Assignment 1: Bayes' Rule

Machine Learning

Fall 2019

TODO: make these more concrete based on Connor's comment in the course feedback survey. (note: Perhaps breaking these down in more detail when we get to a particular problem / reading is better).

Learning Objectives

- Gain familiarity with key ideas in ML (with a focus on probabilistic methods).
- Formal definition of a probability.
- Bayes' Rule
- Applications of Bayesian Inference

1 Motivation

The theme of uncertainty has cropped up a number of times so far this semester. For instance, we learned about the logistic regression model, which computes the probability that the output for a given input is 1. In this way, the logistic regression model explicitly represents the idea of the uncertainty of its predictions and uses the concept of probability to express these uncertainties. We've also seen examples of models that don't fit the data perfectly. For example, the plot in Figure 1 shows a line of best fit that doesn't go through all of the data. The fit might be imperfect because we don't have the right model (a form of uncertainty) or because the thing we are trying to predict is inherently random in some way (another form of uncertainty).

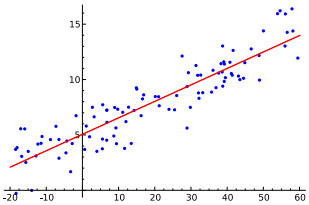


Figure 1: A dataset with one independent variable (x-axis) and one dependent variable. Also shown is the line of best fit.

It turns out that the notion of probability provides us with a precise definition of the concept of uncertainty and allows us to reason about all many forms of uncertainty in a unified way. Formalizing uncertainty using probability theory will enable us to do a bunch of really cool things with respect to machine learning, including

- make explicit our assumptions about where data comes, the uncertainties present within it, and the impact of uncertainty on predictions based on that data.
- allow us to quantify our confidence in our model. For example, instead of returning the one best fitting model (as we did in the first module), we assign a probability to each possible model, and models that fit the data better will get a higher probability.
- give us a way to incorporate expert knowledge into our machine learning models. This expert knowledge will allow for us to more easily interpret machine learning models.
- give us powerful tools to reason about fairness and bias in machine learning models.

2 Six Big Ideas in (Probabilistic) Machine Learning

Back by popular demand[1]... Six Big ideas in ML! Here we're going to highlight ideas in ML that have specific connections to probabilistic machine learning.

[1] there was actually no specific demand for this, but here it is anyway :)

⚠ Notice

Our intent is to provide a larger framework for you to interpret the content you are learning. You should not expect to understand all of these ideas in any sort of detail. Please take these as a "wow isn't this cool / interesting," and avoid getting too lost in the weeds.

this article in Nature Reviews.

The pymc examples are also super cool (e.g., GP Mauna Loa)

Idea 1: Probabilistic Methods Can Help Us Understand Learning in Biological and Artificial Systems

One common challenge that humans and robots (and other animals for that matter) face is the need to reason about and interpret ambiguous information from the world around them. For instance, the photons of light that hit our retinas create a 2-dimensional map of the 3-dimensional world. Despite the fact that there are many possible 3D worlds that could have created the same pattern of light on our retinas, our brain has no trouble figuring out the most likely 3-dimensional structure of the world[2].

[2] optical illusions are a good example of when this does not hold (i.e., our brain is tricked into inferring the wrong 3D structure).

One approach that scientists have taken to understand how humans resolve ambiguities in the world in order to accurately extract structure and meaning is by applying what is known as the computational approach. In the computational approach, one formalizes the goal of the agent as well as the constraints facing it. For example, we might posit that a person is trying to accurately understand the 3D structure of their environment but is constrained to only have access to a 2D projection of that environment (i.e., the 2D projection is what is captured on our retina). Once we have formalized the problem facing the agent, we can then use computational tools to compute the optimal solution to the problem and compare this solution with what people actually do. Through this comparison we may learn more about the nature of human cognition and the reasons why humans think and act the way they do. In David Marr's classic book "Vision," he states the core ideas in the following way,.

