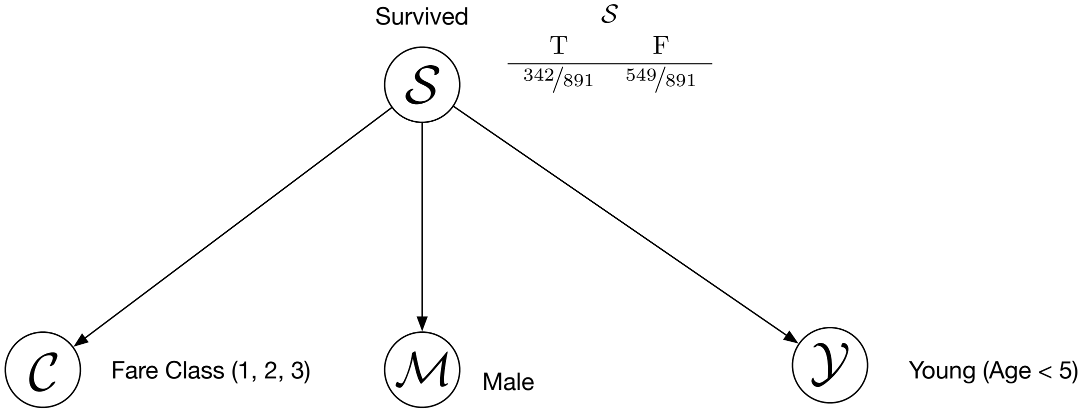
For a discriminative model you need $p(Y = 1|X = \mathbf{X})$. For a generative model you need $p(Y = 1)$ and $p(X = \mathbf{x}|Y = 1)$. Bayes' rule provides the linkage between these two models since one can be transformed into the other simply by applying Bayes' rule.

8 *Meet the Naïve Bayes Algorithm*

Now that we've learned the basic concept of a probabilistic graphical model, let's make things concrete and learn about a specific PGM. For our first PGM, we're going to learn about the Naïve Bayes algorithm (we'll be doing some work with Naïve Bayes in the next assignment as well, so you'll have some time to get accustomed to it). The reason it is called Naïve Bayes is that it assumes that all of the observed data $(X_1, X_2, \ldots, X_n)$ are conditionally independent given $\mathcal{Y}$. The BN for the Naïve Bayes algorithm is shown below.



As a motivating example, let's look back at the Titanic dataset from the last module. A potential BN for the Titanic dataset is shown below.



$\mathcal{S}$	
T	F
342/891	549/891

$\mathcal{S}$	$\mathcal{C}$ 1	2	3
T	136/342	87/342	119/342
F	80/549	97/549	372/549

$\mathcal{S}$	$\mathcal{M}$ T	F
T	109/342	233/342
F	468/549	81/549

$\mathcal{S}$	$\mathcal{Y}$ T	F
T	27/342	315/342
F	13/549	536/549

The probabilities in this BN were computed by looking at the training set and counting the appropriate passengers that fell into each category. For instance, to compute $p(\mathcal{Y}|\mathcal{S})$ since $p(\mathcal{Y}|\mathcal{S}) = \frac{p(\mathcal{Y},\mathcal{S})}{p(\mathcal{S})}$, we can approximate this probability by counting the number of passengers under 5 who survived and dividing by the total number who survived (note that there are some subtle and important modifications to this method of fitting these probabilities that we'll discuss in the next assignment). This process was repeated for each conditional probability. Since we assume that all of the features are conditionally independent given the output ($\mathcal{S}$ in this case), this process is done independently for each feature.

8.1 Inference

While the Naïve Bayes Algorithm might sound fancy, once we have the BN, all we need to do to run the algorithm is to use Bayes' rule. We'll let you work through this on your own via an exercise.

Exercise 11 (45 minutes)

(a) Using the BN shown above, what is the probability that a young, male in first class would survive the Titanic disaster? Hint: write this as a conditional probability and then use Bayes' rule. Hint 2: leverage the fact that $\mathcal{C}, \mathcal{Y}, \mathcal{M}$ are all conditionally independent of each other given $\mathcal{S}$.

