

How to Use Probabilities

The Crash Course

1

Goals of this lecture

- Probability notation like $p(X | Y)$:
 - What does this expression mean?
 - How can I manipulate it?
 - How can I estimate its value in practice?
- Probability models:
 - What is one?
 - Can we build one for language ID?
 - How do I know if my model is any good?

600.465 – Intro to NLP – J. Eisner

2

3 Kinds of Statistics

- **descriptive**: mean Hopkins SAT (or median)
- **confirmatory**: statistically significant?
- **predictive**: wanna bet?

this course – why?

600.465 – Intro to NLP – J. Eisner

3

Fugue for Tinhorns

- Opening number from *Guys and Dolls*
 - 1950 Broadway musical about gamblers
 - Words & music by Frank Loesser
- Video:
<http://www.youtube.com/watch?v=NxAX74eM8DY>
- Lyrics:
http://www.lyricsmania.com/fugue_for_tinhorns_lyrics_guys_and_dolls.html

600.465 – Intro to NLP – J. Eisner

4

Notation for Greenhorns



probability model

0.9

$$p(\text{Paul Revere wins} | \text{weather's clear}) = 0.9$$

600.465 – Intro to NLP – J. Eisner

5

What does that really mean?

$$p(\text{Paul Revere wins} | \text{weather's clear}) = 0.9$$

- Past performance?
 - Revere's won 90% of races with clear weather
 - Hypothetical performance?
 - If he ran the race in many parallel universes ...
 - Subjective strength of belief?
 - Would pay up to 90 cents for chance to win \$1
 - Output of some computable formula?
 - Ok, but then which formulas should we trust?
- $p(X | Y)$ versus $q(X | Y)$

600.465 – Intro to NLP – J. Eisner

6

p is a function on sets of "outcomes"

$p(\text{win} \mid \text{clear}) \equiv p(\text{win, clear}) / p(\text{clear})$

All Outcomes (races)

600.465 – Intro to NLP – J. Eisner 7

p is a function on sets of "outcomes"

$p(\text{win} \mid \text{clear}) \equiv p(\text{win, clear}) / p(\text{clear})$

syntactic sugar logical conjunction of predicates predicate selecting races where weather's clear

p measures total probability of a set of outcomes (an "event").

600.465 – Intro to NLP – J. Eisner 8

Required Properties of p (axioms) ^{most of the}

- $p(\emptyset) = 0$ $p(\text{all outcomes}) = 1$
- $p(X) \leq p(Y)$ for any $X \subseteq Y$
- $p(X) + p(Y) = p(X \cup Y)$ provided $X \cap Y = \emptyset$
e.g., $p(\text{win} \& \text{clear}) + p(\text{win} \& \neg \text{clear}) = p(\text{win})$

p measures total probability of a set of outcomes (an "event").

600.465 – Intro to NLP – J. Eisner 9

Commas denote conjunction

$p(\text{Paul Revere wins, Valentine places, Epitaph shows} \mid \text{weather's clear})$

what happens as we add conjuncts to left of bar ?

- probability can only decrease
- numerator of historical estimate likely to go to zero:
times Revere wins AND Val places... AND weather's clear
times weather's clear

600.465 – Intro to NLP – J. Eisner 10

Commas denote conjunction

$p(\text{Paul Revere wins, Valentine places, Epitaph shows} \mid \text{weather's clear})$

$p(\text{Paul Revere wins} \mid \text{weather's clear, ground is dry, jockey getting over sprain, Epitaph also in race, Epitaph was recently bought by Gonzalez, race is on May 17, ...})$

what happens as we add conjuncts to right of bar ?

