Module 2: Review of Main Concepts

Machine Learning

Fall 2019

Learning Objectives

- Review some of the key formulas we've learned.
- Try some practice problems to firm up the details.

1 Motivation for the Creation of this Document

Some folks have expressed that they are having a hard time remembering all of the things that they've learned this module. We are putting this document together to have a single place that lists the main mathematical and algorithms content from this module. We are aiming for a relatively concise resource, so we are avoiding long explanations. Eventually, we might (or you might via NB?) add pointers to the original assignments or external resources that explain this stuff more fully. An exception to this is that in some cases we have put an additional worked example (e.g., in the maximum likelihood estimation section).

Notice

By creating this document we are not trying to elevate the importance of the math / algorithms portion of this class over the context and ethics or programming parts. We are creating this resource in response to specific requests from students who wanted more opportunities to reinforce the math and algorithms content.

2 What this Document Is and Is Not

This document is not: a standalone resource for learning all of these topics (the explanations provided here are complementary to those given in previous assignments on these same topics), representative of all three streams of the class, a resource for learning new material / concepts not covered in other assignments, something you need to engage with in any particular way (see the next section for some suggestions on how to engage).

This document is: a place to assess your understanding of the math and algorithms parts of this module, a place to practice these concepts, a place to ask clarifying / enriching questions (via NB).

3 How to Engage

You should use this document as a way to solidify concepts which you feel a bit shaky about. As a result, if you feel good about a topic, there is no need for you to engage with a particular section. For instance, if you have the basic idea of probability spaces down, don't even worry about reading that section. If you would like reinforcement on a topic, read the appropriate section and try some of the practice problems. You may find it useful to consult the original

assignment that introduced these concepts. Additionally, **you should post on NB if you have a question or would just like more detail on a particular topic in this document.**

4 Probability

Probability gives us a formal language to express various forms of uncertainty. This is hugely valuable when doing machine learning, which often involves many forms of uncertainty (e.g., missing data, model uncertainty, noise in training data, etc.).

4.1 Probability Space

Suppose we want to describe some random process in terms of probability. To do so, we **define a probability space**. A probability space consists of two things.

- **Events:** these are things that may or may not occur as a result a random process. For example, the event $\mathcal{H}$ might represent the event that when a coin is flipped it comes up heads.
- **Probability measure:** this is a function, often called p, that assigns a probability to any event.

  In order for p to be a valid probability measure, it must satisfy these three properties.

1. For any event $\mathcal{E}$, $0 \leq p(\mathcal{E}) \leq 1$ (probabilities range from 0, for an impossible event, to 1, for a certain event).
2. For any set of disjoint events, $\mathcal{E}_1, \mathcal{E}_2, \ldots \mathcal{E}_n$ (disjoint events are those that cannot co-occur),

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

3. For any set of exhaustive events, $\mathcal{E}_1, \mathcal{E}_2, \ldots \mathcal{E}_n$ (an exhaustive set of events means at least one *must* occur),

$$p(\mathcal{E}_1 \text{ or } \mathcal{E}_2 \ldots \text{ or } \mathcal{E}_n) = 1 \ . \tag{2}$$

Exercise 1

Consider the random process of rolling a fair, 6-sided die. Define a probability space (events and a probability measure) to represent this random process. You can define your own set of events, but an easy default is to define an event for each possible outcome of the roll (e.g., $\mathcal{E}_1$ is the die came out as a 1, $\mathcal{E}_2$ is the die came out as 2, etc.).

Exercise 2

Construct a probability space to model the 2020 U.S. Congressional Elections. Define some events of interest (e.g., a particular party controlling one of the houses of congress) and define a valid probability measure over these events (just use your intuition to pick reasonable values, but do adhere to the rules for probability measure specified above).

4.2 Random Variables

A random variable is a generalization of an event. **A random variable is an entity that takes on a value as a result of a random process.** For instance, one might define a random variable D that captures rolling a 6-sided die (D would take on a value from the set $\{1, 2, 3, 4, 5, 6\}$). (Note: we can also express this random process using events (as we did in the exercise above); both formulations are valid). A random variable consists of two things.

- **A mapping from each possible outcome of the random process to a value for the random variable** (e.g., our random variable D takes on the value 1 when the die roll comes up 1, value 2 when the die roll comes up 2, etc.).
- **A probability mass function (PMF), which gives the probability that a random variable takes on a particular value.** For example, $p(D = 1)$ is the probability the 6-sided die comes up 1. If the die is fair, $p(D = 1) = \frac{1}{6}$.

  Similar to the conditions outlined for a probability measure, a PMF must satisfy the following conditions.

  1. If V is the set of all possible values that the random variable X can take on, then $0 \leq p(X = x) \leq 1$ for any value x in the set V.
  2. If we add the probability of all possible values that X can take on, we should get 1. That is, $\sum_{x \in V} p(X = x) = 1$.

Exercise 3

Choose a random process and a quantity that would result from that random process. Define that quantity as a random variable (that is, what values can it take on, what would be a reasonable PMF). If you pick a pretty simple system you may be able to specify the PMF exactly (e.g., for rolling a die), but for a more complicated problem consider specifying what the PMF looks like at a qualitative level (the solution has an example like this).

⚠ Notice

In pretty much all of the content in this module any rule that works for events will also work for random variables. For instance, Bayes' rule for events $\mathcal{A}$ and $\mathcal{B}$ is $p(\mathcal{A}|\mathcal{B}) = \frac{p(\mathcal{B}|\mathcal{A})p(\mathcal{A})}{p(\mathcal{B})}$. This equation looks the same as Bayes' rule for two random variables X and Y, $p(X = x|Y = y) = \frac{p(Y=y|X=x)p(X=x)}{p(Y=y)}$. As a result, when we present a rule for manipulating the probability of various events, you can also assume that it will work with little modification for random variables. In order to make this document simpler, we won't explicitly give the analogous formula for random variables, but if you have any questions on what it would look like, please post on NB.

