Assignment 2: Probabilistic Graphical Models

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 Todo

- It would be nice to standardize on random variables instead of events and random variables (events are just a special case).

2 Motivation and Context

- We've learned how probabilities can be used to describe uncertainty in the world
- We've learned how Bayes' rule can be used to reason about hypotheses, models, or other things that cannot be directly observed.

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 it using the product rule in the following way.

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

Notice that this looks pretty much identical to Equation 1 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).

You may also 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 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 2 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 4 is a special case of the Equation 5 where $\mathcal{Y} = \{1, 2, \ldots, k\}$.

Exercise 1 (15 minutes)

This exercise is based off of the example given in 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 that 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.

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

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

☆ Solution

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

4 Some Twists on Bayes' Rule

You should be pretty proficient with the vanilla form of Bayes' rule. There are a few 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).

$$p(\mathcal{A}|\mathcal{B}) = \frac{p(\mathcal{B}|\mathcal{A})p(\mathcal{A})}{p(\mathcal{B})} \qquad \text{vanilla Bayes' rule} \tag{7}$$

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

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

5 Independence and Conditional Independence

Two of the most important concepts in probability theory are independence and the closely related concept of conditional independence. The reason that these ideas are important is that 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. This saves makes your life much, 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 10 is that if $\mathcal{A} \perp\!\!\!\perp \mathcal{B}$, then

$$p(\mathcal{A}|\mathcal{B}) = p(\mathcal{A}) \qquad \text{and}$$
$$p(\mathcal{B}|\mathcal{A}) = p(\mathcal{B}) \ .$$

A very similar equation to Equation 10 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 2 (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.

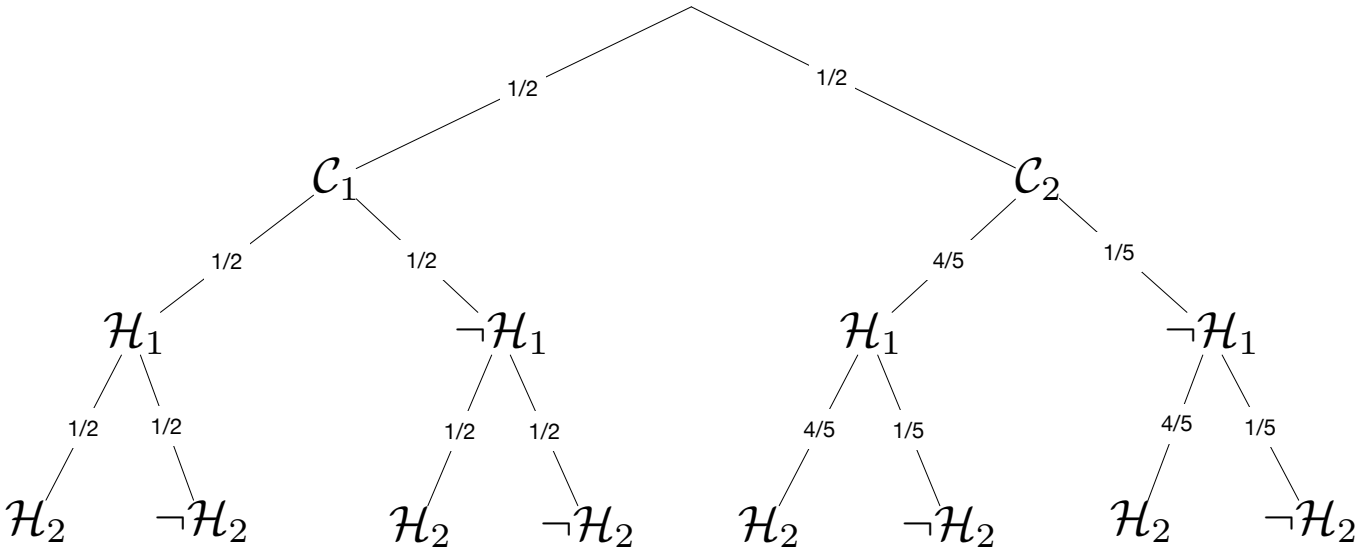
- The daily increase in the Dow Jones Industrial average and the daily increase in the NASDAQ.
- The event that the American League wins the World Series in 2019 and the event that the National League wins the World Series in 2019.
- Testing positive for a disease and having that disease.