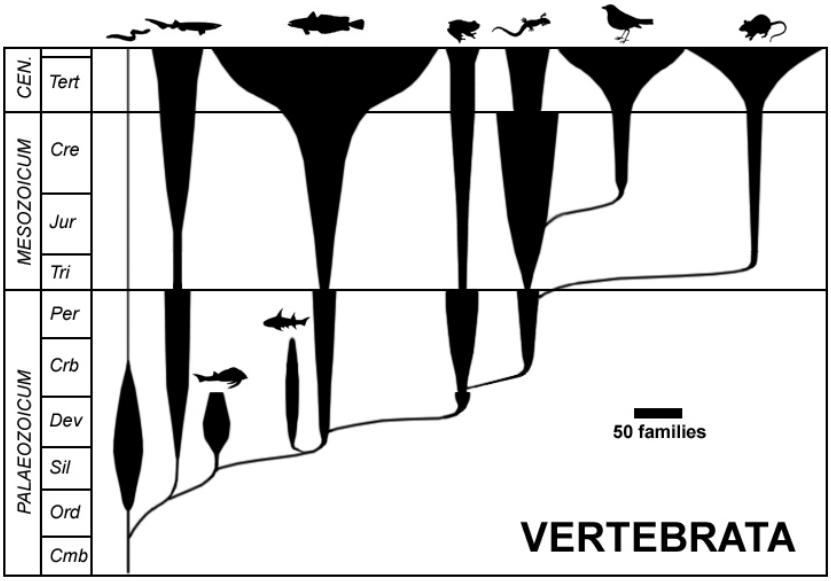
> Trying to understand perception by studying only neurons is like trying to understand bird flight by studying only feathers: It just cannot be done. In order to understand bird flight, we have to understand aerodynamics; only then do the structure of feathers and the different shapes of birds' wings make sense. [Marr, 1982] p. 27

The science of "aerodynamics" is being invoked in the quote to reference a fundamental theory that governs all systems that might try to fly. When it comes to cognition, the idea of Bayesian inference, which will form the basis of probabilistic machine learning and which you'll start learning about later in this assignment, has increasingly emerged as the "aerodynamics" of thought and reasoning. Scientists and engineers have used Bayesian approaches to understand a vast array of human cognition and behavior.

- TODO (millions of ideas to put here)

Idea 2: Probabilistic Methods Can Illuminate Hidden Structures

One of the places where probabilistic machine learning really shines is in determining the hidden structures that underlie data. One very cool example of this is in the field of bioinformatics. A key idea in evolutionary biology is the notion of a phylogenetic tree. A phylogenetic tree captures the relationships between various entities (e.g., individual members of a species, different species, genes, etc.) in a hierarchy (or tree). For instance, the figure below shows a phylogenetic tree of the vertebrates.



You might imagine that such a phylogenetic tree would be put together using a combination of the fossil record along with an analysis of modern day versions of these species. It turns out that probabilistic machine learning has been used to automatically infer phylogenetic trees from a wide variety of data sources. One example is using mitochondrial DNA collected from people all around the world and then using machine learning analysis to reconstruct the most likely phylogenetic tree that underlies these DNA samples. The result is an ancestral tree that helps reconstruct the ancient history of humans and snapshot of how they came to inhabit planet Earth. Reconstructions of this type have given support to the "out of Africa" theory of the origin of modern humans (namely that all people alive today can trace their lineage back to a group of people in subsaharan African from about 150,000 - 200,000 years ago).

The usage of Bayesian approaches to inferring phylogenetic trees has been used in a vast number of places. The abstract of A biologist's guide to Bayesian phylogenetic analysis describes it in the following way.

