Lecture 11: PROBABILITY [*]

10-606 MATHEMATICAL FOUNDATIONS FOR MACHINE LEARNING

1 Probability

Many events can't be predicted with certainty. We use the tools of probability to quantify how likely they are to happen. To reason about probabilities, we'll use a data type called a *probability space*.

1.1 Atomic events

We start from a *universe* or *sample space*, a set U of mutually exclusive and exhaustive possible outcomes. The elements of U are called *simple* or *atomic* events. Often these simple events will be interpretable, e.g., which face of a die lands on top when we roll it or which card we draw from a deck. We assign probabilities to atomic events: $P(a)$ is the probability that $a \in U$ will occur.

For our purposes, it's best to think of U as finite. If we want to deal with continuous outcomes, a good way to think about them is to discretize: pretend that there are a very large but finite number of events, spaced out so that there is always one very close to any possible continuous outcome. Since this viewpoint is for understanding rather than computation, there's no limit on how big we can make U, e.g., we could have one event for every 64-bit floating point number, or even every tuple of 100 64-bit numbers. Discretization is actually better in many cases than trying to deal directly with continuous outcomes: probability on continuous spaces holds a lot of mathematical traps, such as hidden infinities and counterintuitive behaviors.

We'll take the viewpoint that probabilities are *subjective*: we can assign any probabilities we want to atomic events, so long as we follow the rules described below for manipulating them. And we can assign probabilities to any outcomes that we can imagine measuring, no matter whether we think of the process that determines the outcomes as being random.

What does the statement $P(a) = \frac{1}{3}$ mean? One common interpretation is that, if we reset the world and ran it forward many times, measuring whether event a happened in each trial, we would see it happen in a third of them. This interpretation makes sense for something like a biased coin that we can flip many times. But it is not necessarily helpful in general: first, it may not really be possible to reset the world, even approximately. Second, we might want to assign a probability to an event that is deterministic but unknown to us, so that repeated trials don't yield independent measurements. Instead, in such cases we interpret the statement as a measure of our subjective confidence in a: we would be willing to take $2 : 1$ odds in a bet on a, which is the same odds that would have us break even if betting on a biased coin with $P(H) = \frac{1}{3}$.

Probabilities are always nonnegative and an impossible event always has probability 0. Probabilities always sum to 1 over the universe:

$$\sum_{e \in U} P(e) = 1$$

That means that each individual event has probability at most 1. An event with probability 1 is certain to happen, since the sum-to-1 rule means that all other events must have probability zero.

[*] Compiled on Tuesday 20th June, 2023 at 18:35

1.2 Compound events

A compound event is a subset of the universe, $E \subseteq U$. The probability of E is the sum of the probabilities of atomic events in E:

$$P(E) = \sum_{a \in E} P(a)$$

The sum-to-1 rule means that we are certain to get some event in the universe:

$$P(U) = \sum_{a \in U} P(a) = 1$$

We'll sometimes abuse notation and conflate a with $\{a\}$: that is, we'll interchange an atomic event and a compound event with only one element.

Since events are sets, we can combine them using set operations: given events A and B,

- $A \cup B$ means that we get some atomic event in the union of A and B; that is, $A \cup B$ is the event that *either* A or B happens.
- $A \cap B$ is the event that *both* A and B happen.
- $A \setminus B$ is the event that A happens but B does not happen.
- $A^C = U \setminus A$ is the event that A does not happen.

As you can see from the examples above, set operations correspond to logical combinations of events. So we'll sometimes use logical notation as well:

- $A \vee B$ is the same as $A \cup B$.
- $A \wedge B$ is the same as $A \cap B$.
- $\neg A$ or $\bar{A}$ is the same as A^C.

We can make the correspondence with logical notation precise:

- Logical predicates correspond to events. For example, in our die-roll example, the predicate "even" corresponds to the set of even-valued outcomes.
- Logical functions can be used to build predicates. For example, if we have a function "suit" that extracts the suit of a card, then the expression "suit $= \spadesuit$" corresponds to the set $\{a \mid \text{suit}(a) = \spadesuit\}$, or

  $$\{A\spadesuit, 2\spadesuit, 3\spadesuit, 4\spadesuit, 5\spadesuit, 6\spadesuit, 7\spadesuit, 8\spadesuit, 9\spadesuit, 10\spadesuit, J\spadesuit, Q\spadesuit, K\spadesuit\}$$

  Note that we have omitted the argument a to the function "suit": by convention, the atomic event is implicitly an argument of every function or predicate.
- General expressions correspond to the set of atomic events that satisfy them. For example, the expression "suit $= \spadesuit \wedge$ even" corresponds to the set

  $$\{2\spadesuit, 4\spadesuit, 6\spadesuit, 8\spadesuit, 10\spadesuit, Q\spadesuit\}$$

  which is $\{a \mid \text{suit}(a) = \spadesuit \wedge \text{even}(a)\}$.