4.3 Complement Rule

If we know the probability of an event $\mathcal{E}$ occurring, then the probability of it not occurring $p(\neg\mathcal{E})$ is given by the formula

$$p(\neg\mathcal{E}) = 1 - p(\mathcal{E}) \ . \tag{3}$$

Exercise 4

If you flip a fair coin 100 times, the probability of getting all 100 heads is $\left(\frac{1}{2}\right)^{100}$. What's the probability of getting at least one tails.

4.4 Conditional Probability

A conditional probability tells us the probability of some event occurring assuming (or conditioned on) another event having occurred. For instance, we could say "what is the probability that we observe a particular symptom given that a person has a disease?" Conditional probability is defined using the following equation.

$$p(\mathcal{A}|\mathcal{B}) = \frac{p(\mathcal{A}, \mathcal{B})}{p(\mathcal{B})} \tag{4}$$

The term $p(\mathcal{A}, \mathcal{B})$ is known as a joint probability, which we will define in the next section.

4.5 Joint Probability

The probability of two events, $\mathcal{A}$ and $\mathcal{B}$, both occurring is called the joint probability of $\mathcal{A}$ and $\mathcal{B}$. We write this as:

$$p(\mathcal{A}, \mathcal{B}) = \text{the probability of both } \mathcal{A} \text{ and } \mathcal{B} \text{ simultaneously occurring} \tag{5}$$

For any two events $\mathcal{A}$ and $\mathcal{B}$, we can write the joint probability in terms of the product of a marginal probability (the probability of one of the events) and a conditional probability.

$$p(\mathcal{A}, \mathcal{B}) = p(\mathcal{A})p(\mathcal{B}|\mathcal{A}) \tag{6}$$

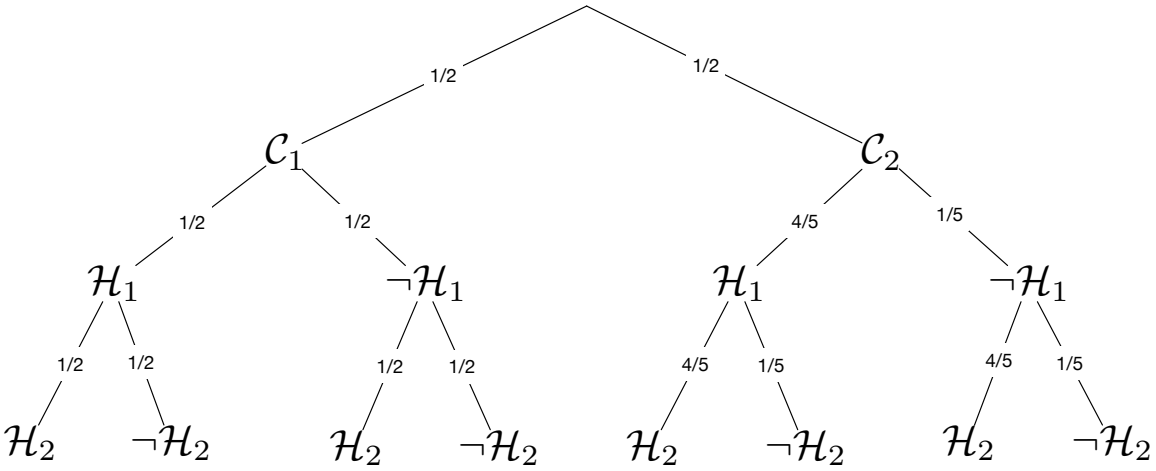
$$= p(\mathcal{B})p(\mathcal{A}|\mathcal{B}) \tag{7}$$

4.6 Marginalization

If we can easily compute the joint distribution of two events (that is $p(\mathcal{A}, \mathcal{B})$), then we can use the technique of marginalization to obtain the *marginal distribution* (the probability of just one of the events in isolation).

$$p(\mathcal{A}) = p(\mathcal{A}, \mathcal{B}) + p(\mathcal{A}, \neg\mathcal{B}) \tag{8}$$

Another way to think about marginalization is to draw a tree. For example, this is tree from a previous assignment that described the process of choosing between two coins ($\mathcal{C}_1$ and $\mathcal{C}_2$) and flipping the chosen coin twice. The first coin is fair and the second coin is biased (probability of heads is 4/5).



If we want to compute the marginal probability of any of the events $\mathcal{H}_1, \mathcal{H}_2, \mathcal{C}_1, \mathcal{C}_2$, we can simply add the probability of each path that contains the event of interest. The probability of a path is the product of the probabilities along it.