> Bayesian phylogenetic methods were introduced in the 1990s and have since revolutionised the way we analyse genomic sequence data3. Examples of such analyses include phylogeographic analysis of virus spread in humans inference of phylogeographic history and

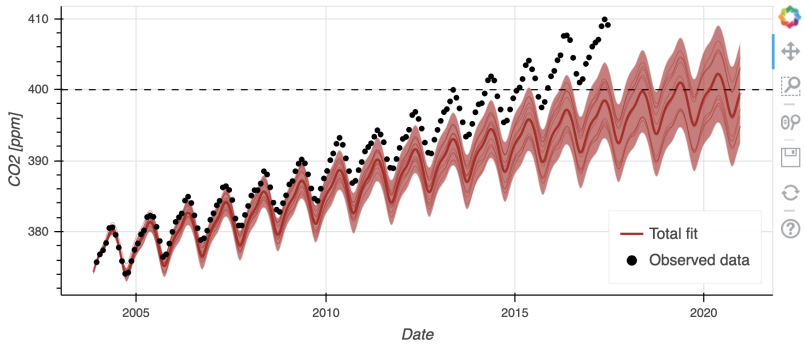
migration between species, analysis of species diversification rates, divergence time estimation, and inference of phylogenetic relationships among species or populations.

(If You're Interested): More examples of finding hidden structures in data Entropic Evidence for Linguistic Structure in the Indus Script

Idea 3: Probabilistic Methods Allow Us to Write Probabilistic Programs

One of the key ideas of probabilistic machine learning is that we can take causal models of how hidden structures give rise to the data that we observe and invert these models in order to allow us to reason about the hidden structures that gave rise to our data. One of the coolest ways in which this fundamental ideas is being leveraged is in the field of probabilistic programming. In a probabilistic program, the programmer writes code to implement the causal model (the one that goes from hidden structures to data) and then the probabilistic programming framework automatically can infer the hidden structure given a sample of data!

For instance, one of the examples listed for the popular probabilistic programming Python library pymc3 shows how a model of CO2 concentration in the atmosphere could be create as a probabilistic program (see Mauna Loa CO2 example). The punchline graph from that notebook is reproduced below. The model in this case was trained on the data up to 2004 and had to predict the rest. In addition to the curve shown, the model also decomposed the curve into various components (including long range and seasonal).



Idea 4: Probabilistic Methods Give Us a Language for Reasoning about Algorithmic Fairness

Idea 5: Probabilistic Methods Allow for Learning More Interpretable Models

https://towardsdatascience.com/what-is-bayesian-statistics-used-for-37b91c2c257c

Idea 6: Probabilistic Methods Let you Learn Hierarchies of Models

TODO (clustering or topic modeling?)

notes: other ideas

- Decision-making under uncertainty
- Probabilistic Programming? (Note: there is an example of inverse computer graphics that is cool). Here is one that references 3D computer vision. Here is probably the clearest version. Another https://arxiv.org/pdf/1607.08128.pdf. This one is really insane (from Science). It uses a few images to predict what the scene would look like from another viewpoint.
- Compression?

Exercise 1 (60 minutes)

Now, we want to hear from you! Choose one of the big ideas above and write a short response to it. Your response could incorporate something surprising you read, a thought-provoking question, your personal experience, an additional resource that builds upon or shifts the discussion. We hope that this reflection will help scaffold class discussions and get you thinking about your interests in the big space that is ML. Also, you have license from us to customize the structure of your response as you see fit. As a rough guide, you should aim for a response of a 1-2 paragraphs.

Solution

There's no one right answer here!

3 Probability

Hopefully the previous section left you feeling excited to learn more about the theory that might underlie these sorts of models and applications. Next, we'll be taking our first steps towards learning this theory.

3.1 Intuition

Most of us are used to thinking that events can be probabilistic, that is we can attach some probability to whether or not they occur. Take for example flipping a coin. We could think of the event that the coin comes up heads as having probability 0.5. That is, there is an even chance that it happens versus doesn't happen. Further, we can talk about events as being observable if we are able to ultimately know whether or not they occurred. For instance, whether a coin comes up heads is an observable event since you can ultimately observe the outcome of the flip. In contrast, some events would be considered unobservable if they are unable to be directly ascertained by human senses. A classic example of this would be whether a scientific theory is true or not. It is impossible to directly observe whether the theory is true, but you might be able to observe events that are consistent or inconsistent with the theory.

Exercise 2 (10 minutes)

Come up with some examples of observable events that are probabilistic in nature. For each event, either provide the probability or explain what factors would determine the probability. Some potential ideas to get you going: sporting events, elections, weather, etc.