Knowledge check

1. **Math**: What is the logical expression that corresponds to the set $\{A\spadesuit, A\heartsuit, A\clubsuit, A\diamondsuit, K\diamondsuit, Q\diamondsuit, J\diamondsuit\}$?
   - **Hint**: Use a function "value" that returns the number associated with a card; assume that value$(J) = 11$, value$(Q) = 12$, value$(K) = 13$, and value$(A) = 1$,
   - **Answer**: "value $= 1 \vee$ (suit $= \wedge$value > 10)"
2. **Math**: What is the set notation for the logical event $\neg(A \rightarrow B)$?
   - **Answer**: $A \setminus B$. Set difference is $(A$ and not $B)$ which we can express as

$$(A \text{ and not } B) == \neg\neg(A \text{ and not } B) == \neg(\text{not } A \text{ or } B) == \neg(A \rightarrow B)$$

1.3 Probability density

If we want to write a probability distribution over a continuous set such as $\mathbb{R}$, we will have infinitely many atomic events: one for each real number. We can't assign positive probability to this many events, since the total probability would be ∞ instead of 1. So instead we define a new function $p \in \mathbb{R} \to \mathbb{R}$, and say that the probability of any event $A \subseteq \mathbb{R}$ is the integral

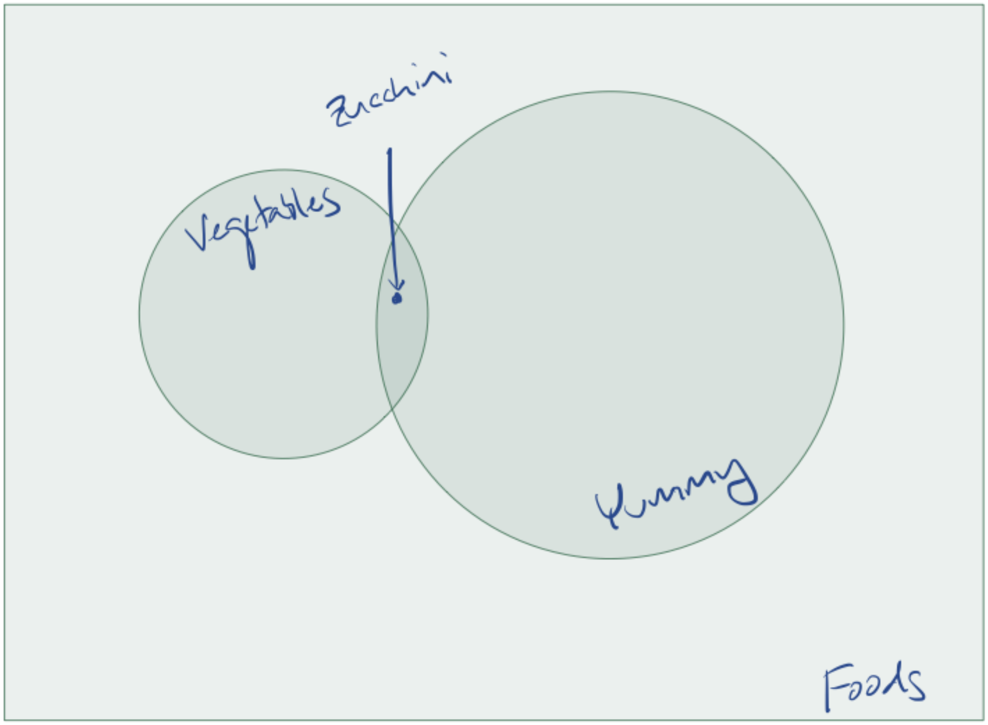
$$P(A) = \int_A p(x)dx$$

The function p is called a *probability density*, and it obeys many (but not all) of the same rules that a probability does. For example, it must be nonnegative, and the integral over the entire universe must be 1.