For example,

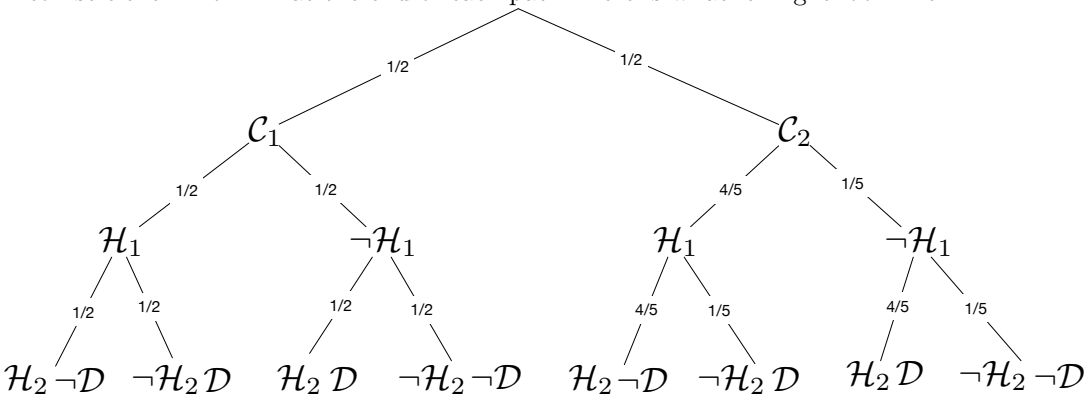
$$\begin{aligned} p(\mathcal{H}_2) &= p(\mathcal{C}_1, \mathcal{H}_1, \mathcal{H}_2) + p(\mathcal{C}_1, \neg\mathcal{H}_1, \mathcal{H}_2) + p(\mathcal{C}_2, \mathcal{H}_1, \mathcal{H}_2) + p(\mathcal{C}_2, \neg\mathcal{H}_1, \mathcal{H}_2) \\ &= \left(\frac{1}{2} \times \frac{1}{2} \times \frac{1}{2}\right) + \left(\frac{1}{2} \times \frac{1}{2} \times \frac{1}{2}\right) + \left(\frac{1}{2} \times \frac{4}{5} \times \frac{4}{5}\right) + \left(\frac{1}{2} \times \frac{1}{5} \times \frac{4}{5}\right) \\ &= \frac{1}{8} + \frac{1}{8} + \frac{8}{25} + \frac{2}{25} \\ &= \frac{13}{20} \end{aligned}$$

It's worth giving the translation of this to random variables explicitly. If X and Y are random variables and V contains all possible values that X can take on, then

$$p(Y = y) = \sum_{x \in V} p(Y = y, X = x) \ . \tag{9}$$

Exercise 5

Let's define the event $\mathcal{D}$ as the two flips in the coin problem above having different outcomes (e.g., one is heads and one is tails). Using the technique of marginalization, what is $p(\mathcal{D})$? If it helps, consider augmenting the tree of the coin problem to list either $\mathcal{D}$ or $\neg\mathcal{D}$ at the end of each path. Here is what it might look like.



4.7 Product Rule

We can decompose the joint probability of a bunch of events into a product of probabilities. Suppose $\mathcal{E}_1, \mathcal{E}_2, \ldots, \mathcal{E}_n$ represent events, then

$$p(\mathcal{E}_1, \mathcal{E}_2, \ldots, \mathcal{E}_n) = p(\mathcal{E}_1)p(\mathcal{E}_2|\mathcal{E}_1)p(\mathcal{E}_3|\mathcal{E}_1, \mathcal{E}_2)\ldots p(\mathcal{E}_n|\mathcal{E}_1, \ldots, \mathcal{E}_{n-1}) \ . \tag{10}$$

The rule we're applying here is to start with an event conditioned on nothing, then multiply by the next event conditioned on the previous event, then multiply by the next event conditioned on the previous two, etc. The order in which you select

the events is also arbitrary, so if $n = 3$, the following are equivalent.

$$\begin{aligned} p(\mathcal{E}_1, \mathcal{E}_2, \mathcal{E}_3) &= p(\mathcal{E}_1)p(\mathcal{E}_2|\mathcal{E}_1)p(\mathcal{E}_3|\mathcal{E}_1, \mathcal{E}_2) \\ &= p(\mathcal{E}_2)p(\mathcal{E}_3|\mathcal{E}_2)p(\mathcal{E}_1|\mathcal{E}_2, \mathcal{E}_3) \\ &= p(\mathcal{E}_3)p(\mathcal{E}_2|\mathcal{E}_3)p(\mathcal{E}_1|\mathcal{E}_3, \mathcal{E}_2) \\ &\ldots \text{ there are three more potential orderings that we won't give explicitly} \end{aligned}$$

Exercise 6

Here is a review problem from Khan academy that can be solved most straightforwardly using the product rule.

A goblet contains 3 red balls, 2 green balls, and 6 blue balls.

If we choose a ball, then another ball without putting the first one back in the goblet, what is the probability that the first ball will be red and the second will be blue?

4.8 Bayes' Rule

Bayes' rule lets you take a conditional probability $p(\mathcal{A}|\mathcal{B})$ and flip the order of the events across the conditioning bar.

$$p(\mathcal{A}|\mathcal{B}) = \frac{p(\mathcal{B}|\mathcal{A})p(\mathcal{A})}{p(\mathcal{B})} \tag{11}$$

There are a few alternate forms of Bayes' rule.

- You can move multiple events through the conditioning bar (here are two examples where we move two events, but you can move any number of events).

$$p(\mathcal{A}, \mathcal{B}|\mathcal{C}) = \frac{p(\mathcal{C}|\mathcal{A}, \mathcal{B})p(\mathcal{A}, \mathcal{B})}{p(\mathcal{C})} \tag{12}$$

$$p(\mathcal{A}|\mathcal{B}, \mathcal{C}) = \frac{p(\mathcal{B}, \mathcal{C}|\mathcal{A})p(\mathcal{A})}{p(\mathcal{B}, \mathcal{C})} \tag{13}$$

- You don't have to swap all of the events across the conditioning bar (e.g., below, we leave $\mathcal{C}$ on the righthand side of the bar).

$$p(\mathcal{A}|\mathcal{B}, \mathcal{C}) = \frac{p(\mathcal{B}|\mathcal{A}, \mathcal{C})p(\mathcal{A}|\mathcal{C})}{p(\mathcal{B}|\mathcal{C})} \tag{14}$$

What's the big deal about Bayes? Here are two potential answers to this.

