Statistical NLP

Mathematical Foundations & Language Modeling

Notions of Probability Theory

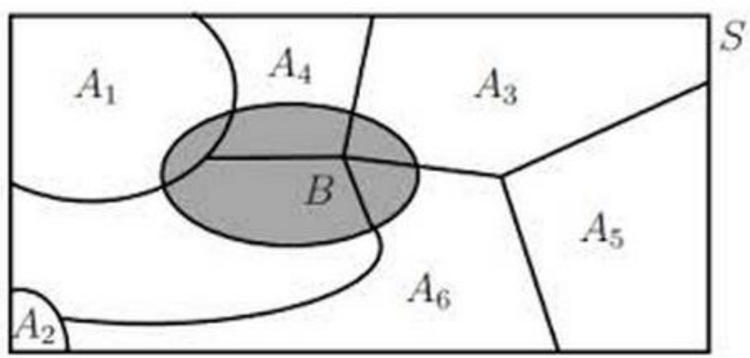
- ***Probability theory*** deals with predicting how likely it is that something will happen.
- The process by which an observation is made is called an ***experiment*** or a ***trial***.
- The collection of ***basic outcomes*** (or ***sample points***) for our experiment is called the ***sample space***.
- An ***event*** is a subset of the sample space.
- Probabilities are numbers between 0 and 1, where 0 indicates impossibility and 1, certainty.
- A ***probability function/distribution*** distributes a probability mass of 1 throughout the sample space.

Conditional Probability and Independence

- ***Conditional probabilities*** measure the probability of events given some knowledge.
- ***Prior probabilities*** measure the probabilities of events before we consider our additional knowledge.
- ***Posterior probabilities*** are probabilities that result from using our additional knowledge.
- The ***chain rule*** relates intersection with conditionalization (important to NLP)
- ***Independence*** and ***conditional independence*** of events are two very important notions in statistics.

Baye's Theorem

- ***Baye's Theorem*** lets us swap the order of dependence between events. This is important when the former quantity is difficult to determine.
- ***P(A|B) = P(B|A)P(A)/P(B)***
- P(B) is a ***normalization constant***.



Random Variables

- A ***random variable*** is a **function**
  X: sample space --> R^n
- A ***discrete random variable*** is a function
  X: sample space --> S
  where S is a countable subset of R.
- If X: sample space --> {0,1}, then X is called a ***Bernoulli trial***.
- The ***probability mass function*** for a random variable X gives the probability that the random variable has different numeric values.

Expectation and Variance

- The ***expectation*** is the ***mean*** or average of a random variable.
- The ***variance*** of a random variable is a measure of whether the values of the random variable tend to be consistent over trials or to vary a lot.

Joint and Conditional Distributions

- More than one random variable can be defined over a sample space. In this case, we talk about a ***joint*** or ***multivariate*** probability distribution.
- The ***joint probability mass function*** for two discrete random variables X and Y is: $p(x,y)=P(X=x, Y=y)$
- The ***marginal probability mass function*** sums up the joint probability masses

Estimating Probability Functions

- What is the P(probability) that sentence “I love her so much” will be uttered? Unknown
  - P must be ***estimated*** from a sample of data.
- An important measure for estimating P is the ***relative frequency*** of the outcome, i.e., the proportion of times a certain outcome occurs.
- Assuming that certain aspects of language can be modeled by one of the well-known distribution is called using a ***parametric*** approach.
- If no such assumption can be made, we must use a ***non-parametric*** approach.

Standard Distributions

- In practice, one commonly finds the same basic form of a probability mass function, but with different constants employed.
- Families of PMFs are called ***distributions*** and the constants that define the different possible PMFs in one family are called ***parameters***.
- Discrete Distributions: the ***binomial distribution***, the ***multinomial distribution***, the ***Poisson distribution***.
- Continuous Distributions: the ***normal distribution***, the ***standard normal distribution***.