It is important to remember that a probability density is **not** a probability. For example, a density can be greater than 1, so long as there's no set A such that $\int_A p(x)dx > 1$. And, the probability of an atomic event x is **not** equal to $p(x)$; as noted above, it's typically zero, so that the probability of x is usually not what we want to ask for. Instead, we might ask for the probability that our random variable will fall in a small interval of length ϵ centered at x; this is approximately $\epsilon\, p(x)$.

1.4 Venn diagrams

We can visualize our universe and the probabilities of various events using a *Venn diagram*: we picture the universe as a large rectangle, and events as shapes that are subsets of the rectangle. Every point in the rectangle corresponds to a different atomic event. The overlap between shapes tells us which compound events intersect, that is, which can occur simultaneously.



For example, if we are picking a random food to eat for dinner, the universe is the set of all possible foods we might pick, corresponding to the whole rectangle. The event that we pick something yummy is the larger circle, and the event that we pick a vegetable is the smaller circle. The overlap between the two circles is the set of yummy vegetables, which includes the atomic event “zucchini” (at least, according to your professor). The area outside the two circles corresponds to the set of outcomes which are neither yummy nor vegetables. Often we try to make the areas of shapes within the Venn diagram proportional to their probabilities.

1.5 Uniform distributions

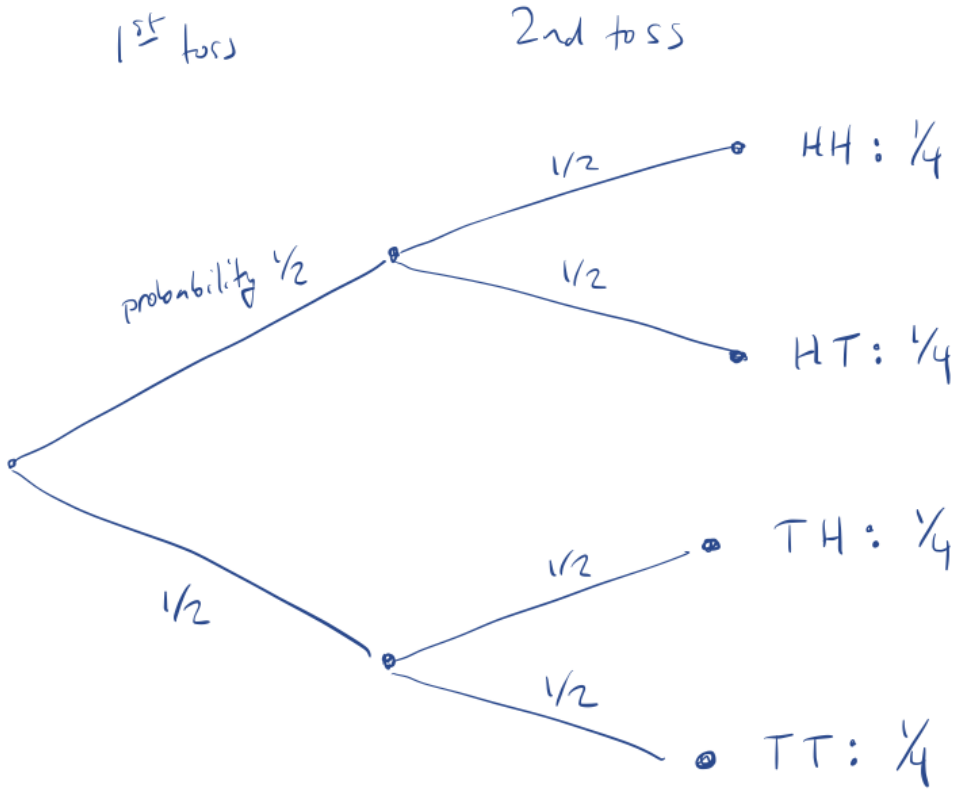
If all atomic events are equally likely, our probability space represents a *uniform distribution*. This distribution is often a default choice but we will also occasionally need *non-uniform distributions*, those in which different atomic events have different probabilities.

For example, a fair die is modeled well by a uniform distribution: $P(1) = P(2) = P(3) = P(4) = P(5) = P(6) = \frac{1}{6}$. But a weighted die needs a non-uniform distribution. If rolling a 6 is four times as likely as any other outcome, we have $P(1) = P(2) = P(3) = P(4) = P(5) = \frac{1}{9}$ and $P(6) = \frac{4}{9}$.