- **It helps us compute probabilities of interest.** Sometimes it is much more natural to compute $p(\mathcal{B}|\mathcal{A})$, then $p(\mathcal{A}|\mathcal{B})$. For example, if $\mathcal{A}$ represents the event that someone has a particular disease and $\mathcal{B}$ represents someone exhibiting a particular symptom, since we think of the disease as causing the symptom it may be easier to model the probability of the symptom given the disease. It is less natural to think of the probability of the disease given the symptom since we don't typically think of a symptom as causing a disease.
- Consider watching Julia Galef's Bayes: How one equation changed the way I think.

Exercise 7

There are lots of practice problems on the web for Bayes' rule. Here are some suggestions.

(a) Former Olin Professor Sanjoy Mahajan has a bunch of practice problems on Bayes' rule available on his website. Each problem allows you to submit your answer and have it checked automatically. He also has the solution for each problem posted. Some of the problems are more standalone than others (e.g., some rely on material communicated in class). We'll suggest some that seem to be standalone here. First, visit the HW02 page on Sanjoy's website (if you get a privacy error from your browser, go ahead and allow your browser to access the site. What could go wrong?). On this page, consider doing **problem 2** (this one is using Bayes' rule in reverse, so it is a bit of a twist), **problem 3** (part (a) is really a question about marginalization), **problem 4**, and **problem 5**.

(b) Here are some more practice problems from a class on Bayesian statistics.

4.9 Independence

Two events $\mathcal{A}$ and $\mathcal{B}$ are independent (written as $\mathcal{A} \perp\!\!\!\perp \mathcal{B}$) if and only if

$$p(\mathcal{A}, \mathcal{B}) = p(\mathcal{A})p(\mathcal{B}) \ . \tag{15}$$

This equation also implies the following very useful rule.

$$p(\mathcal{A}|\mathcal{B}) = p(\mathcal{A}) \ . \tag{16}$$

Intuitively, we can drop $\mathcal{B}$ from the right side of the conditioning bar since knowing that $\mathcal{B}$ occurred doesn't change the probability of $\mathcal{A}$.

Exercise 8

Simplify each expression using the fact that $\mathcal{A} \perp\!\!\!\perp \mathcal{B}$ and $\mathcal{C} \perp\!\!\!\perp \mathcal{D}$ (don't worry about doing anything fancy, just apply the product rule and the basic definitions of independence above). To make things simpler, assume that these independence relationships hold no matter what events you condition on.

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

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

4.10 Conditional Independence

Two events $\mathcal{A}$ and $\mathcal{B}$ are conditionally independent given a third event $\mathcal{C}$ (written as $\mathcal{A} \perp\!\!\!\perp \mathcal{B} \mid \mathcal{C}$) if and only if

$$p(\mathcal{A}, \mathcal{B}|\mathcal{C}) = p(\mathcal{A}|\mathcal{C})p(\mathcal{B}|\mathcal{C}) \ . \tag{17}$$

This equation also implies the following very useful rule.

$$p(\mathcal{A}|\mathcal{B}, \mathcal{C}) = p(\mathcal{A}|\mathcal{C}) \ . \tag{18}$$

The intuition is that we can drop $\mathcal{B}$ from the right side of the conditioning bar since knowing that $\mathcal{B}$ occurred doesn't change the probability of $\mathcal{A}$ if we already know that $\mathcal{C}$ occurred.

Exercise 9

Suppose we know the following facts about some events $\mathcal{A}, \mathcal{B}, \mathcal{C}, \mathcal{D}$.

- $\mathcal{A} \perp\!\!\!\perp \mathcal{B}$
- $\mathcal{A} \perp\!\!\!\perp \mathcal{C} \mid \mathcal{B}$
- $\mathcal{A} \perp\!\!\!\perp \mathcal{D} \mid \mathcal{C}$

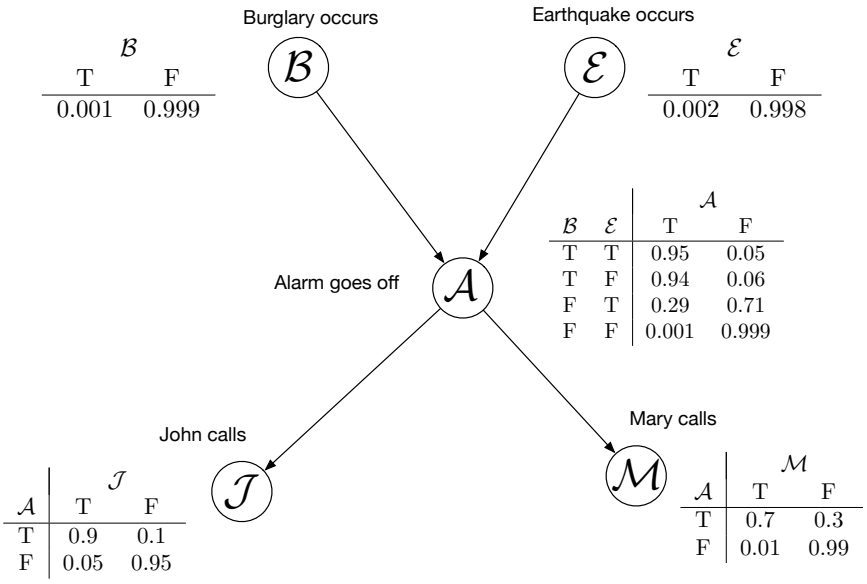
Simplify the expression $p(\mathcal{A}, \mathcal{B}, \mathcal{C}, \mathcal{D})$ as much as possible (again, don't worry about doing anything fancy, just apply the product rule and the basic definitions of independence and conditional independence above).

