Why probability?

Prostitutes Appeal to Pope

Language is ambiguous

Probability theory gives us a tool to model this ambiguity in reasonable ways.

Basic Probability Theory: terminology

An experiment has a set of potential outcomes, e.g., throw a dice, “look at” another sentence

The sample space of an experiment is the set of all possible outcomes, e.g., \{1, 2, 3, 4, 5, 6\}

In NLP our sample spaces tend to be very large

- All words, bigrams, 5-grams
- All sentences of length 20 (given a finite vocabulary)
- All sentences
- All parse trees over a given sentence
Basic Probability Theory: terminology

An event is a subset of the sample space.

Dice rolls:
- (2)
- (3, 6)
- even = (2, 4, 6)
- odd = (1, 3, 5)

NLP:
- a particular word/part of speech occurring in a sentence
- a particular topic discussed (politics, sports)
- sentence with a parasitic gap
- pick your favorite phenomena...

Events

We’re interested in probabilities of events:
- p(2)
- p(even)
- p(odd)
- p(parasitic gap)
- p(first word in a sentence is “banana”)

Random variables

A random variable is a mapping from the sample space to a number (think events).

It represents all the possible values of something we want to measure in an experiment.

For example, random variable, X, could be the number of heads for a coin tossed three times.

<table>
<thead>
<tr>
<th>space</th>
<th>HHH</th>
<th>HHT</th>
<th>HTH</th>
<th>HTT</th>
<th>THH</th>
<th>THT</th>
<th>TTH</th>
<th>TTT</th>
</tr>
</thead>
<tbody>
<tr>
<td>X</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

Really for notational convenience, since the event space can sometimes be irregular.

Random variables

We can then talk about the probability of the different values of a random variable.

The definition of probabilities over all of the possible values of a random variable defines a probability distribution.

<table>
<thead>
<tr>
<th>space</th>
<th>HHH</th>
<th>HHT</th>
<th>HTH</th>
<th>HTT</th>
<th>THH</th>
<th>THT</th>
<th>TTH</th>
<th>TTT</th>
</tr>
</thead>
<tbody>
<tr>
<td>X</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>X</th>
<th>P(X)</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>?</td>
</tr>
<tr>
<td>2</td>
<td>?</td>
</tr>
<tr>
<td>1</td>
<td>?</td>
</tr>
<tr>
<td>0</td>
<td>?</td>
</tr>
</tbody>
</table>
Random variables

We can then talk about the probability of the different values of a random variable.

The definition of probabilities over all of the possible values of a random variable defines a probability distribution.

<table>
<thead>
<tr>
<th>Outcome</th>
<th>P(X)</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>1/8</td>
</tr>
<tr>
<td>2</td>
<td>3/8</td>
</tr>
<tr>
<td>1</td>
<td>3/8</td>
</tr>
<tr>
<td>0</td>
<td>1/8</td>
</tr>
</tbody>
</table>

Probability distribution

To be explicit:
- A probability distribution assigns probability values to all possible values of a random variable.
- These values must be $\geq 0$ and $\leq 1$.
- These values must sum to 1 for all possible values of the random variable.

<table>
<thead>
<tr>
<th>Outcome</th>
<th>P(X)</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>1/2</td>
</tr>
<tr>
<td>2</td>
<td>1/2</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Unconditional/prior probability

Simplest form of probability distribution is:
- $P(X)$

Prior probability: without any additional information:
- What is the probability of heads on a coin toss?
- What is the probability of a sentence containing a pronoun?
- What is the probability of a sentence containing the word “banana”?
- What is the probability of a document discussing politics?
- …

Prior probability

What is the probability of getting HHH for three coin tosses, assuming a fair coin?

$1/8$

What is the probability of getting THT for three coin tosses, assuming a fair coin?

$1/8$
Joint distribution

We can also talk about probability distributions over multiple variables

\[ P(X,Y) \]

- probability of \( X \) and \( Y \)
- a distribution over the cross product of possible values

<table>
<thead>
<tr>
<th>( \text{NLPPass} )</th>
<th>( P(\text{NLPPass}) )</th>
<th>( P(\text{NLPPass, EngPass}) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>true</td>
<td>0.89</td>
<td>true, true 0.88</td>
</tr>
<tr>
<td>false</td>
<td>0.11</td>
<td>true, false 0.01</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>( \text{EngPass} )</th>
<th>( P(\text{EngPass}) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>true</td>
<td>0.92</td>
</tr>
<tr>
<td>false</td>
<td>0.08</td>
</tr>
</tbody>
</table>

What is \( P(\text{ENGPass}) \)?

13

Joint distribution

Still a probability distribution
- all values between 0 and 1, inclusive
- all values sum to 1

All questions/probabilities of the two variables can be calculated from the joint distribution

<table>
<thead>
<tr>
<th>( \text{NLPPass, EngPass} )</th>
<th>( P(\text{NLPPass, EngPass}) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>true, true</td>
<td>0.88</td>
</tr>
<tr>
<td>true, false</td>
<td>0.01</td>
</tr>
<tr>
<td>false, true</td>
<td>0.04</td>
</tr>
<tr>
<td>false, false</td>
<td>0.07</td>
</tr>
</tbody>
</table>

How did you figure that out?

