Probabilities and Data

Computational Linguistics I: Jordan Boyd-Graber

University of Maryland

September 9, 2013

COLLEGE OF INFORMATION STUDIES

Slides adapted from Dave Blei and Lauren Hannah

Roadmap

- Introduction to the Course
- Administrivia
- Python
- What are probabilities
- How to manipulate probabilities
- Properties of probabilities

Outline

1. Computational Linguistics
2. Administrivia and Introductions
3. Introducing Python and NLTK
4. Probability
5. Properties of Probability Distributions
6. Working with probability distributions
7. Recap

Machine Learning is Doing Great!

- Can drive a million miles without an accident
- Can beat any living chess player

Machine Learning is Doing Great!

- Can drive a million miles without an accident
- Can beat any living chess player
- Automated call center vs. five-year old?

Machine Learning is Doing Great!

- Can drive a million miles without an accident
- Can beat any living chess player
- Automated call center vs. five-year old?
- We'll learn why we're so far away

What computational linguistics is

- Computational approaches to understand, generate, and process natural language
- Cross-discipline
  - Computer science: implement algorithms
  - Linguistics: develop theory / data
  - Statistics: learn patterns from data
  - Experts in specific languages: get a computer to handle a new language
  - Psychologists: how does our brain process language
  - Sociologists: how do social constraints change how we process language

What computational linguistics can do!

Automatic solutions to . . .

- Explain why the "ly" in "al**ly**" and "quick**ly**" are different (morphology)
- Tell the difference in category between "**water** the flowers" and "drink the **water**" (part of speech tagging)
- Why "saw the sun with the telescope" is different from "saw the astronomer with the telescope"
- Translate "My hovercraft is full of eels" into Hungarian (machine translation)

Outline

1. Computational Linguistics
2. **Administrivia and Introductions**
3. Introducing Python and NLTK
4. Probability
5. Properties of Probability Distributions
6. Working with probability distributions
7. Recap

What you need for this course

- Flipped classroom: watch lecture online, come to class ready to ask questions
- Helps to have a laptop to bring to class
- Math background
  - Will ask you to manipulate equations
  - Will expect you to be able to do basic derivatives
  - Work with functions like exponentiation and logs
  - Probability: review next week (hugely important)
- Computer / programming skills
  - You will need to write python programs
  - You will need to interact with a Unix command line
  - You will need to interact with data files

Administrivia

- Sign up on Piazza (use a photo)
- Keep track of course webpage
- 5 late days
- Let me know about special needs

Course reading

- We will provide reading assignments, mostly from the book. (Read them **before** associated class.)
- The reading will cover more than we cover in class.
- Don't buy the first edition

Communicating with Piazza

We will use Piazza to manage all communication

`https://piazza.com/umd/fall2013/cmsc723/home`

- Questions answered within 1 day (hopefully sooner)
- Hosts discussions among yourselves
- Use for any kind of technical question
- Use for **most** administrative questions
- Can use to send us private questions too

How to ask for help

- Explain what you're trying to do
- Give a minimal example
  - Someone else should be able to replicate the problem easily
  - Shouldn't require any data / information that only you have
- Explain what you **think** should happen
- Explain what you get instead (copy / paste or screenshot if you can)
- Explain what else you've tried

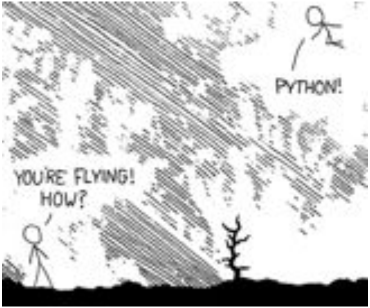
Me

- Fourth year assistant professor
  - iSchool and UMIACS
  - Offices: 2118C Hornbake / 3219 AV Williams
- First time teaching the class (taught second course in sequence before)
- Born in Colorado (where all my family live)
- Grew up in Iowa (hometown: Keokuk, Iowa)
- Went to high school in Arkansas
- Undergrad in California
- Grad school in New Jersey
- Brief jobs in between:
  - Working on electronic dictionary in Berlin
  - Worked on Google Books in New York
- ying / jbg / jordan / boyd-graber

Outline

Why Python?