5.2 Conditional Independence

Sometimes two events (or two random variables) that are not independent might become independent when conditioned on another event. As a motivating example, consider a variant of the coin problem we saw last assignment. A bag contains two coins ($\mathcal{C}_1$ and $\mathcal{C}_2$). Coin 1 is fair $p(\mathcal{H}|\mathcal{C}_1) = \frac{1}{2}$. Coin 2 is not fair ($p(\mathcal{H}|\mathcal{C}_2) = \frac{4}{5}$). 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. We then flip the coin twice. 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. The question is are $\mathcal{H}_1$ and $\mathcal{H}_2$ independent (i.e., is $\mathcal{H}_1 \perp\!\!\!\perp \mathcal{H}_2$)?

Exercise 3 (20 minutes)

In order to determine whether $\mathcal{H}_1 \perp\!\!\!\perp \mathcal{H}_2$, let's make a tree diagram.



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:

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

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, we can state that they are what's called *conditionally independent* given $\mathcal{C}_1$ (or $\mathcal{C}_2$). More formally, events $\mathcal{A}$ and $\mathcal{B}$ are considered conditionally independent given $\mathcal{C}$ (written as $\mathcal{A} \perp\!\!\!\perp \mathcal{B} \mid \mathcal{C}$) if and only if

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

Exercise 4 (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.

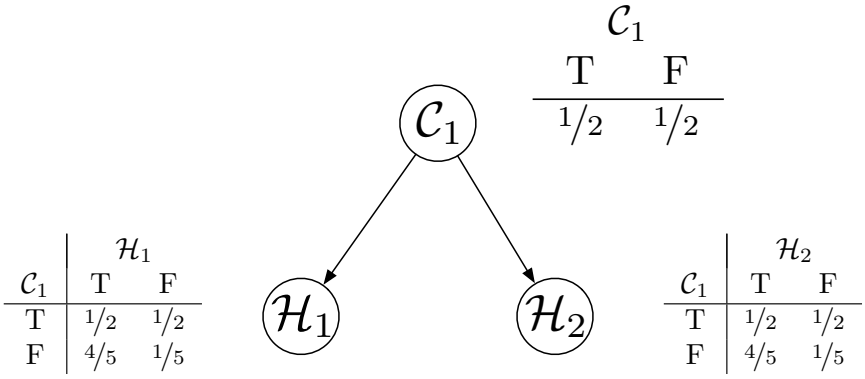
$$\begin{aligned}
p(\mathcal{H}_1, \mathcal{H}_2|\mathcal{C}_2) &= \frac{4}{5} \times \frac{4}{5} = \frac{16}{25} \\
p(\mathcal{H}_1|\mathcal{C}_2) &= p(\mathcal{H}_1, \mathcal{H}_2|\mathcal{C}_2) + p(\mathcal{H}_1, \neg\mathcal{H}_2|\mathcal{C}_2) \\
&= \left(\frac{4}{5} \times \frac{4}{5}\right) + \left(\frac{4}{5} \times \frac{1}{2}\right) = \frac{4}{5} \\
p(\mathcal{H}_2|\mathcal{C}_2) &= p(\mathcal{H}_1, \mathcal{H}_2|\mathcal{C}_2) + p(\neg\mathcal{H}_1, \mathcal{H}_2|\mathcal{C}_2) \\
&= \left(\frac{4}{5} \times \frac{4}{5}\right) + \left(\frac{1}{5} \times \frac{4}{5}\right) = \frac{4}{5} \\
p(\mathcal{H}_1, \mathcal{H}_2|\mathcal{C}_1) &= \frac{4}{5} \times \frac{4}{5} \\
&= p(\mathcal{H}_1|\mathcal{C}_2)p(\mathcal{H}_2|\mathcal{C}_2)
\end{aligned}$$

The definition of the conditional independence of events can be easily extended to random variables. We say that 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 equation 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. (TODO: maybe redo this notation to remove the T's and F's).



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 this information from the graph. That's coming later in the assignment). The tables that are listed by each node tell us the probability of the event happening versus not happening conditioned on whether or not the events listed on the nodes parents (a "parent" of a node, A, is a node B where there is an edge pointing from A to B) happened (happening T stands for *True* or that the event does happen and F stands for *False* or that the event does not 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 want to write the joint distribution of all of the events or random variables (the relationship is the same for either) in the network, it must factorize in the following way. We'll use $X_1, X_2, \ldots X_n$ 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.

$$p(X_1, X_2, \ldots, X_n) = p(X_1|Pa(X_1)) \times p(X_2|PaX_2)) \times \ldots p(X_n|Pa(X_n)) \tag{14}$$

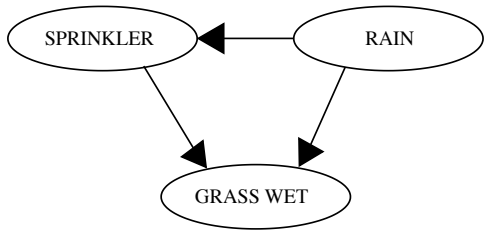
Back to our coin BN, this means that we can write the joint distribution like so.