5 Bayesian Networks

A Bayesian Network (BN) defines a model consisting of one or more events or random variables. A BN consists of the following components.

- A **graph, that encodes the dependencies between one or more random variables and events.** This graph contains nodes that represent random variables or events and directed edges that indicate causal relationships between nodes (e.g., an edge from from node A to node B indicates that A causes B).
- A **conditional probability table that indicates the probability that a node takes on a particular value given the values of its parents.**

Here is an example of a BN that represents how two people, Mary and John, respond to an alarm in their apartment complex. The alarm is triggered by earthquakes and burglaries.



$\mathcal{B}$	
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

The circles in this graph (called nodes) are the various events. Each of these can either be true or false (as stated before, nodes can also represent random variables that can take on multiple values). The arrows indicate causal relationships

(e.g., the alarm going off is caused by burglaries and earthquakes). The tables next to each node give us the probability of the node conditioned on its parents.

The condition that must hold for any BN is that if we write the joint distribution of all of the random variables (or events, since 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).

$$p(X_1, X_2, \dots, X_n) = p(X_1|Pa(X_1))p(X_2|Pa(X_2)) \dots p(X_n|Pa(X_n)) \tag{19}$$

Exercise 10

For practice, use Equation 19 to factor $p(\mathcal{B}, \mathcal{E}, \mathcal{A}, \mathcal{J}, \mathcal{M})$ for the alarm network.

5.1 D-Separation

One of the beautiful things about BNs is that they allow us to infer independence relationships from the graph structure. In assignment 2 we gave some resources to learn about d-separation. We encourage you to take a second look at those if you are having trouble (one was a video and one was a document). If you find any other good resources on d-separation, please share them here. Next, we'll try to state the rules of d-separation as clearly and concisely as possible.

🔁 Recall: Definition of D-Separation

Two nodes, $\mathcal{A}$ and $\mathcal{B}$, in a BN are d-separated when conditioning on a set of nodes, $\mathcal{C}$, if and only if *all* paths between $\mathcal{A}$ and $\mathcal{B}$ are blocked when conditioning on $\mathcal{C}$ (the concept of *blocked* will be defined below). Note that a path consists of traversing one or more edges, in sequence, of a BN (irrespective of the direction of the edge). A path between $\mathcal{A}$ and $\mathcal{B}$ conditioned on nodes $\mathcal{C}$ **is blocked if any of the following conditions are met**.

1. The path contains a collider and neither the collider nor any of its descendants are in the set $\mathcal{C}$ (a collider is a node where the path consists of two incoming arrows, e.g., $\mathcal{D} \rightarrow \mathcal{E} \leftarrow \mathcal{F}$).
2. The path contains a segment that looks like $\mathcal{D} \rightarrow \mathcal{E} \rightarrow \mathcal{F}$ and $\mathcal{E}$ is in the set $\mathcal{C}$ (i.e., we are conditioning on $\mathcal{E}$).
3. The path contains a segment that looks like $\mathcal{D} \leftarrow \mathcal{E} \rightarrow \mathcal{F}$ and $\mathcal{E}$ is in the set $\mathcal{C}$ (i.e., we are conditioning on $\mathcal{E}$).

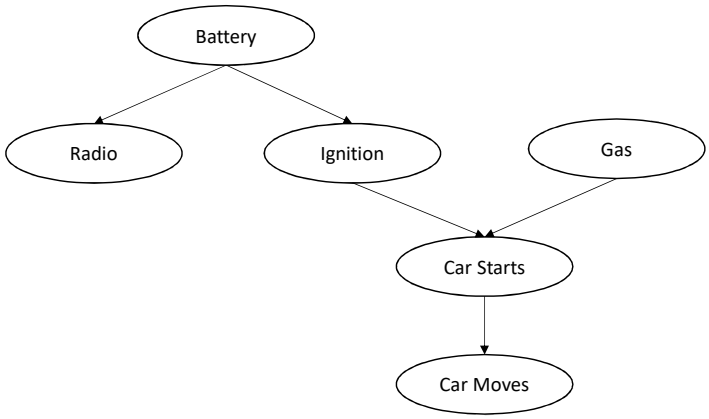
If $\mathcal{A}$ and $\mathcal{B}$ are d-separated conditioned on $\mathcal{C}$, then $\mathcal{A} \perp\!\!\!\perp \mathcal{B} \mid \mathcal{C}$.

Exercise 11

Choose two nodes in the alarm network. Are they independent? How would this change if you conditioned on each of the other three nodes in the network? Repeat this exercise for other pairs of nodes in the network.

Exercise 12

Consider this network that describes the operation of various parts of a car (this is from Slide 20 of Sven Koenig's lecture on Bayesian Networks).



Using the rules of d-separation, determine whether the following conditional independence relationships are true or false (excuse the upper case letters, the cool calligraphic math font doesn't support lowercase letters).

- $\mathcal{RADIO} \perp\!\!\!\perp \mathcal{IGNITION}$?
- $\mathcal{RADIO} \perp\!\!\!\perp \mathcal{IGNITION} \mid \mathcal{BATTERY}$?
- $\mathcal{RADIO} \perp\!\!\!\perp \mathcal{GAS}$?
- $\mathcal{RADIO} \perp\!\!\!\perp \mathcal{GAS} \mid \mathcal{CARMOVES}$?
- This one is a pretty rich network. You could keep going to create new practice problems.