You have just derived the Naïve Bayes inference rule!

☆ Solution

$$\begin{aligned}
p(\mathcal{S}|\mathcal{Y},\mathcal{C}=1,\mathcal{M}) &= \frac{p(\mathcal{Y},\mathcal{C}=1,\mathcal{M}|\mathcal{S})p(\mathcal{S})}{p(\mathcal{Y},\mathcal{C}=1,\mathcal{M})}\\
&= \frac{p(\mathcal{Y},\mathcal{C}=1,\mathcal{M}|\mathcal{S})p(\mathcal{S})}{p(\mathcal{Y},\mathcal{C}=1,\mathcal{M}|\mathcal{S})p(\mathcal{S})+p(\mathcal{Y},\mathcal{C}=1,\mathcal{M}|\neg\mathcal{S})p(\neg\mathcal{S})}\\
&= \frac{p(\mathcal{Y}|\mathcal{S})p(\mathcal{C}=1|\mathcal{S})p(\mathcal{M}|\mathcal{S})p(\mathcal{S})}{p(\mathcal{Y}|\mathcal{S})p(\mathcal{C}=1|\mathcal{S})p(\mathcal{M}|\mathcal{S})p(\mathcal{S})+p(\mathcal{Y}|\neg\mathcal{S})p(\mathcal{C}=1|\neg\mathcal{S})p(\mathcal{M}|\neg\mathcal{S})p(\neg\mathcal{S})}\\
&= \frac{\left(\frac{27}{342}\times\frac{136}{342}\times\frac{109}{342}\times\frac{342}{891}\right)}{\left(\frac{27}{342}\times\frac{136}{342}\times\frac{109}{342}\times\frac{342}{891}\right)+\left(\frac{13}{549}\times\frac{80}{549}\times\frac{468}{549}\times\frac{549}{891}\right)}\\
&= 0.6794
\end{aligned}$$

(b) Naïve Bayes is often more conveniently expressed using odds ratios. Instead of computing $p(\mathcal{S}|\mathcal{Y},\mathcal{C}=1,\mathcal{M})$ let's compute the following.

$$\begin{aligned}
\frac{p(\mathcal{S}|\mathcal{Y},\mathcal{C}=1,\mathcal{M})}{p(\neg\mathcal{S}|\mathcal{Y},\mathcal{C}=1,\mathcal{M})} &= \frac{\frac{p(\mathcal{Y},\mathcal{C}=1,\mathcal{M}|\mathcal{S})p(\mathcal{S})}{p(\mathcal{Y},\mathcal{C}=1,\mathcal{M})}}{\frac{p(\mathcal{Y},\mathcal{C}=1,\mathcal{M}|\neg\mathcal{S})p(\neg\mathcal{S})}{p(\mathcal{Y},\mathcal{C}=1,\mathcal{M})}}\\
&= \frac{p(\mathcal{Y},\mathcal{C}=1,\mathcal{M}|\mathcal{S})p(\mathcal{S})}{p(\mathcal{Y},\mathcal{C}=1,\mathcal{M}|\neg\mathcal{S})p(\neg\mathcal{S})}\\
&= \frac{p(\mathcal{Y}|\mathcal{S})p(\mathcal{C}=1|\mathcal{S})p(\mathcal{M}|\mathcal{S})p(\mathcal{S})}{p(\mathcal{Y}|\neg\mathcal{S})p(\mathcal{C}=1|\neg\mathcal{S})p(\mathcal{M}|\neg\mathcal{S})p(\neg\mathcal{S})}
\end{aligned}$$

What must be true about this odds ratio in order to predict that the passenger survived?

☆ Solution

The odds ratio must be greater than 1, which implies that

$$p(\mathcal{S}|\mathcal{Y},\mathcal{C}=1,\mathcal{M}) > p(\neg\mathcal{S}|\mathcal{Y},\mathcal{C}=1,\mathcal{M})\ .$$