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 as appendices to this document (these will be referenced in the text).

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, $\mathcal{E}_1, \mathcal{E}_2, \ldots \mathcal{E}_n$ (i.e., events 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, $\mathcal{E}_1, \mathcal{E}_2, \ldots \mathcal{E}_n$, (i.e., events where at least one *must* occur),

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

2.2 Random Variables

A random variable is a generalization of an event. Think of a random variable as an entity that takes on some 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 above 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 specific 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 $\mathcal{B}$, $p(\mathcal{A}|\mathcal{B}) = \frac{p(\mathcal{B}|\mathcal{A})p(\mathcal{A})}{p(\mathcal{B})}$, 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 that 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.

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

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.

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

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

2.6 Marginalization

Suppose we can easily compute the joint distribution of two events (that is $p(\mathcal{A}, \mathcal{B})$), 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 it doesn't change the probability of $\mathcal{A}$.

2.10 Conditional Independence

Two events $\mathcal{A}$ and $\mathcal{B}$ are 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 it doesn't change the probability of $\mathcal{A}$ if we already know $\mathcal{C}$.

3 Bayesian Networks

3.1 D-Separation

4 Fairness Criteria for ML

5 Naïve Bayes

6 N-Grams and Sequence Analysis

7 Maximum Likelihood Estimation

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.