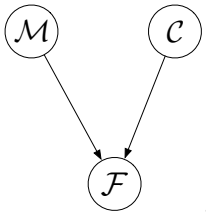
Exercise 13

Consider the following BN that represents a situation facing a doctor trying to diagnose a newborn that comes into the hospital with a fever. In this scenario there are two possible underlying diagnoses that could cause this symptom of having a fever (represented by the event $\mathcal{F}$): a common cold virus (represented by the event $\mathcal{C}$) or meningitis (represented by the event $\mathcal{M}$).

Note: these are not real numbers, but display some of the correct trends.

$\mathcal{M}$

T	F
$1/10000$	$9999/10000$



$\mathcal{C}$

T	F
$1/100$	$99/100$

$\mathcal{M}$	$\mathcal{C}$	$\mathcal{F}$: T	$\mathcal{F}$: F
T	T	$99/100$	$1/100$
F	T	$20/100$	$80/100$
T	F	$99/100$	$1/100$
F	F	$1/100$	$99/100$

(a) Determine $p(\mathcal{C}|\mathcal{F})$

(b) Determine $p(\mathcal{C}|\mathcal{F}, \mathcal{M})$. Why is the effect that conditioning on $\mathcal{M}$ known as *explaining away*?

If you're interested in exploring this idea further (e.g., in the final project), you can visit a symptom checker (like the Mayo Clinic symptom checker or the WebMD Symptom Checker). Also, consider reading (well probably skimming, it's pretty dense) Screening tests: a review with examples.

6 *Fairness Criteria for ML*

Chapter 2 from *Fairness and machine learning* laid out some formal frameworks for fairness. In all of these definitions, we refer to the following random variables.

Variable	Meaning
R	This is the prediction, or response, of the algorithm (e.g., predict recidivate)
Y	This is the actual outcome (e.g., did the person recidivate or not)
A	This is a sensitive attribute (e.g., it might encode the race of the person being evaluated)

- Independence: $R \perp\!\!\!\perp A$. This requires that the prediction is independent of the sensitive attribute (without conditioning on anything). Suppose R is binary (e.g., $R = 1$ if you approve someone for a loan and $R = 0$ otherwise). Further, suppose A can take on two values $A = a$ corresponds to some group of people and $A = b$ corresponds to another group of people.

$$p(R = 1 \mid A = a) = P(R = 1 \mid A = b) \tag{20}$$

  This means that the loan approval rate must be identical for both groups (irrespective of any correlations between being a member of these groups and other variables that may be relevant for predicting the riskiness of a loan).

- Separation: $R \perp\!\!\!\perp A \mid Y$. This condition requires that the response and the sensitive attribute must be independent given the actual outcome. Supposing that R is binary (e.g., loan approved or not), A can take on two values (a or b), and Y is binary (e.g., $Y = 1$ means loan was repaid, $Y = 0$ means loan was defaulted upon), we require the following must hold.

$$p(R = 1 \mid A = a, Y = 1) = p(R = 1 \mid A = b, Y = 1) \tag{21}$$
$$p(R = 1 \mid A = a, Y = 0) = p(R = 1 \mid A = b, Y = 0) \tag{22}$$
$$p(R = 0 \mid A = a, Y = 1) = p(R = 0 \mid A = b, Y = 1) \tag{23}$$
$$p(R = 0 \mid A = a, Y = 0) = p(R = 0 \mid A = b, Y = 0) \tag{24}$$

  These equations state that the true positive rate (Equation 21), false positive rate (Equation 22), false negative rate (Equation 23), and true negative rate (Equation 24) must be equal for group $A = a$ and group $A = b$. This was the definition of fairness that Propublica used in their analysis.

- Sufficiency: $Y \perp\!\!\!\perp A \mid R$. This condition requires that the outcome and the attribute must be conditionally independent given the algorithm's prediction. Supposing that R is binary (e.g., loan approved or not), A can take on two values (a

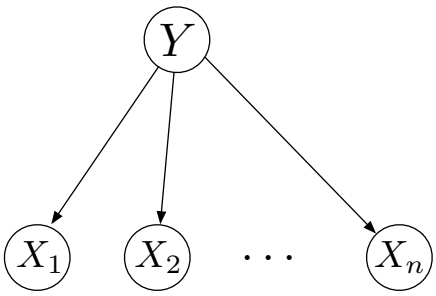
or b), and Y is binary (e.g., $Y = 1$ means loan was repaid, $Y = 0$ means loan was defaulted upon), we require the following must hold.

$$p(Y = 1 \mid A = a, R = 1) = p(Y = 1 \mid A = b, R = 1) \tag{25}$$
$$p(Y = 0 \mid A = a, R = 0) = p(Y = 0 \mid A = b, R = 0) \tag{26}$$

These equations state that the positive predictive value (Equation 25) and negative predictive value (Equation 26) must be equal for group $A = a$ and group $A = b$. This was the definition of fairness that NorthPointe used in their analysis