- Easy to learn
- Widespread
- Can be fast if you need it (cython)



Why NLTK?

- Handy code for accessing data
- Implementations of standard algorithms
- Easy to quickly process text and try things out

Why NLTK?

- Handy code for accessing data
- Implementations of standard algorithms
- Easy to quickly process text and try things out
- Chapter 1 of NLTK book
- Ask questions on Piazza

Outline

Preface: Why make us do this?

- Probabilities are the language we use to describe data
- A reasonable (but geeky) definition of data science is how to get probabilities we care about from data
- Later classes will be about how to do this for different probability models and different types of data
- But first, we need key definitions of probability

Preface: Why make us do this?

- Probabilities are the language we use to describe data
- A reasonable (but geeky) definition of data science is how to get probabilities we care about from data
- Later classes will be about how to do this for different probability models and different types of data
- But first, we need key definitions of probability
- So pay attention!

The Statistical Revolution in NLP

- Speech recognition
- Machine translation
- Part of speech tagging
- Parsing

Solution?

They share the same solution: probabilistic models.

Outline

Random variable

- Probability is about *random variables.*
- A random variable is any "probabilistic" outcome.
- For example,
  - The flip of a coin
  - The height of someone chosen randomly from a population
- We'll see that it's sometimes useful to think of quantities that are not strictly probabilistic as random variables.
  - The temperature on 11/12/2013
  - The temperature on 03/04/1905
  - The number of times "streetlight" appears in a document

Random variable

- Random variables take on values in a *sample space*.
- They can be *discrete* or *continuous*:
  - Coin flip: $\{H, T\}$
  - Height: positive real values $(0, \infty)$
  - Temperature: real values $(-\infty, \infty)$
  - Number of words in a document: Positive integers $\{1, 2, \ldots\}$
- We call the outcomes *events*.
- Denote the random variable with a capital letter; denote a realization of the random variable with a lower case letter.
- E.g., X is a coin flip, x is the value (H or T) of that coin flip.
- This class will focus on **discrete** events.

Definition of Discrete Distribution

- A discrete distribution assigns a probability to every event in the sample space
- For example, if X is an (unfair) coin, then

$$\begin{aligned} P(X=H) &= 0.7 \\ P(X=T) &= 0.3 \end{aligned}$$

- The probabilities over the entire space must sum to one

$$\sum_x P(X=x) = 1$$

- And probabilities have to be greater than 0
- Probabilities of disjunctions are sums over part of the space. E.g., the probability that a die is bigger than 3:

$$P(D>3) = P(D=4) + P(D=5) + P(D=6)$$

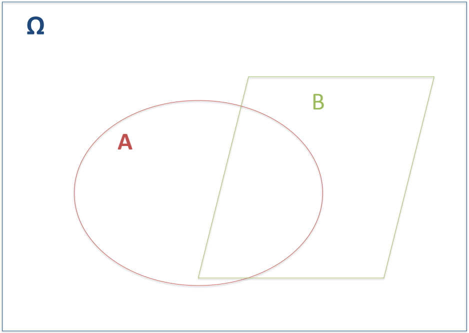
Outline

Events

An *event* is a set of outcomes to which a probability is assigned, for example, getting a card with Red on both sides.