- probability could increase or decrease
- probability gets more relevant to our case (less *bias*)
- probability *estimate* gets less reliable (more *variance*)
times Revere wins AND weather clear AND ... it's May 17
times weather clear AND ... it's May 17

600.465 – Intro to NLP – J. Eisner 11

Simplifying Right Side: Backing Off

$p(\text{Paul Revere wins} \mid \text{weather's clear, } \cancel{\text{ground is dry, jockey getting over sprain, Epitaph also in race, Epitaph was recently bought by Gonzalez, race is on May 17, ...}})$

not exactly what we want but at least we can get a reasonable estimate of it!
(i.e., more bias but less variance)

try to *keep* the conditions that we suspect will have the most influence on whether Paul Revere wins

600.465 – Intro to NLP – J. Eisner 12

Simplifying Left Side: Backing Off

$p(\text{Paul Revere wins, Valentine places, Epitaph shows} \mid \text{weather's clear})$

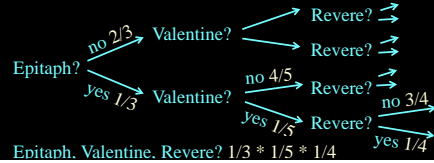
NOT ALLOWED!

but we can do something similar to help ...

Factoring Left Side: The Chain Rule

$$\begin{aligned}
 & p(\text{Revere, Valentine, Epitaph} \mid \text{weather's clear}) = \text{RVEW/W} \\
 = & p(\text{Revere} \mid \text{Valentine, Epitaph, weather's clear}) = \text{RVEW/VEW} \\
 & * p(\text{Valentine} \mid \text{Epitaph, weather's clear}) = * \text{VEW/EW} \\
 & * p(\text{Epitaph} \mid \text{weather's clear}) = * \text{EW/W}
 \end{aligned}$$

True because numerators cancel against denominators
 Makes perfect sense when read from bottom to top



Factoring Left Side: The Chain Rule

$$\begin{aligned}
 & p(\text{Revere, Valentine, Epitaph} \mid \text{weather's clear}) = \text{RVEW/W} \\
 = & p(\text{Revere} \mid \text{Valentine, Epitaph, weather's clear}) = \text{RVEW/VEW} \\
 & * p(\text{Valentine} \mid \text{Epitaph, weather's clear}) = * \text{VEW/EW} \\
 & * p(\text{Epitaph} \mid \text{weather's clear}) = * \text{EW/W}
 \end{aligned}$$

True because numerators cancel against denominators
 Makes perfect sense when read from bottom to top
 Moves material to right of bar so it can be ignored

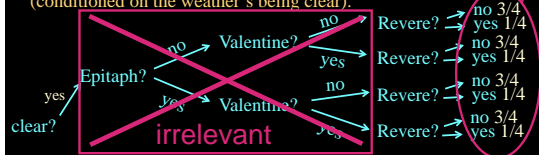
If this prob is unchanged by backoff, we say Revere was **CONDITIONALLY INDEPENDENT** of Valentine and Epitaph (conditioned on the weather's being clear). Often we just **ASSUME** conditional independence to get the nice product above.

Factoring Left Side: The Chain Rule

$$p(\text{Revere} \mid \text{Valentine, Epitaph, weather's clear})$$

conditional independence lets us use backed-off data from all four of these cases to estimate their shared probabilities

If this prob is unchanged by backoff, we say Revere was **CONDITIONALLY INDEPENDENT** of Valentine and Epitaph (conditioned on the weather's being clear).



“Courses in handicapping [i.e., estimating a horse's odds] should be required, like composition and Western civilization, in our universities. For sheer complexity, there's nothing like a horse race, excepting life itself ... To weigh and evaluate a vast grid of information, much of it meaningless, and to arrive at sensible, if erroneous, conclusions, is a skill not to be sneezed at.”

- Richard Russo, *The Risk Pool* (a novel)

Remember Language ID?

- “Horses and Lukasiewicz are on the curriculum.”
- Is this English or Polish or what?
- We had some notion of using n-gram models ...
- Is it “good” (= likely) English?
- Is it “good” (= likely) Polish?
- Space of outcomes will be not races but character sequences (x_1, x_2, x_3, \dots) where $x_n = \text{EOS}$

Remember Language ID?