2 Experiments

We can use probabilities to describe *experiments* that we are thinking of conducting. We make a probability space that describes what we think might happen; atomic events represent possible outcomes of the experiment, and the universe contains all possible outcomes.

For example, suppose we flip a coin twice. We can make a probability space that says that each sequence of heads and tails is equally likely:



Here the universe is $\{HH, HT, TH, TT\}$, and we’ve assigned probability $\frac{1}{4}$ to each of these atomic events.

In a realistic experiment, we might not observe the distinctions between all atomic events. For example, above we might just observe the total number of heads instead of the exact sequence of coin flips. We can describe what we observe using compound events: in this case, $\{HH\}$, $\{HT, TH\}$, and $\{TT\}$.

2.1 Samples

A typical experiment is *repeatable*: we can imagine doing the same experiment more than once, with the distribution over outcomes $P(X)$ being the same each time. If we repeat the experiment T times, we can collect the outcomes into a *sample* or *data set*: $X_1, X_2, \ldots, X_T$.

By assumption, order doesn't matter: the sample is *exchangeable*. That is, we can permute the indices $1 : T$ without changing the information contained in the sample.

Despite the name, a data set is not a set: like a set, order is unimportant, but unlike a set, it matters if we see the same outcome more than once.

For example, if we flip a single coin seven times, we might see the sample H, H, T, H, T, T, T (that is, $X_1 = H$, $X_2 = H$, $X_3 = T$, and so forth). This outcome is equivalent to T, H, T, H, T, H, T or T, T, T, T, H, H, H: all that matters is that we saw three heads and four tails.

For another example, we might flip a pair of coins three times, recording the number of heads each time. Then we might see the samples $0, 0, 2$ or $1, 1, 1$. What matters here is the number of times we saw 0, the number of times we saw 1, and the number of times we saw 2.

3 Working with probabilities

In small probability spaces we can just list out all the atomic events, and calculate probabilities of compound events by the sum rule. But it's common to have so many atomic events that we have no hope of listing them all: for example, there are more possible shuffles of a 52-card deck than there are atoms in the Earth. So, we need tools to help us do the calculations without explicitly enumerating everything. One such tool is the manipulation of unions and our ability to complex compound events as the union of simpler events.

3.1 Disjoint union

If A and B are disjoint events, meaning that $A \cap B = \emptyset$, then

$$P(A \cup B) = P(A) + P(B)$$

This follows directly from our definitions:

$$P(A \cup B) = \sum_{e \in A \cup B} P(e) = \sum_{e \in A} P(e) + \sum_{e \in B} P(e) = P(A) + P(B)$$

Disjoint events are also called *mutually exclusive*.

The disjoint union rule also works for multiple sets: if the sets A_i are mutually exclusive, meaning that they are pairwise disjoint, then