☆ Solution

TODO

3.2 Formal Definition

Next, we'll define more formally[3] what we mean by a probability. Having this formal definition will give us the ability to useful rules for manipulating and reasoning about probabilities. To define the concept of a probability, we'll need to specify two ingredients.

[3] this is not the full definition of a probability space used in modern mathematics. For the purposes of most people that *use* probability theory on actual problems, this definition is needlessly complex. Use the following link if you want a more in-depth discussion of the parts of the formal definition that are tricky (our expectation is that you won't want this discussion!).

3.3 Events

The first is the notion we need to fine is an **event**. Think of an event as something that may or may not occur in response to some random process. For instance, we could define the event that a coin comes up heads when it is flipped. We often use capital letters to indicate events. Since we've been using capital letters to also represent matrices, in our materials we'll use a cool mathy-looking calligraphic font to represent events. For instance, we might use the symbol $\mathcal{H}$ to refer to the event that a coinflip comes up heads. It's important to emphasize that a single random process can have many associated events. For instance, for the coin flip example we might also define $\mathcal{T}$ to be the event that the coin comes up tails.

Further, events don't necessarily have to be mutually exclusive. For instance, we might define the $\mathcal{R}_h$ to indicate the event that the Republican party controls the majority in the house of representatives following the 2020 election and $\mathcal{D}_s$ to indicate the event that the Democratic party controls the majority in the senate following the 2020 election. Both (or none) of these events could occur.

3.4 Probability Measure Function

The probability measure function assigns a probability to the occurrence of any particular event. We can think of this probability measure as a function that takes as input an event and outputs a probability. For instance, $p(\mathcal{E})$ provides the probability that event $\mathcal{E}$ occurs according to probability measure p. All probability measure functions must satisfy the following properties.

- $0 \leq p(\mathcal{E}) \leq 1$ (the probability of an event ranges from 0 (for an impossible event) to 1 (an event that will always occur).

- Given a set of n events $\mathcal{E}_1, \mathcal{E}_2, \ldots, \mathcal{E}_n$ that are disjoint (i.e., no two can occur simultaneously)

$$p(\mathcal{E}_1 \text{ or } \mathcal{E}_2 \text{ or } \ldots \text{ or } \mathcal{E}_n) = \sum_{i=1}^{n} p(\mathcal{E}_i) \ . \tag{1}$$

  The equation above specifies what's sometimes called the union rule of probability. It states that the probability of 1 of these disjoint events occurring must be equal to the sum of the probability of each of the events occurring. You will also sometimes see Equation 1 written as

$$p(\mathcal{E}_1 \cup \mathcal{E}_2 \cup \ldots \cup \mathcal{E}_n) = \sum_{i=1}^{n} p(\mathcal{E}_i) \ . \tag{2}$$

  If you're not familiar with the symbol $\cup$ it is the symbol for a union of two sets. The reason you'll sometimes see this is that in the most rigorous definition of a probability space, an event is defined formally as a set (as stated in the margin note earlier in this section, you need not worry about the most rigorous definition in this course).

- Given a set of (not necessarily disjoint) events $\mathcal{E}_1, \mathcal{E}_2, \ldots, \mathcal{E}_n$ where at least one of these n events must occur

$$p(\mathcal{E}_1 \text{ or } \mathcal{E}_2 \text{ or } \ldots \text{ or } \mathcal{E}_n) = 1 \ . \tag{3}$$

3.5 Complement Rule for Probability

Given the definition of probability detailed above, it follows that if the probability of an event happening is $p(\mathcal{E})$ then the probability of the event *NOT* happening is $1 - p(\mathcal{E})$. The following are common ways of expression this relationship (we'll use Equation 4 in this class). These all say the same thing (the only difference is notation).

$$\begin{aligned} p(\neg\mathcal{E}) &= 1 - p(\mathcal{E}) \\ p(\text{not } \mathcal{E}) &= 1 - p(\mathcal{E}) \\ p(\overline{\mathcal{E}}) &= 1 - p(\mathcal{E}) \\ p(\mathcal{E}') &= 1 - p(\mathcal{E}) \end{aligned} \tag{4}$$

We point out these alternate notations not to confuse you (we'd never do that!) but to help you interpret various external resources you might find on these topics.