Intersections and unions:

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



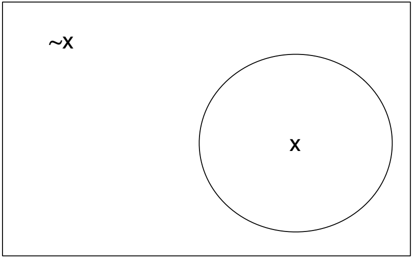
Joint distribution

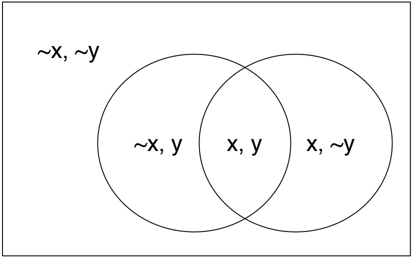
- Typically, we consider collections of random variables.
- The joint distribution is a distribution over the configuration of all the random variables in the ensemble.
- For example, imagine flipping 4 coins. The joint distribution is over the space of all possible outcomes of the four coins.

$$
\begin{aligned}
P(HHHH) &= 0.0625 \\
P(HHHT) &= 0.0625 \\
P(HHTH) &= 0.0625 \\
&\dots
\end{aligned}
$$

- You can think of it as a single random variable with 16 values.

Visualizing a joint distribution

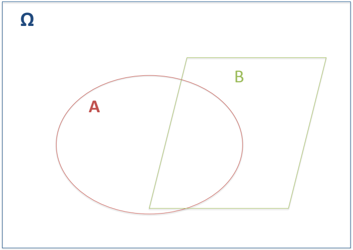




Conditional Probabilities

The *conditional probability* of event A given event B is the probability of A when B is known to occur,

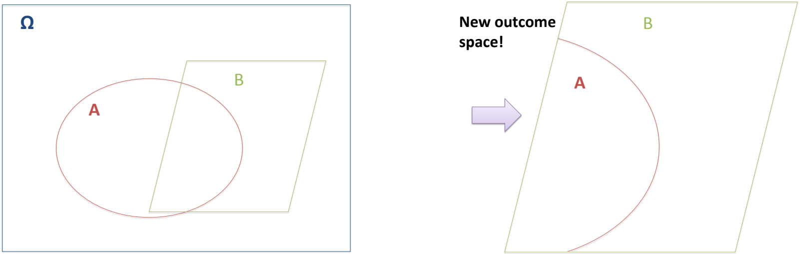
$$P(A|B) = \frac{P(A \cap B)}{P(B)}.$$



Conditional Probabilities

The *conditional probability* of event A given event B is the probability of A when B is known to occur,

$$P(A|B) = \frac{P(A \cap B)}{P(B)}.$$



Conditional Probabilities

Example

What is the probability that the sum of two dice is six given that the first is greater than three?

Conditional Probabilities

Example

What is the probability that the sum of two dice is six given that the first is greater than three?

- $A \equiv$ First die
- $B \equiv$ Second die

	B=1	B=2	B=3	B=4	B=5	B=6
A=1	2	3	4	5	6	7
A=2	3	4	5	6	7	8
A=3	4	5	6	7	8	9
A=4	5	6	7	8	9	10
A=5	6	7	8	9	10	11
A=6	7	8	9	10	11	12

Conditional Probabilities

Example

What is the probability that the sum of two dice is six given that the first is greater than three?

- $A \equiv$ First die
- $B \equiv$ Second die

	B=1	B=2	B=3	B=4	B=5	B=6
A=1	2	3	4	5	6	7
A=2	3	4	5	6	7	8
A=3	4	5	6	7	8	9
A=4	5	6	7	8	9	10
A=5	6	7	8	9	10	11
A=6	7	8	9	10	11	12

Conditional Probabilities

Example

What is the probability that the sum of two dice is six given that the first is greater than three?

- $A \equiv$ First die
- $B \equiv$ Second die

	B=1	B=2	B=3	B=4	B=5	B=6
A=1	2	3	4	5	6	7
A=2	3	4	5	6	7	8
A=3	4	5	6	7	8	9
A=4	5	6	7	8	9	10
A=5	6	7	8	9	10	11
A=6	7	8	9	10	11	12

$$P(A > 3 \cap B + A = 6) = \frac{2}{36}$$

$$P(B > 3) = \frac{3}{6}$$

$$P(A > 3 | B + A = 6) = \frac{\frac{2}{36}}{\frac{3}{6}} = \frac{2}{36}\frac{6}{3}$$

$$= \frac{1}{9}$$

The chain rule

- The definition of conditional probability lets us derive the *chain rule*, which let's us define the joint distribution as a product of conditionals:

$$P(X, Y)$$

The chain rule

- The definition of conditional probability lets us derive the *chain rule*, which let's us define the joint distribution as a product of conditionals:

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

The chain rule

- The definition of conditional probability lets us derive the *chain rule*, which let's us define the joint distribution as a product of conditionals:

$$\begin{aligned} P(X,Y) &= P(X,Y)\frac{P(Y)}{P(Y)} \\ &= P(X|Y)P(Y) \end{aligned}$$

The chain rule

- The definition of conditional probability lets us derive the *chain rule*, which let's us define the joint distribution as a product of conditionals:

$$
\begin{aligned}
P(X,Y) &= P(X,Y)\frac{P(Y)}{P(Y)} \\
&= P(X|Y)P(Y)
\end{aligned}
$$

- For example, let Y be a disease and X be a symptom. We may know $P(X|Y)$ and $P(Y)$ from data. Use the chain rule to obtain the probability of having the disease and the symptom.

The chain rule

- The definition of conditional probability lets us derive the *chain rule*, which let's us define the joint distribution as a product of conditionals:

$$\begin{aligned} P(X,Y) &= P(X,Y)\frac{P(Y)}{P(Y)} \\ &= P(X|Y)P(Y) \end{aligned}$$

- For example, let Y be a disease and X be a symptom. We may know $P(X|Y)$ and $P(Y)$ from data. Use the chain rule to obtain the probability of having the disease and the symptom.
- In general, for any set of N variables

$$P(X_1,\dots,X_N) = \prod_{n=1}^{N} P(X_n|X_1,\dots,X_{n-1})$$

Marginalization

If we are given a joint distribution, what if we are only interested in the distribution of one of the variables?

We can compute the distribution of $P(X)$ from $P(X, Y, Z)$ through *marginalization*:

$$\sum_y \sum_z P(X, Y = y, Z = z)$$

Marginalization

If we are given a joint distribution, what if we are only interested in the distribution of one of the variables?

We can compute the distribution of $P(X)$ from $P(X,Y,Z)$ through *marginalization*:

$$\sum_y \sum_z P(X, Y=y, Z=z) = \sum_y \sum_z P(X) P(Y=y, Z=z \mid X)$$

Marginalization

If we are given a joint distribution, what if we are only interested in the distribution of one of the variables?

We can compute the distribution of $P(X)$ from $P(X, Y, Z)$ through *marginalization*:

$$\begin{aligned}\sum_y \sum_z P(X, Y = y, Z = z) &= \sum_y \sum_z P(X) P(Y = y, Z = z | X) \\ &= P(X) \sum_y \sum_z P(Y = y, Z = z | X)\end{aligned}$$

Marginalization

If we are given a joint distribution, what if we are only interested in the distribution of one of the variables?

We can compute the distribution of $P(X)$ from $P(X,Y,Z)$ through *marginalization*:

$$\begin{aligned}\sum_y \sum_z P(X, Y=y, Z=z) &= \sum_y \sum_z P(X)P(Y=y, Z=z \,|\, X) \\ &= P(X) \sum_y \sum_z P(Y=y, Z=z \,|\, X) \\ &= P(X)\end{aligned}$$

Marginalization (from Leyton-Brown)

Joint distribution

temperature (T) and weather (W)

	T=Hot	T=Mild	T=Cold
W=Sunny	.10	.20	.10
W=Cloudy	.05	.35	.20

Marginalization allows us to compute distributions over smaller sets of variables:

- $P(X,Y) = \sum_z P(X,Y,Z=z)$
- Corresponds to summing out a table dimension
- New table still sums to 1

- Marginalize out weather
- Marginalize out temperature

Marginalization (from Leyton-Brown)

Joint distribution

temperature (T) and weather (W)