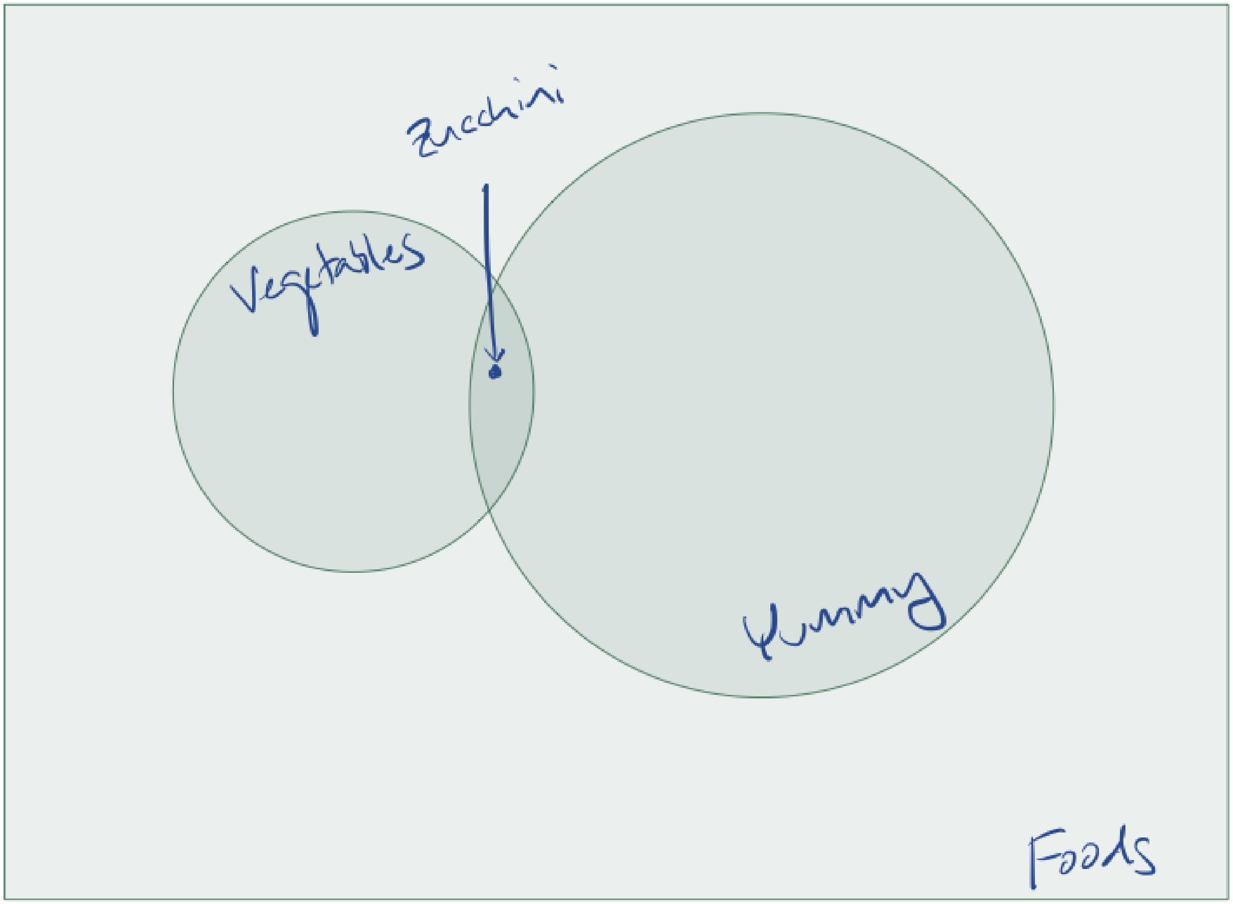
$$P\left(\bigcup_i A_i\right) = \sum_i P(A_i)$$

A useful generalization of the disjoint union rule is the *union bound*: for any A and B, disjoint or not, we have

$$P(A \cup B) \leq P(A) + P(B)$$

3.2 Non-disjoint union

We can use the tools above to figure out what happens when we take a union of sets that overlap, like in our food example from before.



Let A be the event that we pick a vegetable, and B be the event that we pick a yummy food. How can we calculate $P(A \cup B)$?

We can split the set $A \cup B$ into three disjoint parts: $A \setminus B$, $B \setminus A$, and $A \cap B$. So,

$$P(A \cup B) = P(A \setminus B) + P(B \setminus A) + P(A \cap B)$$

We can also split A and B into two pieces each:

$$P(A) = P(A \setminus B) + P(A \cap B)$$

$$P(B) = P(B \setminus A) + P(A \cap B)$$

Adding these together,

$$P(A) + P(B) = P(A \setminus B) + P(B \setminus A) + 2P(A \cap B)$$

Since $P(A \cap B) \geq 0$, we see that $P(A) + P(B) \geq P(A \cup B)$; this is the union bound we described above. By subtracting the above expressions for $P(A)+P(B)$ and $P(A \cup B)$ we can figure out the exact difference:

$$P(A) + P(B) - P(A \cup B) = P(A \cap B)$$

Knowledge check

1. **Numerical answer**: Suppose you survey a large group of people and ask them two questions: "do you like computer science?" and "do you like math?" You find that 70% of people respond yes to the first question, and 90% respond yes to the second. Furthermore, 65% respond yes to both questions. What is the probability of finding a person who responds no to both questions?
   - **Answer**: Define the event A to be someone responds yes to the first question and the event B to be someone responds yes to the second question. Using the information provided in the question, we have $P(A) = 0.7$, $P(B) = 0.9$, and $P(A \cap B) = 0.65$. Rearranging the equation above, we get

     $$P(A)+P(B)-P(A\cup B) = P(A\cap B) \rightarrow P(A)+P(B)-P(A\cap B) = P(A\cup B)$$

     Plugging in the values we know, we can conclude that $P(A \cup B) = 0.95$. This means that 95% of people responded yes to at least one question. Therefore, only 5% of people responded no to both questions, so the probability of finding a person who responds no to both questions is 0.05.