7 *Naïve Bayes*

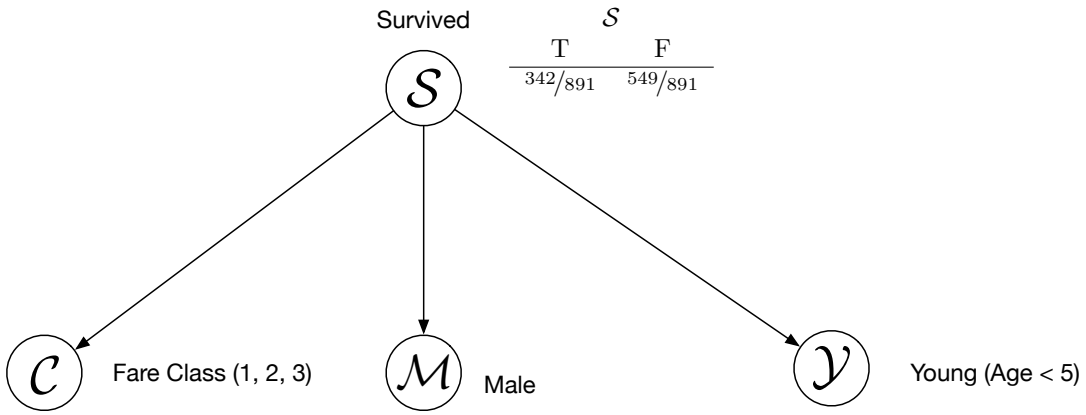
The Naïve Bayes model can be described using the following BN.



Remember, that in order to define the BN we need to also specify the conditional probability tables (the probability of each node conditioned on its parents). For the Naïve Bayes model this consists of the following probabilities.

- If Y is a random variable, we need $p(Y = i)$ (for all possible values, i, that Y could take on). If $\mathcal{Y}$ is an event, we would need $p(\mathcal{Y})$.
- If the X_i's are random variables, we need $p(X_i = j|Y = k)$ (for all values, k, that Y can take on and all possible values, j, that the X_i can take on). If $\mathcal{X}_i$ is an event, we would need $p(\mathcal{X}_i \mid \mathcal{Y})$ and $p(\mathcal{X}_i \mid \neg\mathcal{Y})$.

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	$^{13}/_{549}$	$^{536}/_{549}$

You'll notice that one of the nodes in this graph is a random variable (*fare class*) and the rest are events (*survived*, *male*, and *young*). The conditional probabilities were determined using the technique of maximum likelihood estimation (MLE) (which we will describe later). In this case, MLE simply consists of counting the number of times one of the nodes takes on a particular value when the parent also takes a particular value and normalizing over all possible values the node can take on (more on this in the next section).

8 Maximum Likelihood Estimation

In assignment 4, we derived the maximum likelihood estimates for the parameters of the Naïve Bayes model. Rather than redo the derivation here, we're going to do two things. First, we're going to derive maximum likelihood estimates for a simpler problem. Next, we're going to apply the maximum likelihood technique to data from the Titanic network. We hope that these two things will help you wrap your minds around MLE.

8.1 Flipping Coins

Suppose we have a coin and we want to know what the probability of it coming up heads is. Let's call the probability that the coin comes up heads Θ. Suppose we flip the coin n times. Let's define a random variable X_i that takes on value 1 when the coin comes up heads and value 0 when the coin comes up tails. We'll use lower case x_i to reference the specific value that we observed for the ith flip. We'd like to determine the maximum likelihood estimate of Θ. We can write this goal formally using the following equation.

$$\Theta^\star = \arg\max_\Theta p(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n | \Theta) \tag{27}$$

We can simplify the equation above by making the independently and identically distributed (i.i.d.) assumption. That is, we'll assume that once we know our model parameter, Θ, each of the random variables X_i is conditionally independent of any other random variable X_j. More formally, we have $X_i \perp\!\!\!\perp X_j \mid \Theta$ for all $i \neq j$. Intuitively this means that once we know the bias of the coin, the outcome of each flip has no bearing on the outcome of any other flips. Applying the i.i.d. assumption gives us

$$\Theta^\star = \arg\max_\Theta p(X_1 = x_1|\Theta)p(X_2 = x_2|\Theta)\dots p(X_n = x_n|\Theta) \ . \tag{28}$$

We can now take the log of the expression inside of our arg max without affecting the maximum value.

$$\begin{aligned}\Theta^\star &= \arg\max_\Theta \log\left(p(X_1 = x_1|\Theta)p(X_2 = x_2|\Theta)\dots p(X_n = x_n|\Theta)\right) \\ &= \arg\max_\Theta \sum_{i=1}^{n} \log p(X_i = x_i|\Theta)\end{aligned} \tag{29}$$

The term $p(X_i = x_i|\Theta)$ will either be Θ if x_i is 1 or $1-\Theta$ if x_i is 0. We can indicate this in our equation by using the indicator function, $\mathbb{I}$, that takes on value 1 when the condition inside is true and 0 when the condition inside is false.

$$\Theta^\star = \arg\max_\Theta \sum_{i=1}^{n} \left(\mathbb{I}[x_i = 1]\log\Theta + \mathbb{I}[x_i = 0]\log\left(1-\Theta\right)\right) \tag{30}$$

Examining Equation 30, we can see that the value inside the summation only depends on whether x_i is 1 or 0. Therefore, if n_H is the number of heads we observed (heads corresponds to $x_i = 1$) and n_T is the number of tails we observed (tails corresponds to $x_i = 0$), we can rewrite Equation 30 as

$$\Theta^\star = \arg\max_\Theta n_H \log\Theta + n_T \log\left(1-\Theta\right) \ . \tag{31}$$

We can solve this equation by taking the derivative with respect to Θ and setting it to 0 (in Exercise 1 on Assignment 4 we had to use Lagrange multipliers since it was a harder problem; we don't need to do that here).

$$\begin{aligned}\frac{d}{d\Theta}\left(n_H \log\Theta + n_T \log\left(1-\Theta\right)\right) &= \frac{n_H}{\Theta} - \frac{n_T}{1-\Theta} \\ 0 &= \frac{n_H}{\Theta^\star} - \frac{n_T}{1-\Theta^\star} && \Theta^\star \text{ occurs at a critical point, i.e., the derivative is 0} \\ 0 &= n_H(1-\Theta^\star) - n_T\Theta^\star && \text{assuming } \Theta^\star \text{ is not 0 or 1, we can multiply by } \Theta^\star(1-\Theta^\star) \\ 0 &= n_H - n_H\Theta^\star - n_T\Theta^\star \\ (n_H + n_T)\Theta^\star &= n_H \\ \Theta^\star &= \frac{n_H}{n_H + n_T}\end{aligned} \tag{32}$$

While there's a lot of math to get to this final answer (which, tells us to estimate $\Theta^\star$ as the fraction of heads observed in the training data... pretty reasonable!) the steps are very formulaic.