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

Exercise 5 (20 minutes)

Consider the belief network below (source: https://en.wikipedia.org/wiki/Bayesian_network#Example). (TODO: not crazy about this notation)

RAIN	SPRINKLER T	SPRINKLER F
F	0.4	0.6
T	0.01	0.99



RAIN T	RAIN F
0.2	0.8

SPRINKLER	RAIN	GRASS WET T	GRASS WET 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 (i.e., is true)).

(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 the $\mathcal{R}$ node. 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

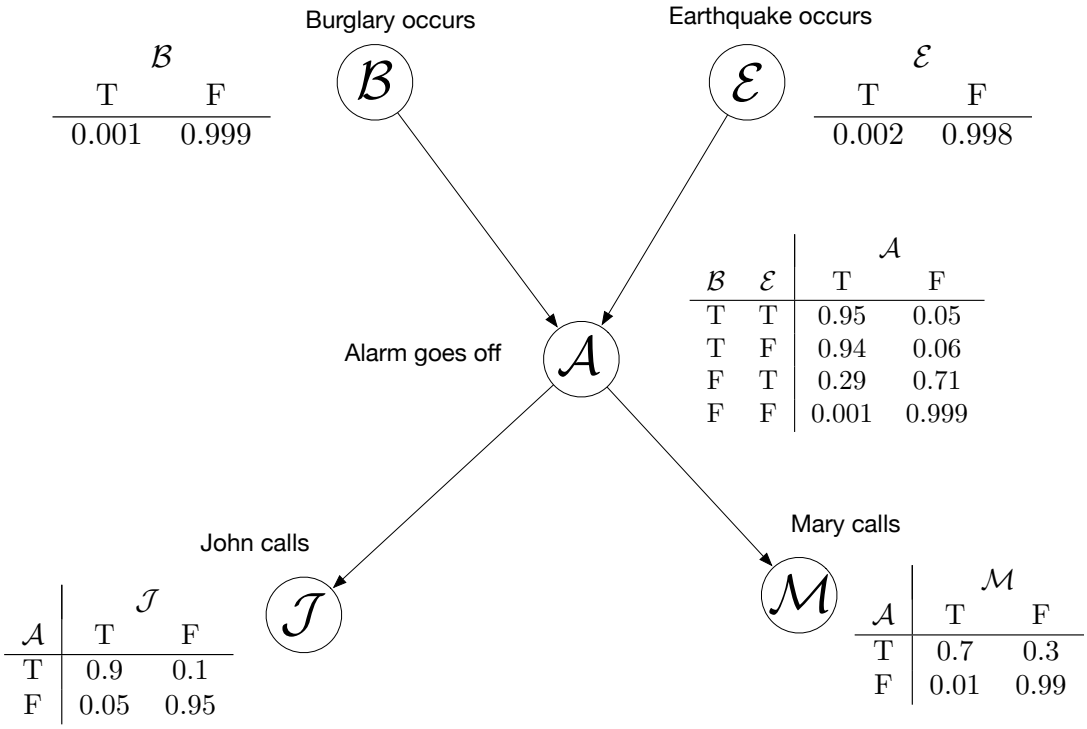
While the graphical structure of the BN is useful for decomposing the joint distribution of the random variables in the graph, it can also be used to reason about the conditional independence relationships in the graph. For instance, it's possible that given the BN for the coin problem that we can determine that $\mathcal{H}_1 \perp\!\!\!\perp \mathcal{H}_2 \mid \mathcal{C}_1$ simply by looking at the graph. In order to figure out these sorts of conditional independence relationships, 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).
- This one seems pretty good
- Pieter Abbeel Lecture (not sure how clear this is)

Exercise 6 (15 minutes)

Consider the following BN that describes how two people John and Mary respond to an alarm going off 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

$\mathcal{B}$	$\mathcal{E}$	$\mathcal{A}$	
		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

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

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

For each of the following statements of conditional independence, 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 one path between these nodes.

Exercise 7 (45 minutes)

Consider the following BN from the previous problem that describes how two people John and Mary respond to an alarm going off 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 when conditioning on $\mathcal{A}$. Therefore, $\mathcal{B} \perp\!\!\!\perp \mathcal{E}$.

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

(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 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.