Baysian Statistics I: Bayesian Updating

- Assume that the data are coming in sequentially and are independent.
- Given an a-priori probability distribution, we can update our beliefs when a new datum comes in by calculating the ***<u>Maximum A Posteriori (MAP)</u>*** distribution.

$$
\begin{aligned}
h_{MAP} &\equiv \underset{h \in H}{\operatorname{argmax}}\, P(h|D) \\
&= \underset{h \in H}{\operatorname{argmax}}\, \frac{P(D|h)\,P(h)}{P(D)} \\
&= \underset{h \in H}{\operatorname{argmax}}\, P(D|h)\,P(h)
\end{aligned}
$$

- The MAP probability becomes the new prior and the process repeats on each new datum.

Bayesian Statistics II: Bayesian Decision Theory

- If we assume prior probabilities of h are same, then it is Maximum Likelyhood

$$h_{ML} \equiv \underset{h \in H}{\operatorname{argmax}} P(D|h)$$

Entropy

- The entropy is the average uncertainty of a single random variable.
- Let $\boldsymbol{p(x)=P(X=x)};$ where $\boldsymbol{x} \in \mathscr{X}$.
- $\boldsymbol{H(p)= H(X)= - \Sigma_{x \in \mathscr{X}} p(x) \log_2 p(x)}$
- In other words, entropy measures the amount of information in a random variable. It is normally measured in bits.

Joint Entropy and Conditional Entropy

- The ***joint entropy*** of a pair of discrete random variables X, Y ~ p(x,y) is the amount of information needed on average to specify both their values.
- $H(X,Y) = - \Sigma_{x \in \mathcal{X}} \Sigma_{y \in \mathcal{Y}} p(x,y) \log_2 p(X,Y)$
- The ***conditional entropy*** of a discrete random variable Y given another X, for X, Y ~ p(x,y), expresses how much extra information you still need to supply on average to communicate Y given that the other party knows X.
- $H(Y|X) = - \Sigma_{x \in \mathcal{X}} \Sigma_{y \in \mathcal{Y}} p(x,y) \log_2 p(y|x)$
- ***Chain Rule*** for Entropy: $H(X,Y)=H(X)+H(Y|X)$

Mutual Information (1/3)

- By the chain rule for entropy, we have $\boldsymbol{H(X,Y) = H(X) + H(Y|X) = H(Y) + H(X|Y)}$
- Therefore, $\boldsymbol{H(X) - H(X|Y) = H(Y) - H(Y|X)}$
- This difference is called the ***mutual information between X and Y***.
- It is the reduction in uncertainty of one random variable due to knowing about another, or, in other words, the amount of information one random variable contains about another.

Mutual Information (2/3)

- MI measures the mutual dependence of two random variables.
  - The higher it is, the more dependent the two random variables are with each other.
  - Its value is always positive.
- The equation of MI:

$$MI(X;Y) = \sum_{x,y} p(x,y) \log\left(\frac{p(x,y)}{p(x)\,p(y)}\right)$$

Mutual Information (3/3)

- For example, say a discrete random variable *X* represents visibility at a certain moment in time and random variable *Y* represents wind speed at that moment.
- The mutual information between *X* and *Y:*

$$
\begin{aligned}
MI(X;Y) = {} & p(X=good, Y=high)\log\left(\frac{p(X=good, Y=high)}{p(X=good)\,p(Y=high)}\right) + \\
& p(X=bad, Y=high)\log\left(\frac{p(X=bad, Y=high)}{p(X=bad)\,p(Y=high)}\right) + \\
& p(X=good, Y=low)\log\left(\frac{p(X=good, Y=low)}{p(X=good)\,p(Y=low)}\right) + \\
& p(X=bad, Y=low)\log\left(\frac{p(X=bad, Y=low)}{p(X=bad)\,p(Y=low)}\right)
\end{aligned}
$$

Pointwise Mutual Information (PMI)

