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 they've learned this module. We are putting together this document to have a single place that lists all of the mathematical and algorithms content we've learned in 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 that explain this stuff more fully. An exception to this is that in some cases we have put additional worked examples into the document.

Notice

By creating this document we are not elevating the math / algs portion of this class over the context and ethics or programming parts. We are creating this resource in response to specific request from a subset of students for more opportunities to reinforce the math / algs for this module.

2 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.).

2.1 Probability Space

Suppose we want to describe some random process in terms of probability. In order 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 of our 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 function 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, $\mathcal{E}_1, \mathcal{E}_2, \dots \mathcal{E}_n$ (disjoint events are those that cannot co-occur),

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

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

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

2.2 Random Variables

A random variable is a generalization of an event. Think of a random variable as 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 the rolling a 6-sided die (D would take on a value from the set $\{1, 2, 3, 4, 5, 6\}$. A random variable consists of two things.

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

  Similar to the conditions for outlined for a probability measure function, 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, $\sum_{x \in V} p(X = x) = 1$.

⚠ 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.

2.3 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

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

2.4 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.

2.5 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.

$$\begin{aligned} p(\mathcal{A},\mathcal{B}) &= p(\mathcal{A})p(\mathcal{B}|\mathcal{A}) && (6) \\ &= p(\mathcal{B})p(\mathcal{A}|\mathcal{B}) && (7) \end{aligned}$$

2.6 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).

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

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

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

TODO: we can also think of marginalization using tree.

2.7 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.

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

2.8 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

2.9 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 $\mathcal{A}$.

2.10 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.

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

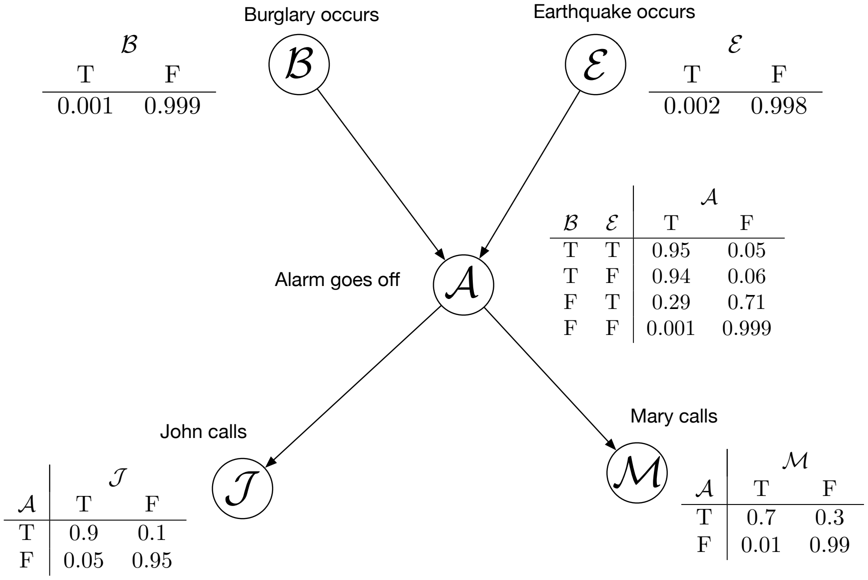
The intuition is that we can drop $\mathcal{B}$ from the right side of the conditioning bar since knowing that $\mathcal{B}$ occurred doesn't change the probability of $\mathcal{A}$ if we already know that $\mathcal{C}$ occurred.

3 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 the random variables and events. This graph contains nodes and directed edges. We can think of an edge that goes from node A to node B as indicating 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. Note: that on the next assignment when we generalize these networks to contain continuous values, these probability tables will be modified.

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

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

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

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

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, 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).

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

Exercise 1

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

3.1 D-Separation

One of the beautiful things at 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 $\mathcal{C}$ 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 2

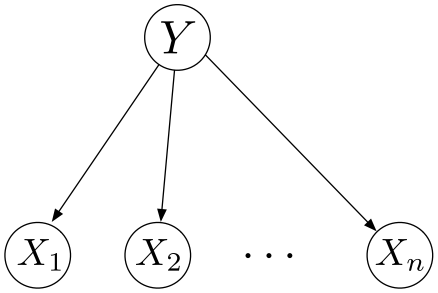
Choose two nodes in the alarm network. Are they independent? How would this change if you conditioned on each of the other three nodes in the network? Repeat this exercise for other pairs of nodes in the network.

4 Fairness Criteria for ML

TODO

5 Naïve Bayes

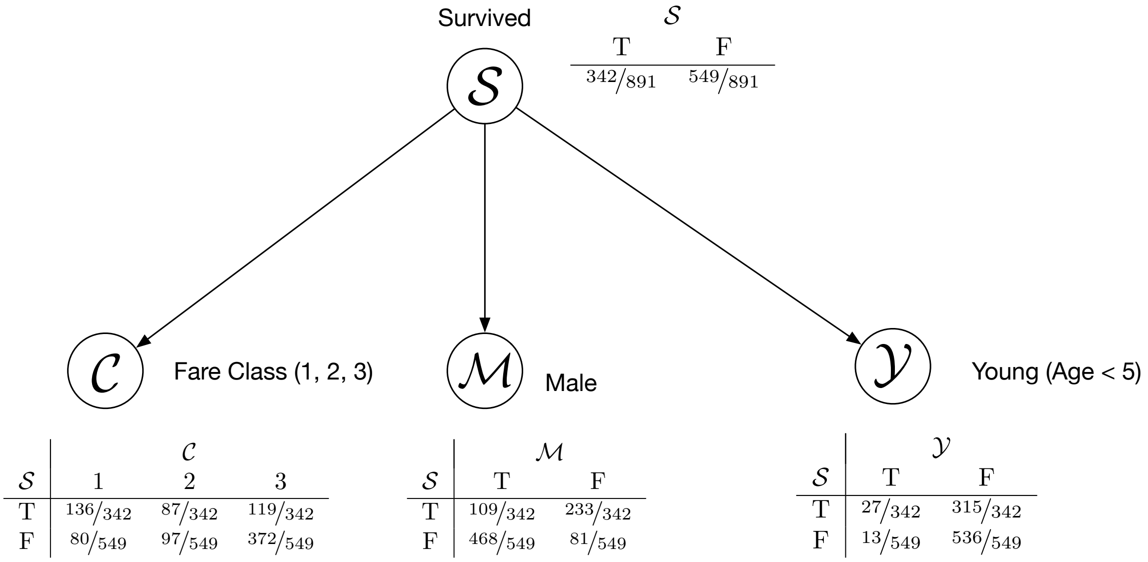
The Naïve Bayes model can be described using the following BN.