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

If we marginalize out $\mathcal{E}$ we are left with the following.

$$\begin{aligned}
p(\mathcal{A}|\mathcal{B}) &= p(\mathcal{A}, \mathcal{E}|\mathcal{B}) + p(\mathcal{A}, \neg\mathcal{E}|\mathcal{B}) \\
&= p(\mathcal{E}|\mathcal{B})p(\mathcal{A}|\mathcal{E}, \mathcal{B}) + p(\neg\mathcal{E}|\mathcal{B})p(\mathcal{A}|\neg\mathcal{E}, \mathcal{B}) \\
&= p(\mathcal{E})p(\mathcal{A}|\mathcal{E}, \mathcal{B}) + p(\neg\mathcal{E})p(\mathcal{A}|\neg\mathcal{E}, \mathcal{B}) \\
&= 0.002 \times 0.95 + 0.998 \times 0.94 \\
&= 0.94002 \\
p(\mathcal{A}) &= p(\mathcal{B}, \mathcal{E}, \mathcal{A}) + p(\mathcal{B}, \neg\mathcal{E}, \mathcal{A}) + p(\neg\mathcal{B}, \mathcal{E}, \mathcal{A}) + p(\neg\mathcal{B}, \neg\mathcal{E}, \mathcal{A}) \\
&= p(\mathcal{B})p(\mathcal{E}|\mathcal{B})p(\mathcal{A}|\mathcal{B}, \mathcal{E}) + p(\mathcal{B})p(\neg\mathcal{E}|\mathcal{B})p(\mathcal{A}|\mathcal{B}, \neg\mathcal{E}) \\
&\qquad + p(\neg\mathcal{B})p(\mathcal{E}|\neg\mathcal{B})p(\mathcal{A}|\neg\mathcal{B}, \mathcal{E}) + p(\neg\mathcal{B})p(\neg\mathcal{E}|\neg\mathcal{B})p(\mathcal{A}|\neg\mathcal{B}, \neg\mathcal{E}) \\
&= p(\mathcal{B})p(\mathcal{E})p(\mathcal{A}|\mathcal{B}, \mathcal{E}) + p(\mathcal{B})p(\neg\mathcal{E})p(\mathcal{A}|\mathcal{B}, \neg\mathcal{E}) \\
&\qquad + p(\neg\mathcal{B})p(\mathcal{E})p(\mathcal{A}|\neg\mathcal{B}, \mathcal{E}) + p(\neg\mathcal{B})p(\neg\mathcal{E})p(\mathcal{A}|\neg\mathcal{B}, \neg\mathcal{E}) \\
&= 0.001 \times 0.002 \times 0.95 + 0.001 \times 0.998 \times 0.94 \\
&\qquad + 0.999 \times 0.002 \times 0.29 + 0.999 \times 0.998 \times 0.001 \\
&= 0.002516 \\
p(\mathcal{A}|\mathcal{B}) &= \frac{0.94002 \times 0.001}{0.002516} \\
&= 0.3736
\end{aligned}$$

(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 (the | symbol). 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 $\mathcal{B}$.

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

7 *Generative versus Discriminative Models*

TODO: write some intro here

7.1 *Discriminative Models: a Look Back at Logistic Regression*

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 utilized a weight vector $\mathbf{w}$ to compute the probability that the corresponding output y_i was 1 using the $\sigma(\mathbf{w}^\top \mathbf{x_i}) = \frac{1}{1+e^{-\mathbf{w}^\top \mathbf{x_i}}}$ (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 really doing at the time was computing a conditional probability. We can think of Y_i as a random variable that represents the output that corresponds to the the input point $\mathbf{x_i}$ (Y_i is either 0 or 1 since we are dealing with binary classification). We can also think of the input as a random variable $\mathbf{X_i}$ (thinking of the input in this way will be helpful later in this section). In this way, we can think of what the logistic regression algorithm is doing as computing the following conditional probability:

$$p(Y_i = 1 | X_i = \mathbf{x_i}) = \sigma(\mathbf{w}^\top \mathbf{x_i}) \ . \tag{15}$$

We then defined a loss function that would let us find the best weights, $\mathbf{w}$, given a training set of corresponding input output pairs $(\mathbf{x_1}, y_1), (\mathbf{x_2}, y_2), \ldots, (\mathbf{x_n}, y_n)$. 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

TODO

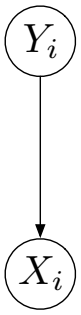
7.2 Generative Models

While the approach outlined above is totally logical, it is not the only way to approach supervised machine learning. Using a simple application of Bayes' rule, we can derive a whole new approach to the problem! 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 $p(Y_i = 1|\mathbf{x_i})$. Instead of modeling that distribution directly, we can instead use Bayes' rule to transform this probability distribution.