1. Write down the likelihood of the data given the parameters.
2. Simplify the likelihood and substitute in the parameters (the parameters will be symbolic at this point, e.g., Θ).
3. Solve for the optimal value of the parameters (e.g., using Lagrange multipliers or setting derivative to 0 to find a critical point).

8.2 MLE and the Titanic

Some of you requested an example of applying the equations that we worked out in Assignment 4 to an actual example of Naïve Bayes. In that spirit, we will use training data to fit the parameters of our Titanic BN.

The MLE equations that we derived in Assignment 4 are as follows.

- To fit the parameters for the outcome variables Y, we can use this equation.

$$p(Y = i|\boldsymbol{\Theta}^\star) = \Theta^\star(i) = \frac{\text{ycount}(i)}{\sum_{j=1}^{c} \text{ycount}(j)} \tag{33}$$

- To fit the conditional probabilities of the features (e.g., young, male, and fare class), we can use the following equation.

$$p(X_j = k \mid Y = i, \boldsymbol{\Theta}^\star) = \Theta_j^\star(k|i) = \frac{\text{xcount}_j(k|i)}{\sum_{u=1}^{r} \text{xcount}_j(u|i)} \tag{34}$$

Let's apply these equations to the Titanic BN (the BN is shown earlier in this document).

- To determine $p(\mathcal{S})$ we need to count up the number of times each possible outcome occurred. We can use $\text{ycount}(\mathcal{S})$ and $\text{ycount}(\neg\mathcal{S})$ to denote the number of passengers who survived versus didn't survive.

$$\begin{aligned} p(\mathcal{S}) &= \frac{\text{ycount}(\mathcal{S})}{\text{ycount}(\mathcal{S}) + \text{ycount}(\neg\mathcal{S})} \\ &= \frac{342}{342 + 549} \\ &= \frac{342}{891} \end{aligned}$$

$p(\neg\mathcal{S})$ could be determined similarly, or you could just use $1 - p(\mathcal{S})$.

- In order to determine $p(\mathcal{M}|\mathcal{S})$ and $p(\mathcal{M}|\neg\mathcal{S})$, we need to know four counts: $\text{xcount}(\mathcal{M}|\mathcal{S}), \text{xcount}(\neg\mathcal{M}|\mathcal{S}), \text{xcount}(\mathcal{M}|\neg\mathcal{S})$, and $\text{xcount}(\neg\mathcal{M}|\neg\mathcal{S})$ (note: we dropped the subscripts on xcount since it is clear from the arguments which node in the BN we're talking about).

$$\begin{aligned} \Theta^\star(\mathcal{M}|\mathcal{S}) &= \frac{\text{xcount}(\mathcal{M}|\mathcal{S})}{\text{xcount}(\mathcal{M}|\mathcal{S}) + \text{xcount}(\neg\mathcal{M}|\mathcal{S})} \\ &= \frac{109}{109 + 233} \\ &= \frac{109}{342} \end{aligned}$$

$$\begin{aligned} \Theta^\star(\mathcal{M}|\neg\mathcal{S}) &= \frac{\text{xcount}(\mathcal{M}|\neg\mathcal{S})}{\text{xcount}(\mathcal{M}|\neg\mathcal{S}) + \text{xcount}(\neg\mathcal{M}|\neg\mathcal{S})} \\ &= \frac{468}{468 + 81} \\ &= \frac{468}{549} \end{aligned}$$

Exercise 14

Apply the formulas to determine the probabilities of $\mathcal{Y}$ and C given $\mathcal{S}$. Here are the relevant counts.

- xcount$(\mathcal{Y}|\mathcal{S}) = 27$, xcount$(\neg\mathcal{Y}|\mathcal{S}) = 315$
- xcount$(\mathcal{Y}|\neg\mathcal{S}) = 13$, xcount$(\neg\mathcal{Y}|\neg\mathcal{S}) = 536$
- xcount$(C = 1|\mathcal{S}) = 136$, xcount$(C = 2|\mathcal{S}) = 87$, xcount$(C = 3|\mathcal{S}) = 119$
- xcount$(C = 1|\neg\mathcal{S}) = 80$, xcount$(C = 2|\neg\mathcal{S}) = 97$, xcount$(C = 3|\neg\mathcal{S}) = 372$

9 N-Grams

This was primarily coding based, so we don't have too much here. Here are some basic ideas to remember about n-grams.

- N-Grams consist of N continuous symbols from a sequence. A common example of a sequence is text where the symbols are words. $N = 2$ gives us, as a special case, the bigram model and $N = 1$ gives us, as a special case, the unigram model.
- N-Grams can be used as features to represent a sequence (e.g., encoding the presence or absence of a particular N-Gram to summarize a document).
- These features can be used to classify the document or to compute various conditional probabilities (e.g., the probability that some symbol follows another).

10 Are we missing anything?

Keep in mind that this document's scope is within the math / algorithms part of this class (not on the context and ethics or programming side). Comment on NB here to request stuff to add.