Exercise 3 (10 minutes)

Here are some diagnostic questions to make sure that you go the basic ideas.

(a) Suppose $\mathcal{E}_1, \mathcal{E}_2, \mathcal{E}_3$ are disjoint events. Further, suppose that one of these events must occur. Which of the following functions are valid probability measure

functions?

$$p_1(\mathcal{E}_1) = \frac{1}{10}, p_1(\mathcal{E}_2) = \frac{1}{5}, p_1(\mathcal{E}_3) = \frac{7}{10}$$

$$p_2(\mathcal{E}_1) = \frac{11}{10}, p_2(\mathcal{E}_2) = \frac{-1}{10}, p_2(\mathcal{E}_3) = 0$$

$$p_3(\mathcal{E}_1) = \frac{1}{10}, p_3(\mathcal{E}_2) = \frac{1}{5}, p_3(\mathcal{E}_3) = \frac{1}{2}$$

$$p_4(\mathcal{E}_1) = 1, p_4(\mathcal{E}_2) = 0, p_4(\mathcal{E}_3) = 0$$

☆ Solution

p_1 is a valid probability measure function since the probabilities add up to 1 and all are non-negative. p_2 is not a valid probability measure function since two of the probabilities are outside of the appropriate range $[0, 1]$. p_3 is not a valid probability measure function since the probabilities of the three events add up to less than 1. p_4 is a valid probability measure function since the probabilities add to 1 and are in the appropriate range.

(b) The Birthday Problem is a well-known probability problem often used in discrete math courses. According to the Wikipedia article, the probability that at least two students among the 70 students in machine learning this semester share the same birthday is 0.999. What is the probability that no two students share the same birthday?

☆ Solution

Notice that the event *no two students share a birthday* only happens when the event *at least two students share a birthday* does not happen. Therefore, these events are complements.

$$\begin{aligned} p(\text{no two students share a birthday}) &= 1 - p(\neg\text{no two students share a birthday}) \\ &= 1 - p(\text{at least two students share a birthday}) \\ &= 1 - 0.999 \\ &= 0.001 \end{aligned}$$

4 Bayes' Rule

External Resource(s) (60 minutes)

Learning Objectives

Note that these learning objectives have been written to be very specific (based on feedback from the course survey). When you first read them, you probably won't know what they mean in detail. As you return to them hopefully the more precise statement of these learning objectives will be useful for assessing your understanding of the provided resources.

- When Bayes' rule is useful (i.e., when $p(A|B)$ is easier to work with than $p(B|A)$).
- The idea of a conjoint probability $p(\mathcal{A}, \mathcal{B})$ (note: alternate notations include $p(\mathcal{A}$ and $\mathcal{B})$ and $p(\mathcal{A} \cap \mathcal{B})$).
- The definition of a conditional probability $p(\mathcal{A}|\mathcal{B}) = \frac{p(\mathcal{A} \text{ and } \mathcal{B})}{p(\mathcal{B})}$.
- The equation for Bayes' rule $p(\mathcal{A}|\mathcal{B}) = \frac{p(\mathcal{B}|\mathcal{A})p(\mathcal{A})}{p(\mathcal{B})}$.
- The equation for the product rule $p(\mathcal{A} \text{ and } \mathcal{B}) = p(\mathcal{B})p(\mathcal{A}|\mathcal{B})$.

Allen Downey (ever heard of him?) wrote an excellent book called Think Bayes that introduces Bayesian analysis. The first chapter starts with a less formal definition of probability than we gave earlier. The chapter then gives intuitions around conjoint probability (the probability that multiple events occur simultaneously), conditional probability (the probability that some event occurs conditioned on another event having occurred), and finally to Bayes' rule (a surprisingly easy theorem to derive that allows you to write one conditional probability distribution in terms of another). The Monty Hall problem in section 1.7 is probably okay to gloss over (see Allen's note at the end of that section for why this is the case).