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, you need $p(Y = i)$ (for all possible values, i, that Y could take on). If $\mathcal{Y}$ is an event, you would need $p(\mathcal{Y})$.
- If the X_i's are random variables, you 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).

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
$^{342}/_{891}$	$^{549}/_{891}$

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

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

	$\mathcal{Y}$	
$\mathcal{S}$	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 variables (*fare class*) and some 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).

6 *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 an even simpler problem. Next, we're going to apply the maximum likelihood technique to a specific example of data from the Titanic network. We hope that by these two strategies provide new ways for you to understand MLE.

6.1 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's to reference the specific values that we observed for the n flips. We'd like to determine the maximum likelihood estimate of Θ. We can write this goal formally using the following equation.

$$\Theta^\star = \arg\max_\Theta p(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n | \Theta) \tag{20}$$

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 $i \neq j$. Applying that assumption gives us

$$\Theta^\star = \arg\max_\Theta p(X_1 = x_1|\Theta)p(X_2 = x_2|\Theta)\ldots p(X_n = x_n|\Theta) \ \ . \tag{21}$$

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

$$\begin{aligned}\Theta^\star &= \arg\max_\Theta \log\left(p(X_1 = x_1|\Theta)p(X_2 = x_2|\Theta)\ldots p(X_n = x_n|\Theta)\right) \\ &= \arg\max_\Theta \sum_{i=1}^{n} \log p(X_i = x_i|\Theta)\end{aligned} \tag{22}$$

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.

$$\Theta^\star = \arg\max_\Theta \sum_{i=1}^{n} \left(\mathbb{I}[x_i = 1]\log\Theta + \mathbb{I}[x_i = 0]\log(1 - \Theta)\right) \tag{23}$$

Examining Equation 23, we can see that the value inside the summation only depends on whether x_i is 1 or 0. Therefore, if n_H $(x_i = 1)$ is the number of heads we observed and n_T is the number of tails we observed $(x_i = 0)$, we can rewrite Equation 23 as

$$\Theta^\star = \arg\max_\Theta n_H \log\Theta + n_T \log(1 - \Theta) \ \ . \tag{24}$$

We can solve this equation by taking the derivative with respect to Θ and setting it to 0 (on 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).

$$\begin{aligned}
\frac{d}{d\Theta}\left(n_H \log \Theta + n_T \log\left(1-\Theta\right)\right) &= \frac{n_H}{\Theta} - \frac{n_T}{1-\Theta} \\
0 &= \frac{n_H}{\Theta^\star} - \frac{n_T}{1-\Theta^\star} && \text{$\Theta^\star$ occurs at a critical point} \\
&= \frac{n_H(1-\Theta^\star)}{\Theta^\star(1-\Theta^\star)} - \frac{n_T\Theta^\star}{\Theta^\star(1-\Theta^\star)} && \text{put over a common denominator} \\
0 &= n_H(1-\Theta^\star) - n_T\Theta^\star && \text{assuming $\Theta^\star$ is not 0 or 1, we can multiply by $\Theta^\star(1-\Theta^\star)$} \\
0 &= n_H - n_H\Theta^\star - n_T\Theta^\star \\
(n_H + n_T)\Theta^\star &= n_H \\
\Theta^\star &= \frac{n_H}{n_H + n_T}
\end{aligned} \tag{25}$$

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 likelihood of data given parameters.
2. Simplify likelihood and substitute in 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 solving for a critical point).

6.2 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 fo Naïve Bayes. In that spirit, we will use training data to fir the parameters of our Titanic BN.

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

- In order to fit the parameters for the outcome variables Y, we can use this equation.

$$\Theta^\star(i) = \frac{\text{ycount}(i)}{\sum_{j=1}^{c} \text{ycount}(j)} \tag{26}$$

- In order to fit the conditional probabilities of the features (in this case young, male, and fare class), we can use the following equation.

$$\Theta_j^\star(k|i) = \frac{\text{xcount}_j(k|i)}{\sum_{u=1}^{r} \text{xcount}_j(u|i)} \tag{27}$$

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 ycount$(\mathcal{S})$ and 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 $p(\mathcal{M}|\mathcal{S})$, we need to know four counts $\text{xcount}(\mathcal{M}|\mathcal{S}), \text{xcount}(\neg\mathcal{M}|\mathcal{S}), \text{xcount}(\mathcal{M}|\neg\mathcal{S}), \text{and}, \text{xcount}(\neg\mathcal{M}|\neg\mathcal{S})$ (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 3

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

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

7 *N-Grams*

This was primarily coding based, so for now we don't have anything here. Let us know if there are parts you want explained and we'll update this document.

8 *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.