14

Joint distribution

Still a probability distribution
- all values between 0 and 1, inclusive
- all values sum to 1

All questions/probabilities of the two variables can be calculated from the joint distribution

\[ P(x) = \sum_{y \in Y} p(x,y) \]

Called “marginalization”, aka summing over a variable

<table>
<thead>
<tr>
<th>( \text{NLPPass, EngPass} )</th>
<th>( P(\text{NLPPass, EngPass}) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>true, true</td>
<td>0.88</td>
</tr>
<tr>
<td>true, false</td>
<td>0.01</td>
</tr>
<tr>
<td>false, true</td>
<td>0.04</td>
</tr>
<tr>
<td>false, false</td>
<td>0.07</td>
</tr>
</tbody>
</table>

15

Joint distribution

Still a probability distribution
- all values between 0 and 1, inclusive
- all values sum to 1

All questions/probabilities of the two variables can be calculated from the joint distribution

16
Conditional probability

As we learn more information, we can update our probability distribution.

$P(X \mid Y)$ models this (read “probability of $X$ given $Y$”)
- What is the probability of heads given that both sides of the coin are heads?
- What is the probability the document is about politics, given that it contains the word “Clinton”?
- What is the probability of the word “banana” given that the sentence also contains the word “split”?

Notice that it is still a distribution over the values of $X$.

17

Conditional probability

$p(X \mid Y) = \frac{P(X,Y)}{P(Y)}$

In terms of prior and joint distributions, what is the conditional probability distribution?

18

Conditional probability

$p(X \mid Y) = \frac{P(X,Y)}{P(Y)}$

Given that $Y$ has happened, what proportion of those events does $X$ also happen.

19

Conditional probability

$p(X \mid Y) = \frac{P(X,Y)}{P(Y)}$

Given that $Y$ has happened, what proportion of those events does $X$ also happen.

19

<table>
<thead>
<tr>
<th>NLPPass, EngPass</th>
<th>P(NLPPass, EngPass)</th>
</tr>
</thead>
<tbody>
<tr>
<td>true, true</td>
<td>.88</td>
</tr>
<tr>
<td>true, false</td>
<td>.01</td>
</tr>
<tr>
<td>false, true</td>
<td>.04</td>
</tr>
<tr>
<td>false, false</td>
<td>.07</td>
</tr>
</tbody>
</table>

What is $p(\text{NLPPass}=\text{true} \mid \text{EngPass}=\text{false})$?

20
Conditional probability

\[ p(X \mid Y) = \frac{P(X,Y)}{P(Y)} \]

<table>
<thead>
<tr>
<th>NLPass, EngPass</th>
<th>P(NLPass, EngPass)</th>
</tr>
</thead>
<tbody>
<tr>
<td>true, true</td>
<td>.88</td>
</tr>
<tr>
<td>true, false</td>
<td>.01</td>
</tr>
<tr>
<td>false, true</td>
<td>.04</td>
</tr>
<tr>
<td>false, false</td>
<td>.07</td>
</tr>
</tbody>
</table>

What is: \( p(\text{NLPass} = \text{true} \mid \text{EngPass} = \text{false}) \)?

\[ P(\text{true, false}) = 0.01 \]
\[ P(\text{EngPass} = \text{false}) = 0.01 + 0.07 = 0.08 \]
\[ 0.125 = 0.125 \]

Notice this is very different than \( p(\text{NLPass} = \text{true}) = 0.89 \).

A note about notation

When talking about a particular assignment, you should technically write \( p(X = x) \), etc.

However, when it’s clear, we’ll often shorten it.

Also, we may also say \( P(X) \) or \( p(x) \) to generically mean any particular value, i.e. \( P(X = x) \).

\[ P(\text{true, false}) = 0.01 \]
\[ P(\text{EngPass} = \text{false}) = 0.01 + 0.07 = 0.08 \]
\[ 0.125 = 0.125 \]

Properties of probabilities

\[ P(A \text{ or } B) = ? \]

\[ P(A \text{ or } B) = P(A) + P(B) - P(A,B) \]

Notice this is very different than \( P(\text{NLPass} = \text{true}) = 0.89 \).
Properties of probabilities

\[ P(\neg E) = 1 - P(E) \]

More generally:

\[ P(E_1, E_2) \leq P(E_1) \]

\[ \sum_{j \neq i} p(e_j) = 1 - p(e_i) \]

Chain rule (aka product rule)

\[ p(X | Y) = \frac{p(X, Y)}{p(Y)} \]

We can view calculating the probability of \( X \) AND \( Y \) occurring as two steps:

1. \( Y \) occurs with some probability \( p(Y) \)
2. Then, \( X \) occurs, given that \( Y \) has occurred

or you can just trust the math... 😊

Chain rule

\[
p(X,Y,Z) = P(X,Y|Z)P(Y,Z) \\
p(X,Y,Z) = P(X|Y,Z)P(Y,Z)P(Z) \\
p(X,Y,Z) = P(Y,Z|X)P(X) \\
\]

\[ p(X_1, X_2, ..., X_n) = ? \]

Applications of the chain rule

We saw that we could calculate the individual prior probabilities using the joint distribution

\[ p(x) = \sum_{y} p(x,y) \]

What if we don't have the joint distribution, but do have conditional probability information:

\[ p(Y) \]
\[ p(X | Y) \]

\[ p(x) = \sum_{y} p(y)p(x|y) \]
Bayes’ rule (theorem)

\[ p(X | Y) = \frac{P(X, Y)}{P(Y)} \quad \Rightarrow \quad p(X, Y) = P(X | Y)P(Y) \]

\[ p(Y | X) = \frac{P(X, Y)}{P(X)} \quad \Rightarrow \quad p(X, Y) = P(Y | X)P(X) \]

\[ p(X | Y) = \frac{P(Y | X)P(X)}{P(Y)} \]

Obtaining probabilities

We’ve talked a lot about probabilities, but not where they come from.

How do we calculate:
- the probability of heads?
- the probability that a sentence contains a pronoun?
- the probability that a sentence contains “banana”, given that it also contains the word split?

Estimating probabilities

What is the probability of a sentence contains a pronoun?

We don’t know!

We can estimate it based on data, though:

\[ p(y) = \frac{\text{count}(y)}{n} \quad \text{number of examples with thing } y \quad \text{total number of examples} \]

\[ p(x | y) = \frac{\text{count}(x, y)}{\text{count}(y)} \quad \text{number of examples with thing } y \text{ and thing } x \quad \text{number of examples with thing } y \]

Maximum likelihood estimates

This is called the maximum likelihood estimation. Why?
Bayes rule

 Allows us to talk about $P(Y|X)$ rather than $P(X|Y)$

 Sometimes this can be more intuitive

 Why?

 $p(X|Y) = \frac{P(Y|X)P(X)}{P(Y)}$

 Bayes rule

 \[ p(\text{disease} | \text{symptoms}) \propto \frac{\text{count}(\text{disease, symptoms})}{\text{count}(\text{symptoms})} \]

 How would you estimate this?

 Find a bunch of people with those symptoms and see how many have the disease

 Is this feasible?

 Bayes rule

 \[ p(\text{linguistic phenomena} | \text{features}) \]

 How would you estimate this?

 For all examples that had those features, how many had that phenomena?

 For all the examples with that phenomena, how many had this feature

 $p(\text{cause} | \text{effect})$ vs. $p(\text{effect} | \text{cause})$
Gaps

I just won't put these away.
These, I just won't put away.
I just won't put away.

37

Gaps

What did you put away?
The socks that I put away.

38

Gaps

Whose socks did you fold and put away?
Whose socks did you fold?
Whose socks did you put away?

39

Parasitic gaps

These I'll put away without folding.
These I'll put away.
These without folding.

40
Parasitic gaps

These I'll put ___ away without folding ___.

gap
gap

1. Cannot exist by themselves (parasitic)

These I'll put my pants away without folding ___.

gap

2. They're optional

These I'll put ___ away without folding them.

gap

Frequency of parasitic gaps

Parasitic gaps occur on average in 1/100,000 sentences

Problem:
You have developed a complicated set of regular expressions to try and identify parasitic gaps. If a sentence has a parasitic gap, it correctly identifies it 95% of the time. If it doesn’t, it will incorrectly say it does with probability 0.005. Suppose we run it on a sentence and the algorithm says it has a parasitic gap; what is the probability it actually does?

Prob of parasitic gaps

You have developed a complicated set of regular expressions to try and identify parasitic gaps. If a sentence has a parasitic gap, it correctly identifies it 95% of the time. If it doesn’t, it will incorrectly say it does with probability 0.005. Suppose we run it on a sentence and the algorithm says it has a parasitic gap, what is the probability it actually does?

\[ G \equiv \text{gap} \]
\[ T \equiv \text{test positive} \]

What question do we want to ask?
You have developed a complicated set of regular expressions to try and identify parasitic gaps. If a sentence has a parasitic gap, it correctly identifies it 95% of the time. If it doesn’t, it will incorrectly say it does with probability 0.005. Suppose we run it on a sentence and the algorithm says it has a parasitic gap, what is the probability it actually does?

\[ p(g | t) = \ ? \]

\[ p(t | g) = 0.95 \]
\[ p(\overline{g}) = 0.005 \]
\[ p(t | \overline{g}) = 0.001 \]

\[ p(g | t) = \frac{p(t | g)p(g)}{p(t)} \]
\[ = \frac{p(t | g)p(g)}{p(t | g)p(g) + p(t | \overline{g})p(\overline{g})} \]
\[ = \frac{0.95 \times 0.0001}{0.0001 \times 0.95 + 0.9999 \times 0.005} \]
\[ = 0.002 \]