Allen's treatment of the material is, of course, not the only one out there (we like it for its focus on building intuition and focusing on the key ideas). Here are some other resources you might consider checking out (they are optional).

- Khan Academy Video on Bayes' Theorem shows some simple applications of Bayes' rule and explains why it is a convenient way to reason about the probability of hypothesis given data).
- Veritasium Episode on Bayes' Theorem has a bit more history and philosophy of Bayes' Theorem along with some nice visualizations. It also includes the presenter walking on a very scenic mountain (for some reason), so there's that if nothing else.
- I (Paul) ran across this example of applying Bayes' rule to a real world prob-

lem. It was created by a grad school friend of mine and is hilarious (lots of Cat Memes). I did notice that there is a mistake in the math at the 8:12 mark in the video (he states that $p(\text{alarm}|\text{no theft}) = 1 - p(\text{alarm}|\text{theft})$, which is not necessarily the case). It's still a good video though.

Exercise 4

It would be great to have something.

5 Marginalization Rule for Probabilities

The application of Bayes' rule often proceeded something like this. Let's define the event $\mathcal{D}$ as whether or not a person has a disease and $\mathcal{S}$ as the event that a particular symptom is observed. If we want to know $p(\mathcal{D}|\mathcal{S})$ we can apply Bayes' rule.

$$p(\mathcal{D}|\mathcal{S}) = \frac{p(\mathcal{S}|\mathcal{D})p(\mathcal{D})}{p(\mathcal{S})} \tag{5}$$

In order to calculate $p(\mathcal{S})$, some of the resources simply gave a number (e.g., in the Khan Academy video you Googled and got this value), used a convenient trick to get it (as in Allen's M&M example), or used the following calculation (as in the Veritasium and Car Alarm videos).

$$p(\mathcal{S}) = p(\mathcal{D})p(\mathcal{S}|\mathcal{D}) + p(\neg\mathcal{D})p(\mathcal{S}|\neg\mathcal{D}) \tag{6}$$

We wanted to revisit this calculation as it is hiding away some pretty powerful and interesting stuff. This calculation can be derived using the technique of marginalizing a probability measure function. The basic idea is that you want to compute the probability of some event, $\mathcal{A}$. However, it may be difficult to directly calculate $p(\mathcal{A})$ instead you can introduce another event, $\mathcal{B}$, and write $p(\mathcal{A})$ as:

$$p(\mathcal{A}) = p(\mathcal{A}, \mathcal{B}) + p(\mathcal{A}, \neg\mathcal{B}) \tag{7}$$

In the equation above we sometimes say that we are *marginalizing out* $\mathcal{B}$ (by summing over the two possibilities: that $\mathcal{B}$ occurred and that $\mathcal{B}$ did not occur).

Exercise 5 (15 minutes)

Using Equation 7 and other rules of probability you've learned thus far, show that Equation 5 is true.

☆ Solution

$$
\begin{aligned}
p(\mathcal{S}) &= p(\mathcal{S}, \mathcal{D}) + p(\mathcal{S}, \neg\mathcal{D}) && \text{marginalization property, Eq 7} \\
&= p(\mathcal{D})p(\mathcal{S}|\mathcal{D}) + p(\neg\mathcal{D})p(\mathcal{S}|\neg\mathcal{D}) && \text{product rule}
\end{aligned}
$$

Note that it was up to us what order we applied the product rule. If we had first split out $\mathcal{S}$ when going from line 1 to line 2 of our solution, we would have been left with $p(\mathcal{S})p(\mathcal{D}|\mathcal{S}) + p(\mathcal{S})p(\neg\mathcal{D}|\mathcal{S})$. This move wouldn't really make any progress towards a solution (since we still don't know $p(\mathcal{S})$.

