Two Sisters Reunited After 18 Years in Checkout Counter  
Deer Kill 130,000  
Prostitutes Appeal to Pope  
Man Struck by Lightning Faces Battery Charge  
Milk Drinkers Are Turning to Powder  
Doctor Who Aided Bin Laden Raid in Jail  
Enraged Cow Injures Farmer with Axe  
Eye Drops Off Shelf  
Drunk Gets Nine Months in Violin Case  
Stolen Paining Found by Tree  
Include Your Children When Baking Cookies  
August Sales Fall for GM as Trucks Lift Chrysler  
Big Rig Carrying Fruit Crashes on 210 Freeway, Creates Jam  
Squad Helps Dog Bite Victims  
Two Foot Ferries May Be Headed to Trinidad  
Astronaut Takes Blame for Gas in Spacecraft  
Reagan Wins on Budget, but More Lies Ahead  
Dealers Will Hear Car Talk at Noon

Assignment 0  
What do you think the average, median, min and max were for the “programming proficiency” question (1-10)?

Assignment advice  
- test individual components of your regex first, then put them all together  
- write test cases

Assignment deadlines

Class participation
Assignment 0

Mean/median: 6.5

Assignment 0: programming languages

Python 9
Java 8
Haskell 1
Ruby 1
Matlab 1
SML 1
C++ 1

Regex revisited

Corpus statistics

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

- \(p(\{2\})\)
- \(p(\text{even})\)
- \(p(\text{odd})\)
- \(p(\text{parasitic gap})\)
- \(p(\text{first word in a sentence is “banana”})\)
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.

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

<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>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

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>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

Random variables

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.
**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?
- …

**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>NLPPass</th>
<th>P(NLPPass)</th>
</tr>
</thead>
<tbody>
<tr>
<td>true</td>
<td>0.89</td>
</tr>
<tr>
<td>false</td>
<td>0.11</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>EngPass</th>
<th>P(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>

<table>
<thead>
<tr>
<th>NLPPass, EngPass</th>
<th>P(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>

**Prior probability**

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

\[ \frac{1}{8} \]

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

\[ \frac{1}{8} \]

**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>NLPPass, EngPass</th>
<th>P(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>

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

P(NLPPass, EngPass) = 0.92

How did you figure that out?

Conditional probability

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

P(X | Y) models this (read "probability of X given Y")
- What is the probability of a 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

Joint distribution

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

Called "marginalization", aka summing over a variable

<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>

Conditional probability

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

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

Given that \( y \) has happened, what proportion of those events does \( x \) also happen

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

Given that \( y \) has happened, what proportion of those events does \( x \) also happen

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

<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 \( p(\text{NLPPass}=\text{true} \mid \text{EngPass}=\text{false}) \)?

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

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

\[
P(\text{true, false}) = 0.01
\]

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

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. \( P(X=x) \)

\[
P(\text{EngPass} = \text{false}) = 0.01 + 0.07 = 0.125
\]
Properties of probabilities

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

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

Properties of probabilities

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

More generally:

- Given events \( E = e_1, e_2, \ldots, e_n \)
  \[ p(e_i) = 1 - \sum_{j \neq i} p(e_j) \]

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

Chain rule (aka product rule)

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

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

We can view calculating the probability of \( X \text{ AND } Y \) 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

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

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

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)

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

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

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

\[ p(\text{disease} \mid \text{symptoms}) \]

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

\[ p(\text{disease} \mid \text{symptoms}) \propto p(\text{symptoms} \mid \text{disease}) \]

How would you estimate this?

Find a bunch of people with the disease and see how many have this set of symptoms. Much easier!

Bayes rule

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

- For all examples that had those features, how many had that phenomenon?
- \[ p(\text{features} \mid \text{linguistic phenomena}) \]
- For all the examples with that phenomena, how many had this feature

\[ p(\text{cause} \mid \text{effect}) \text{ vs. } p(\text{effect} \mid \text{cause}) \]

Gaps

I just won't put these away.

These, I just won't put away.

I just won't put away.
Gaps

What did you put ___ away?

The socks that I put ___ away.

Whose socks did you fold ___ and put ___ away?

Whose socks did you fold ___?

Whose socks did you put ___ away?

Parasitic gaps

These I'll put ___ away without folding ___.

These I'll put ___ away.

These without folding ___.

Parasitic gaps

These I'll put ___ away without folding ___.

1. Cannot exist by themselves (parasitic)

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

2. They're optional

These I'll put ___ away without folding them.
Parasitic gaps

http://literalminded.wordpress.com/2009/02/10/dougs-parasitic-gap/

Frequency of parasitic gaps

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

Problem:
Laura Linguist has 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 is a parasitic gap, what is the probability it actually is?

\[ p(g|t) = ? \]
Prob of parasitic gaps

Laura Linguist has 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 is a parasitic gap, what is the probability it actually is?

\[
p(g \mid t) = \frac{p(t \mid g) p(g)}{p(t)} = \frac{p(t \mid g) p(g)}{\sum_{G \in G} p(g) p(t \mid g)} = \frac{p(t \mid g) p(g)}{p(g) p(t \mid g) + p(t \mid \neg g) p(t \mid \neg g)}
\]

\[
p(g \mid t) = \frac{0.95 \times 0.00001}{0.00001 \times 0.95 + 0.99999 \times 0.005} = 0.002
\]