- Let $p(X)$ = probability of text X in English
- Let $q(X)$ = probability of text X in Polish
- Which probability is higher?
 - (we'd also like bias toward English since it's more likely *a priori* – ignore that for now)

"Horses and Lukasiewicz are on the curriculum."

$$p(x_1=h, x_2=o, x_3=r, x_4=s, x_5=e, x_6=s, \dots)$$

600.465 – Intro to NLP – J. Eisner

19

Apply the Chain Rule

$$\begin{aligned} p(x_1=h, x_2=o, x_3=r, x_4=s, x_5=e, x_6=s, \dots) &= p(x_1=h) && 4470/52108 \\ * p(x_2=o | x_1=h) &&& 395/ 4470 \\ * p(x_3=r | x_1=h, x_2=o) &&& 5/ 395 \\ * p(x_4=s | x_1=h, x_2=o, x_3=r) &&& 3/ 5 \\ * p(x_5=e | x_1=h, x_2=o, x_3=r, x_4=s) &&& 3/ 3 \\ * p(x_6=s | x_1=h, x_2=o, x_3=r, x_4=s, x_5=e) &&& 0/ 3 \\ * \dots &= 0 \end{aligned}$$

counts from
Brown corpus
20

600.465 – Intro to NLP – J. Eisner

Back Off On Right Side

$$\begin{aligned} p(x_1=h, x_2=o, x_3=r, x_4=s, x_5=e, x_6=s, \dots) &\approx p(x_1=h) && 4470/52108 \\ * p(x_2=o | x_1=h) &&& 395/ 4470 \\ * p(x_3=r | x_1=h, x_2=o) &&& 5/ 395 \\ * p(x_4=s | x_2=o, x_3=r) &&& 12/ 919 \\ * p(x_5=e | x_3=r, x_4=s) &&& 12/ 126 \\ * p(x_6=s | x_4=s, x_5=e) &&& 3/ 485 \\ * \dots &= 7.3e-10 * \dots \end{aligned}$$

counts from
Brown corpus
21

600.465 – Intro to NLP – J. Eisner

Change the Notation

$$\begin{aligned} p(x_1=h, x_2=o, x_3=r, x_4=s, x_5=e, x_6=s, \dots) &\approx p(x_1=h) && 4470/52108 \\ * p(x_2=o | x_1=h) &&& 395/ 4470 \\ * p(x_i=r | x_{i-2}=h, x_{i-1}=o, i=3) &&& 5/ 395 \\ * p(x_i=s | x_{i-2}=o, x_{i-1}=r, i=4) &&& 12/ 919 \\ * p(x_i=e | x_{i-2}=r, x_{i-1}=s, i=5) &&& 12/ 126 \\ * p(x_i=s | x_{i-2}=s, x_{i-1}=e, i=6) &&& 3/ 485 \\ * \dots &= 7.3e-10 * \dots \end{aligned}$$

counts from
Brown corpus
22

600.465 – Intro to NLP – J. Eisner

Another Independence Assumption

$$\begin{aligned} p(x_1=h, x_2=o, x_3=r, x_4=s, x_5=e, x_6=s, \dots) &\approx p(x_1=h) && 4470/52108 \\ * p(x_2=o | x_1=h) &&& 395/ 4470 \\ * p(x_i=r | x_{i-2}=h, x_{i-1}=o) &&& 1417/14765 \\ * p(x_i=s | x_{i-2}=o, x_{i-1}=r) &&& 1573/26412 \\ * p(x_i=e | x_{i-2}=r, x_{i-1}=s) &&& 1610/12253 \\ * p(x_i=s | x_{i-2}=s, x_{i-1}=e) &&& 2044/21250 \\ * \dots &= 5.4e-7 * \dots \end{aligned}$$

counts from
Brown corpus
23

600.465 – Intro to NLP – J. Eisner

Simplify the Notation

$$\begin{aligned} p(x_1=h, x_2=o, x_3=r, x_4=s, x_5=e, x_6=s, \dots) &\approx p(x_1=h) && 4470/52108 \\ * p(x_2=o | x_1=h) &&& 395/ 4470 \\ * p(r | h, o) &&& 1417/14765 \\ * p(s | o, r) &&& 1573/26412 \\ * p(e | r, s) &&& 1610/12253 \\ * p(s | s, e) &&& 2044/21250 \\ * \dots &&& \end{aligned}$$

counts from
Brown corpus
24

600.465 – Intro to NLP – J. Eisner

Simplify the Notation

$p(x_1=h, x_2=o, x_3=r, x_4=s, x_5=e, x_6=s, \dots)$
 $\approx p(h \mid \text{BOS}, \text{BOS})$
 * $p(o \mid \text{BOS}, h)$
 * $p(r \mid h, o)$
 * $p(s \mid o, r)$
 * $p(e \mid r, s)$
 * $p(s \mid s, e)$
 * ... These basic probabilities are used to define $p(\text{horses})$

the parameters of our old trigram generator! Same assumptions about language.