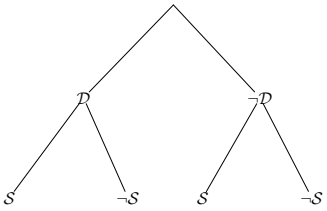
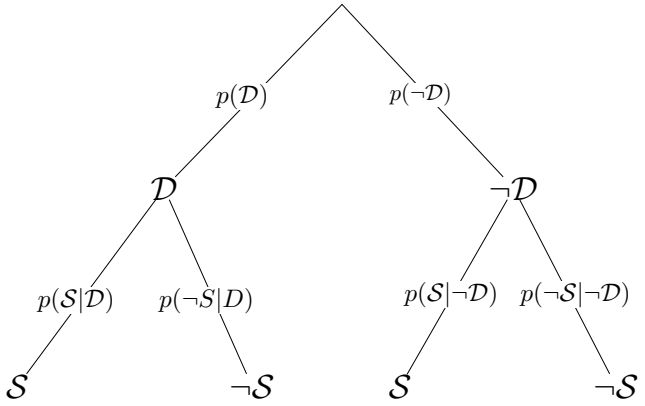


Figure 2: A tree diagram of the events $\mathcal{D}$ (has a disease) and $\mathcal{S}$ (has a symptom).

Another way to think about marginalization is to draw a tree where you have the event, which you'd like to know the probability of ($\mathcal{S}$ in the previous exercise) and the variable you are marginalizing out ($\mathcal{D}$ in the previous exercise) at the next junction in the tree (see Figure 2).

Further, we can annotate the arrows with the conditional probability of the event conditioned on the things further up in the tree.



If you want to find the probability of anyone joint probability (e.g., $p(\mathcal{D}, \neg\mathcal{S})$), you can follow corresponding path, multiplying probabilities as you go. For example, from examining the graph above we get

$$p(\mathcal{D}, \neg\mathcal{S}) = p(\mathcal{D})p(\neg\mathcal{S}|\mathcal{D}) \ . \tag{8}$$

A marginal probability for an event (e.g., $p(\mathcal{S})$) can be found by summing over all paths that arrive at the event. For example, from examining the graph above we get

$$p(\mathcal{S}) = p(\mathcal{D})p(\mathcal{S}|\mathcal{D}) + p(\neg\mathcal{D})p(\mathcal{S}|\neg\mathcal{D}) \ . \tag{9}$$

Exercise 6 (20 minutes)

Do problem 2 from this assignment. They use $\mathcal{E}'$ to refer to the event $\neg\mathcal{E}$.

☆ Solution

Solution is embedded in the link.

Exercise 7 (30 minutes)

Applying Bayes' Rule. Have a few options for where Bayes could be applied. Have them pick one of them and figure out an estimate.

6 Random Variables

We've talked about the concept of an event that captures whether something happens as a result of some random process. It turns out that it is very useful in machine learning and probabilistic modeling to talk about a variable that captures some quantity that is a result of some random process. We call this entity a *random variable*.

External Resource(s) (20 minutes)

Watch the following two videos from Khan Academy on Random Variables.

- Khan Academy Video on Random Variables.
- Discrete and Continuous Random Variables (note: for now we are interested in discrete random variables).

Now that you have a sense of what a random variable is, we'll introduce the notion of probability mass function (or PMF). A PMF is a function that assigns a probability to a discrete random variable taking on a particular value as a result of a random process. We'll use capital, unbolded letters to refer to random variables (e.g., the random variable X) and $p(X = k)$ to refer to the probability that X takes on value k).

For example, if we were flipping a fair coin 3 times, we might define a random variable X as follows.

$$X = \text{the number of coins that come up heads in 3 flips} \tag{10}$$

The probability mass function is then defined over all possible values that X could possibly take on. If don't know how we arrived at these values, that is fine. These are results that can be derived using basic combinatorics (go take Sarah Spence Adams' class and learn about how to do this).

$$p(X = 0) = p(0 \text{ heads come up}) = \frac{1}{8}$$
$$p(X = 1) = p(1 \text{ heads come up}) = \frac{3}{8}$$
$$p(X = 2) = p(2 \text{ heads come up}) = \frac{3}{8}$$
$$p(X = 3) = p(3 \text{ heads come up}) = \frac{1}{8}$$

7 Basic Bayes in Python

External Resource(s) (30 minutes)

Go through the assignment 1 companion notebook.