TODO: not sure if we should introduce a convention along the lines of $p(\mathbf{x_i}) = p(X_i = \mathbf{x_i})$.

$$\begin{aligned} p(Y_i = 1|X_i = \mathbf{x_i}) &= \frac{p(X_i = \mathbf{x_i}|Y_i = 1)p(Y_i = 1)}{p(X_i = \mathbf{x_i})} \\ &= \frac{p(X_i = \mathbf{x_i}|Y_i = 1)p(Y_i = 1)}{p(X_i = \mathbf{x_i}|Y_i = 1)p(Y_i = 1) + p(X_i = \mathbf{x_i}|Y_i = 0)p(Y_i = 0)} \end{aligned} \tag{16}$$

What these equations are telling us is that if we have a model of the probability of the output being 1 *a priori* ($p(Y_i = 1)$) along with a model of the inputs $\mathbf{x_i}$ given the output Y_i ($p(Y_i|\mathbf{x_i})$), then we have all the information we need to compute $p(Y_i = 1|\mathbf{x}_i)$. In a way this amounts to adopting the perspective that the hidden output Y_i causes the input X_i (see Figure **??**). We call this sort of model a probabilistic generative model (PGM). The BN corresponding to this model is given below.



The natural question you might ask yourself 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 straight forward (just modify $p(Y_i = 1)$ in Equation 16).
- Suppose you found out that $p(X_i|Y_i)$ 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 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). More modern versions of this idea are generative adversarial networks (GANs), which are

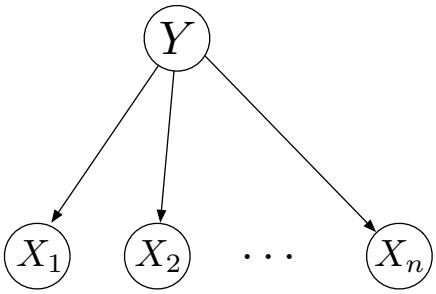
behind such work as this person does not exist and better language models and their implications (the second link is the work of a former Oliner!).

✔ Understanding Check

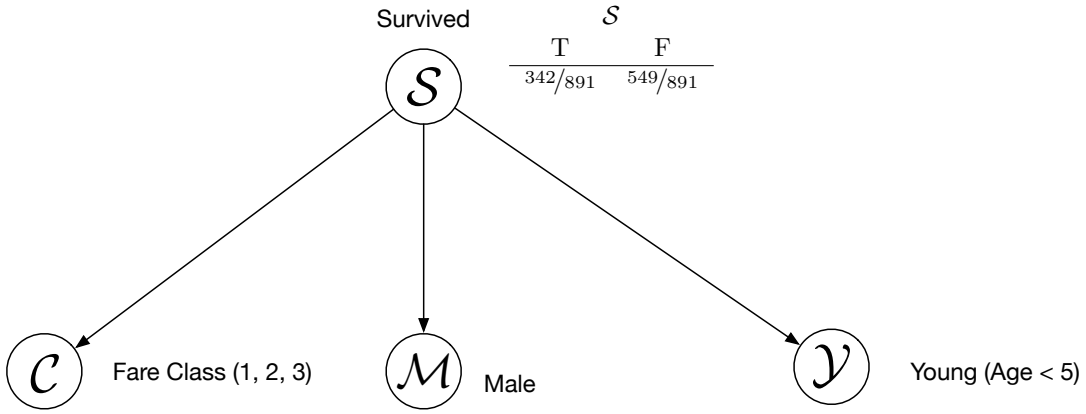
TODO

8 Naïve Bayes: from the top down (maybe do the whole thing...)

Now that we've learned the basic idea of PGMs, let's actually learn about one. Your first PGM is going to be the Naïve Bayes algorithm. 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	$^{13}/_{549}$	$^{536}/_{549}$

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 use Bayes' rule. We'll let you work through this on your own via an exercise.

Exercise 8

(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

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

8.2 Fitting the Probabilities

In this case, the probabilities were simply 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 simply taking the number of passengers under 5 who survived and dividing by the total number who survived.

There are some subtle and important modifications to this method of fitting these probabilities that we'll discuss in the companion notebook. This process can be repeated for each relevant conditional probability. Since we assume all of the features are conditionally independent given the output ($\mathcal{S}$ in this case), this process can be done independently for each feature.

8.3 Naïve Bayes for Text Classification

(punted, almost definitely)

9 Compas Model of Recidivism