How to Use Probabilities

The Crash Course

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?

3 Kinds of Statistics

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

this course – why?

Notation for Greenhorns

0.9

probability model

p(Paul Revere wins | weather's clear) = 0.9

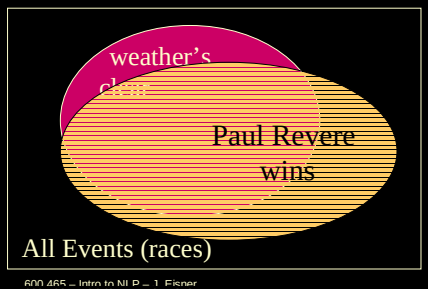
What does that really mean?

p(Paul Revere wins | 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)

p is a function on event sets

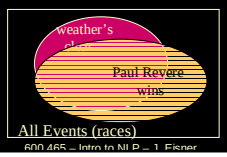
p(win | clear) ≡ p(win, clear) / p(clear)



p is a function on event sets

p(win | clear) ≡ p(win, clear) / p(clear)

- p(win | clear): syntactic sugar
- p(win, clear): logical conjunction of predicates
- p(clear): predicate selecting races where weather's clear



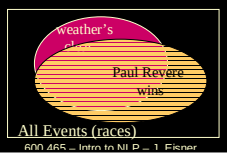
p measures total probability of a set of events.

Required Properties of p (axioms)

most of the

- p(∅) = 0 p(all events) = 1
- p(X) ≤ p(Y) for any X ⊆ Y
- p(X) + p(Y) = p(X ∪ Y) provided X ∩ Y=∅

e.g., p(win & clear) + p(win & ¬clear) = p(win)



p measures total probability of a set of events.

Commas denote conjunction

p(Paul Revere wins, Valentine places, Epitaph shows | 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:

$$\frac{\#\text{ times Revere wins AND Val places… AND weather's clear}}{\#\text{ times weather's clear}}$$

Commas denote conjunction

p(Paul Revere wins, Valentine places, Epitaph shows | weather's clear)

p(Paul Revere wins | 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***)

$$\frac{\#\text{ times Revere wins AND weather clear AND … it's May 17}}{\#\text{ times weather clear AND … it's May 17}}$$

Simplifying Right Side: Backing Off

p(Paul Revere wins | 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, …~~)

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

Simplifying Right Side: Backing Off

p(Paul Revere wins, ~~Valentine places, Epitaph shows~~ | weather's clear)

NOT ALLOWED!

but we can do something similar to help …

Factoring Left Side: The Chain Rule

p(Revere, Valentine, Epitaph | weather's clear) RVEW/W
= p(Revere | ~~Valentine, Epitaph~~, weather's clear) = RVEW/VEW
* p(Valentine | ~~Epitaph~~, weather's clear) * VEW/EW
* p(Epitaph | weather's clear) * EW/W

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.

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 events will be not races but character sequences (x_1, x_2, x_3, …) where x_n = 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, …)

Apply the Chain Rule

p(x_1=h, x_2=o, x_3=r, x_4=s, x_5=e, x_6=s, …)
= 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
* … = 0

counts from Brown corpus

Back Off On Right Side

p(x_1=h, x_2=o, x_3=r, x_4=s, x_5=e, x_6=s, …)
≈ 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
* … = 7.3e-10 * …

counts from Brown corpus

Change the Notation

p(x_1=h, x_2=o, x_3=r, x_4=s, x_5=e, x_6=s, …)
≈ 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
* … = 7.3e-10 * …

counts from Brown corpus

Another Independence Assumption

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

$\approx p(x_1=h)$ 4470/52108

$* p(x_2=o \mid x_1=h)$ 395/ 4470

$* p(x_i=r \mid x_{i-2}=h, x_{i-1}=o)$ 1417/14765

$* p(x_i=s \mid x_{i-2}=o, x_{i-1}=r)$ 1573/26412

$* p(x_i=e \mid x_{i-2}=r, x_{i-1}=s)$ 1610/12253

$* p(x_i=s \mid x_{i-2}=s, x_{i-1}=e)$ 2044/21250

$* \ldots = 5.4e\text{-}7 * \ldots$

counts from Brown corpus

Simplify the Notation

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

$\approx p(x_1=h)$ 4470/52108

$* p(x_2=o \mid x_1=h)$ 395/ 4470

$* p(r \mid h, o)$ 1417/14765

$* p(s \mid o, r)$ 1573/26412

$* p(e \mid r, s)$ 1610/12253

$* p(s \mid s, e)$ 2044/21250

$* \ldots$

counts from Brown corpus

Simplify the Notation

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

$\approx p(h \mid BOS, BOS)$

$* p(o \mid BOS, h)$

$* p(r \mid h, o)$

$* p(s \mid o, r)$

$* p(e \mid r, s)$

$* p(s \mid s, e)$

$* \ldots$

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

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

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

These basic probabilities are used to define p(horses)

counts from Brown corpus

Simplify the Notation

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

$\approx t_{BOS, BOS, h}$

$* t_{BOS, h, o}$

$* t_{h, o, r}$

$* t_{o, r, s}$

$* t_{r, s, e}$

$* t_{s, e, s}$

$* \ldots$

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

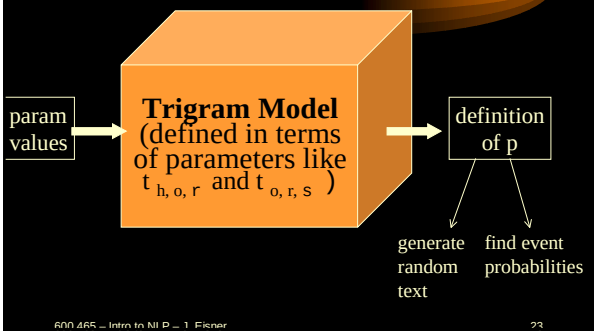
values of those parameters, as naively estimated from Brown corpus.