values of those parameters, as naively estimated from Brown corpus.

counts from Brown corpus

4470/52108
395/ 4470
1417/14765
1573/26412
1610/12253
2044/21250

600.465 - Intro to NLP - J. Eisner 26

Simplify the Notation

$p(x_1=h, x_2=o, x_3=r, x_4=s, x_5=e, x_6=s, \dots)$
 $\approx t_{\text{BOS, BOS, h}}$
 * $t_{\text{BOS, h, o}}$
 * $t_{h, o, r}$
 * $t_{o, r, s}$
 * $t_{r, s, e}$
 * $t_{s, e, s}$
 * ... This notation emphasizes that they're just real variables whose value must be estimated

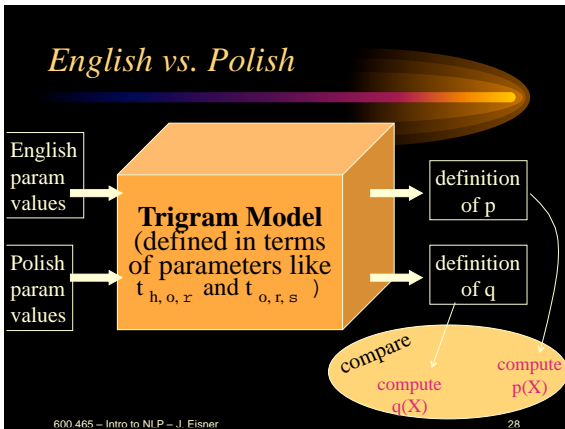
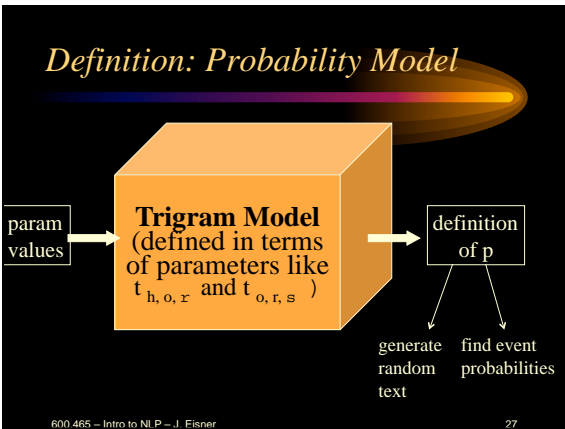
the parameters of our old trigram generator! Same assumptions about language.

values of those parameters, as naively estimated from Brown corpus.

counts from Brown corpus

4470/52108
395/ 4470
1417/14765
1573/26412
1610/12253
2044/21250

600.465 - Intro to NLP - J. Eisner 26



What is "X" in $p(X)$?

- Element (or subset) of some implicit "outcome space"
 - e.g., race
 - e.g., sentence
- What if outcome is a whole text?
 - $p(\text{text})$
 - $= p(\text{sentence 1}, \text{sentence 2}, \dots)$
 - $= p(\text{sentence 1})$
 - * $p(\text{sentence 2} \mid \text{sentence 1})$
 - * ...

definition of p
 definition of q
 compare: compute $q(X)$ compute $p(X)$

600.465 - Intro to NLP - J. Eisner 29

What is "X" in " $p(X)$ "?

- Element (or subset) of some implicit "outcome space"
 - e.g., race, sentence, text ...
- Suppose an outcome is a sequence of letters:
 - $p(\text{horses})$
- But we rewrote $p(\text{horses})$ as
 - $p(x_1=h, x_2=o, x_3=r, x_4=s, x_5=e, x_6=s, \dots)$
 - $\approx p(x_1=h) * p(x_2=o \mid x_1=h) * \dots$
- What does this variable=value notation mean?