	T=Hot	T=Mild	T=Cold
W=Sunny	.10	.20	.10
W=Cloudy	.05	.35	.20

Marginalization allows us to compute distributions over smaller sets of variables:

- $P(X,Y) = \sum_z P(X,Y,Z=z)$
- Corresponds to summing out a table dimension
- New table still sums to 1

- Marginalize out weather

T=Hot	T=Mild	T=Cold
.15	.55	.30

- Marginalize out temperature

Marginalization (from Leyton-Brown)

Joint distribution

temperature (T) and weather (W)

	T=Hot	T=Mild	T=Cold
W=Sunny	.10	.20	.10
W=Cloudy	.05	.35	.20

Marginalization allows us to compute distributions over smaller sets of variables:

- $P(X,Y) = \sum_z P(X,Y,Z=z)$
- Corresponds to summing out a table dimension
- New table still sums to 1

- Marginalize out weather

T=Hot	T=Mild	T=Cold
.15	.55	.30

- Marginalize out temperature

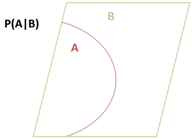
W=Sunny	.40
W=Cloudy	.60

Bayes' Rule

What is the relationship between $P(A|B)$ and $P(B|A)$?

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}$$

1. Start with $P(A|B)$
2. Change outcome space from B to Ω
3. Change outcome space again from Ω to A

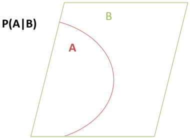


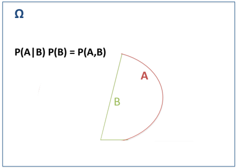
Bayes' Rule

What is the relationship between $P(A|B)$ and $P(B|A)$?

$$P(B|A) = \frac{P(A|B){\color{red}P(B)}}{P(A)}$$

1. Start with $P(A|B)$
2. Change outcome space from B to Ω
3. Change outcome space again from Ω to A



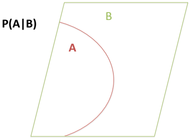


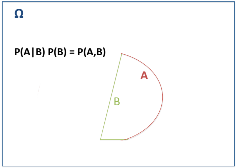
Bayes' Rule

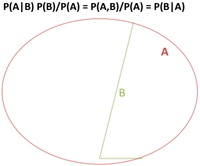
What is the relationship between $P(A|B)$ and $P(B|A)$?

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}$$

1. Start with $P(A|B)$
2. Change outcome space from B to Ω
3. Change outcome space again from Ω to A







Independence

Random variables X and Y are independent if and only if $P(X = x, Y = y) = P(X = x)P(Y = y)$.

Conditional probabilities equal unconditional probabilities with independence:

- $P(X = x \mid Y) = P(X = x)$
- *Knowing Y tells us nothing about X*

Mathematical examples:

- If I draw two socks from my (multicolored) laundry, is the color of the first sock independent from the color of the second sock?
- If I flip a coin twice, is the first outcome independent from the second outcome?

Independence

Intuitive Examples:

- Independent:
  - you use a Mac / the Green line is on schedule
  - snowfall in the Himalayas / your favorite color is blue
- Not independent:
  - you vote for Mitt Romney / you are a Republican
  - there is a traffic jam on the Beltway / the Redskins are playing

Independence

Intuitive Examples:

- Independent:
  - you use a Mac / the Green line is on schedule
  - snowfall in the Himalayas / your favorite color is blue
- Not independent:
  - you vote for Mitt Romney / you are a Republican
  - there is a traffic jam on the Beltway / the Redskins are playing
- But trust math, not your intuition (examples in class!)

Where do we go from here?

- Probability of the next word (language models)
- Probability of meaning given a word (sense disambiguation)
- Probability of a part of speech in a sentence (tagging)
- Probability of a syntactic structure given a sentence (parsing)
- Basically, the whole course is about probability (except for the next class)

Outline

Recap

- Welcome to computational linguistics!
- Intro to python
- Review of probability

In class . . .

- Introductions
- Quiz (answer: Marzipan)
- Installation issues
- Probability / Python questions