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

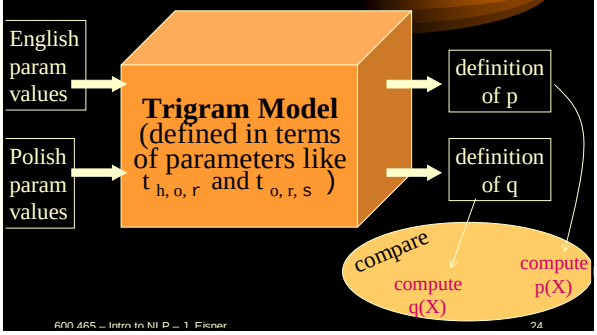
This notation emphasizes that they're just real variables whose value must be estimated

counts from Brown corpus

Definition: Probability Model

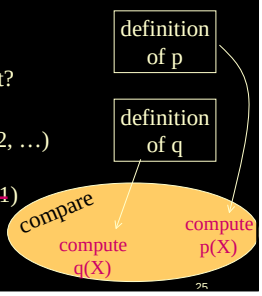


English vs. Polish



What is "X" in p(X)?

- Element of some implicit "event space"
  - e.g., race
  - e.g., sentence
- What if event is a whole text?
  - p(text)
    = p(sentence 1, sentence 2, …)
    = p(sentence 1)
    * p(sentence 2 | ~~sentence 1~~)
    * …



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

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

Random Variables: What is "variable" in "p(variable=value)"?

Answer: variable is really a function of Event

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

Random Variables: What is "variable" in "p(variable=value)"?

Answer: variable is really a function of Event

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

Random Variables: What is "variable" in "p(variable=value)"?

- p(number of heads=2) or just p(H=2)
  - Event is a sequence of 3 coin flips
  - H is the number of heads in the event
  - So p(H=2)
    = p(H(Event)=2) picks out the *set* of events with 2 heads
    = p({HHT,HTH,THH})
    = p(HHT)+p(HTH)+p(THH)

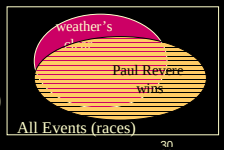
TTT	TTH	HTT	HTH
THT	THH	HHT	HHH

All Events

Random Variables: What is "variable" in "p(variable=value)"?

- p(weather's clear)
  - Event is a race
  - weather's clear is true or false of the event
  - So p(weather's clear)
    = p(weather's clear(Event)=true)
    picks out the *set* of events with clear weather

p(win | clear) ≡ p(win, clear) / p(clear)



Random Variables: What is "variable" in "p(variable=value)"?

- $p(x_1=h) * p(x_2=o \mid x_1=h) * \ldots$
  - Event is a sequence of letters
  - x_2 is the second letter in the sequence
  - So $p(x_2=o)$

    $= p(x_2(\text{Event})=o)$ picks out the *set* of events with …

    $= \Sigma\, p(\text{Event})$ over all events whose second letter …

    = p(horses) + p(boffo) + p(xoyzkklp) + …

Back to trigram model of p(horses)

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

$\approx t_{\text{BOS, BOS, h}}$

$* t_{\text{BOS, h, o}}$

$* t_{\text{h, o, r}}$

$* t_{\text{o, r, s}}$

$* t_{\text{r, s, e}}$

$* t_{\text{s, e, s}}$

$* \ldots$

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

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

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

counts from Brown corpus

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

A Different Model

- Exploit fact that horses is a common word

$p(W_1 = \texttt{horses})$

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

$= p(W_i = \texttt{horses} \mid \sout{i=1})$

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

independence assumption says that sentence-initial words w_1 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?

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.

$p(W_1 = \texttt{horses}) = \Sigma_t\, p(W_1=\texttt{horses}, T_1 = t)$

$= \Sigma_t\, p(W_1=\texttt{horses} \mid T_1 = t) * p(T_1 = t)$

$= \Sigma_t\, p(W_i=\texttt{horses} \mid T_i = t, i=1) * p(T_1 = t)$

$\approx \Sigma_t\, p(W_i=\texttt{horses} \mid T_i = t) * p(T_1 = t)$

$= p(W_i=\texttt{horses} \mid T_i = \texttt{PlNoun}) * p(T_1 = \texttt{PlNoun})$

(if first factor is 0 for any other part of speech)

$\approx (72 / 55912) * (977 / 52108)$

$= 2.4\text{e-}5$

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???

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
- For a different task (e.g., speech transcription instead of language ID), use that task to evaluate the models

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)

Cross-Entropy ("xent")

- Want prob of test data to be high:
  p(h | BOS, BOS) * p(o | BOS, h) * p(r | h, o) * p(s | o, r) …
  1/8 * 1/8 * 1/8 * 1/16 …
- high prob → low xent by 3 cosmetic improvements:
  - Take logarithm (base 2) to prevent underflow:
    log (1/8 * 1/8 * 1/8 * 1/16 …)
    = log 1/8 + log 1/8 + log 1/8 + log 1/16 … = (-3) + (-3) + (-3) + (-4) + …
  - Negate to get a positive value in *bits* 3+3+3+4+…
  - Divide by length of text to get *bits per letter* or *bits per word*
    - Want this to be small (equivalent to wanting good compression!)
    - Lower limit is called ***entropy*** – obtained in principle as cross-entropy of best possible model on an infinite amount of test data
  - Or use ***perplexity*** = 2 to the xent (9.5 choices instead of 3.25 bits)