600.465 - Intro to NLP - J. Eisner 30

Random Variables:
 What is “variable” in “ $p(\text{variable}=\text{value})$ ”?

Answer: variable is really a function of Outcome

- $p(x_1=h) * p(x_2=o | x_1=h) * \dots$
 - Outcome is a sequence of letters
 - x_2 is the second letter in the sequence
- $p(\text{number of heads}=2)$ or just $p(H=2)$ or $p(2)$
 - Outcome is a sequence of 3 coin flips
 - H is the number of heads
- $p(\text{weather's clear}=\text{true})$ or just $p(\text{weather's clear})$
 - Outcome is a race
 - weather's clear is true or false

600.465 – Intro to NLP – J. Eisner 31

Random Variables:
 What is “variable” in “ $p(\text{variable}=\text{value})$ ”?

Answer: variable is really a function of Outcome

- $p(x_1=h) * p(x_2=o | x_1=h) * \dots$
 - Outcome is a sequence of letters
 - $x_2(\text{Outcome})$ is the second letter in the sequence
- $p(\text{number of heads}=2)$ or just $p(H=2)$ or $p(2)$
 - Outcome is a sequence of 3 coin flips
 - $H(\text{Outcome})$ is the number of heads
- $p(\text{weather's clear}=\text{true})$ or just $p(\text{weather's clear})$
 - Outcome is a race
 - weather's clear (Outcome) is true or false

600.465 – Intro to NLP – J. Eisner 32

Random Variables:
 What is “variable” in “ $p(\text{variable}=\text{value})$ ”?

- $p(\text{number of heads}=2)$ or just $p(H=2)$
 - Outcome is a sequence of 3 coin flips
 - H is the number of heads in the outcome
- So $p(H=2)$
 - = $p(H(\text{Outcome})=2)$ picks out *set of outcomes w/2 heads*
 - = $p(\text{HHT,HTH,THH})$
 - = $p(\text{HHT})+p(\text{HTH})+p(\text{THH})$

TTT	TTH	HTT	HTH
THT	THH	HHT	HHH

All Outcomes 33

Random Variables:
 What is “variable” in “ $p(\text{variable}=\text{value})$ ”?

- $p(\text{weather's clear})$
 - Outcome is a race
 - weather's clear is true or false of the outcome
- So $p(\text{weather's clear})$
 - = $p(\text{weather's clear}(\text{Outcome})=\text{true})$
 - picks out the *set of outcomes with clear weather*

$p(\text{win} | \text{clear}) \equiv p(\text{win, clear}) / p(\text{clear})$

600.465 – Intro to NLP – J. Eisner 34

Random Variables:
 What is “variable” in “ $p(\text{variable}=\text{value})$ ”?

- $p(x_1=h) * p(x_2=o | x_1=h) * \dots$
 - Outcome is a sequence of letters
 - x_2 is the second letter in the sequence
- So $p(x_2=o)$
 - = $p(x_2(\text{Outcome})=o)$ picks out *set of outcomes with ...*
 - = $\sum p(\text{Outcome})$ over all outcomes whose second letter ...
 - = $p(\text{horses}) + p(\text{boffo}) + p(\text{xoyzkklp}) + \dots$

600.465 – Intro to NLP – J. Eisner 35

Back to trigram model of $p(\text{horses})$

$p(x_1=h, x_2=o, x_3=r, x_4=s, x_5=e, x_6=s, \dots)$

$\approx t_{\text{BOS, BOS, h}}$
 $* t_{\text{BOS, h, o}}$
 $* t_{\text{h, o, r}}$
 $* t_{\text{o, r, s}}$
 $* t_{\text{r, s, e}}$
 $* t_{\text{s, e, s}}$
 $* \dots$

the parameters of our old trigram generator! Same assumptions about language.

values of those parameters, as naively estimated from Brown corpus.

counts from Brown corpus

4470/52108
395/ 4470
1417/ 14765
1573/26412
1610/ 12253
2044/21250

This notation emphasizes that they're just real variables whose value must be estimated

600.465 – Intro to NLP – J. Eisner 36

A Different Model

- Exploit fact that *horses* is a common word

$$p(W_i = \text{horses})$$

where word vector W is a function of the outcome (the sentence) just as character vector X is.

$$= p(W_i = \text{horses} | i=1)$$

$$\approx p(W_i = \text{horses}) = 7.2e-5$$

independence assumption says that sentence-initial words w_i are just like all other words w_i (gives us more data to use)

Much larger than previous estimate of $5.4e-7$ – why?

Advantages, disadvantages?

600.465 – Intro to NLP – J. Eisner 37

Improving the New Model: Weaken the Indep. Assumption

- Don't totally cross off $i=1$ since it's not irrelevant:
 - Yes, *horses* is common, but less so at start of sentence since most sentences start with determiners.

$$p(W_i = \text{horses}) = \sum_t p(W_i = \text{horses}, T_i = t)$$

$$= \sum_t p(W_i = \text{horses} | T_i = t) * p(T_i = t)$$

$$= \sum_t p(W_i = \text{horses} | T_i = t, i=1) * p(T_i = t)$$

$$\approx \sum_t p(W_i = \text{horses} | T_i = t) * p(T_i = t)$$

$$= p(W_i = \text{horses} | T_i = \text{PlNoun}) * p(T_i = \text{PlNoun})$$

$$+ p(W_i = \text{horses} | T_i = \text{Verb}) * p(T_i = \text{Verb}) + \dots$$

$$= (72 / 55912) * (977 / 52108) + (0 / 15258) * (146 / 52108) + \dots$$

$$= 2.4e-5 + 0 + \dots + 0 = 2.4e-5$$

600.465 – Intro to NLP – J. Eisner 38

Which Model is Better?

- Model 1** – predict each letter X_i from previous 2 letters X_{i-2}, X_{i-1}
- Model 2** – predict each word W_i by its part of speech T_i , having predicted T_i from i
- Models make different independence assumptions that reflect different intuitions
- Which intuition is better???

600.465 – Intro to NLP – J. Eisner 39

Measure Performance!

- Which model does better on language ID?
 - Administer test where you know the right answers
 - Seal up test data until the test happens
 - Simulates real-world conditions where new data comes along that you didn't have access to when choosing or training model
 - In practice, split off a test set as soon as you obtain the data, and never look at it
 - Need *enough* test data to get statistical significance
 - Report *all* results on test data
- For a different task (e.g., speech transcription instead of language ID), use that task to evaluate the models

600.465 – Intro to NLP – J. Eisner 40

Cross-Entropy (“xent”)

- Another common measure of model quality
 - Task-independent
 - Continuous – so slight improvements show up here even if they don't change # of right answers on task
- Just measure probability of (enough) test data
 - Higher prob means model better predicts the future
 - There's a limit to how well you can predict random stuff
 - Limit depends on “how random” the dataset is (easier to predict weather than headlines, especially in Arizona)

600.465 – Intro to NLP – J. Eisner 41

Cross-Entropy (“xent”)

- Want prob of test data to be high:
 - Average? Geometric average of $1/2^1, 1/2^2, 1/2^3, 1/2^4, \dots$
 - $1/8 * 1/8 * 1/8 * 1/16 \dots = 1/2^{3.25} \approx 1/9.5$
- high prob → low xent by 3 cosmetic improvements:
 - Take logarithm (base 2) to prevent underflow:
 - $\log(1/8 * 1/8 * 1/8 * 1/16 \dots)$
 - $= \log 1/8 + \log 1/8 + \log 1/8 + \log 1/16 \dots = (-3) + (-3) + (-3) + (-4) + \dots$
 - Negate to get a positive value in *bits* $3+3+3+4+\dots$
 - Divide by length of text → **3.25 bits per letter (or per word)**
 - Want this to be small (equivalent to wanting good compression!)
 - Lower limit is called *entropy* – obtained in principle as cross-entropy of the *true model* measured on an infinite amount of data
 - Or use **perplexity** = 2 to the xent (≈9.5 choices instead of 3.25 bits)

600.465 – Intro to NLP – J. Eisner 42