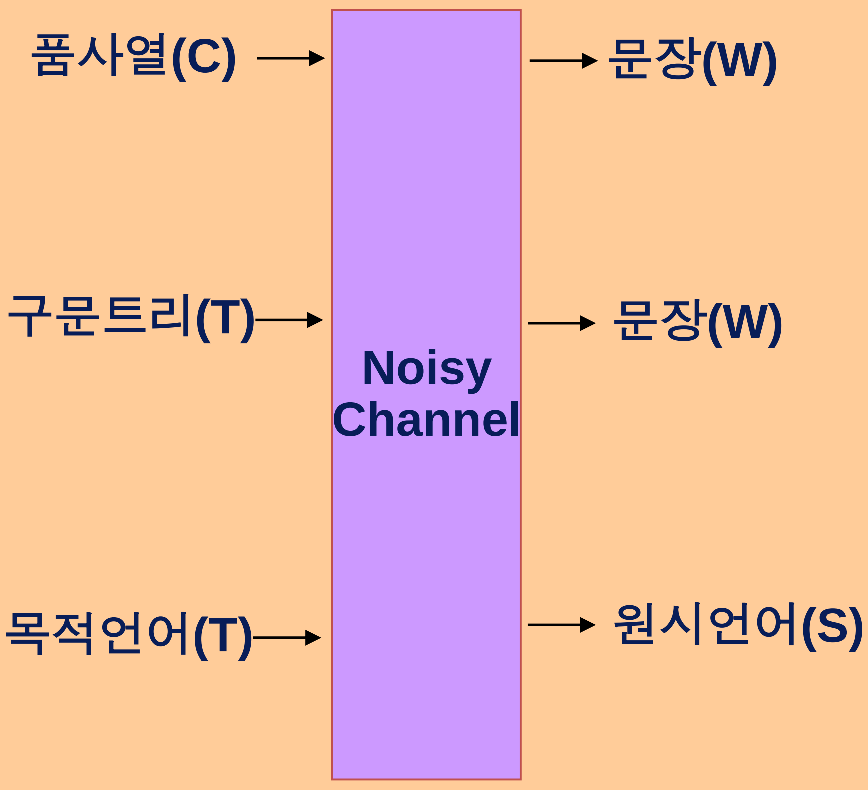
- PMI measures the mutual dependence of between two instances or realizations of random variables.
  - Positive => high correlated
  - Zeros => no information (independence)
  - Negative => opposite correlated

- The equation of PMI:

$$PMI(X=x, Y=y) = \log\left(\frac{p(X=x, Y=y)}{p(X=x)\,p(Y=y)}\right)$$

The Noisy Channel Model

- Noisy channel model for NLP



$$\arg\max_c P(C \mid W) = \arg\max_c \frac{P(C,W)}{P(W)} = \arg\max_c P(C,W)$$

$$\arg\max_t P(T \mid W) = \arg\max_t \frac{P(T,W)}{P(W)} = \arg\max_t P(T,W)$$

$$\arg\max_t P(T \mid S) = \arg\max_t \frac{P(T,S)}{P(S)} = \arg\max_c P(T,S)$$

Relative Entropy or Kullback-Leibler Divergence

- For 2 pmfs, p(x) and q(x), their ***relative entropy*** is:
- $D(p\|q)= \boldsymbol{\Sigma_{x \in \mathscr{X}} p(x)\log(p(x)/q(x))}$
- The relative entropy (also known as the ***Kullback-Leibler divergence***) is **a measure of how different two probability distributions (over the same event space) are**.
- The KL divergence between p and q can also be seen as the average number of bits that are wasted by encoding events from a distribution p with a code based on a not-quite-right distribution q.

The Relation to Language: Cross-Entropy

- Entropy can be thought of as a matter of how surprised we will be to see the next word given previous words we already saw.
- The ***<u>cross entropy</u>*** between a random variable X with true probability distribution p(x) and another pmf q (normally a model of p) is given by: $H(X,q)=H(X)+D(p||q)$.
- Cross-entropy can help us find out what our average surprise for the next word is.

The Entropy of English

- We can model English using ***n-gram models*** (also known a ***Markov chains***).
- These models assume limited memory, i.e., we assume that the next word depends only on the previous k ones [***kth order Markov approximation***].
- What is the Entropy of English?

Perplexity

- A measure related to the notion of cross-entropy and used in the speech recognition community is called the perplexity.
- Perplexity(x_{1n}, m) $=2^{H(x_{1n},m)} = m(x_{1n})^{-1/n}$
- A perplexity of k means that you are as surprised on average as you would have been if you had had to guess between k equiprobable choices at each step.

How to Use Probabilities

Goals of this lecture

- Probability notation like p(Y | X):
  - 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?

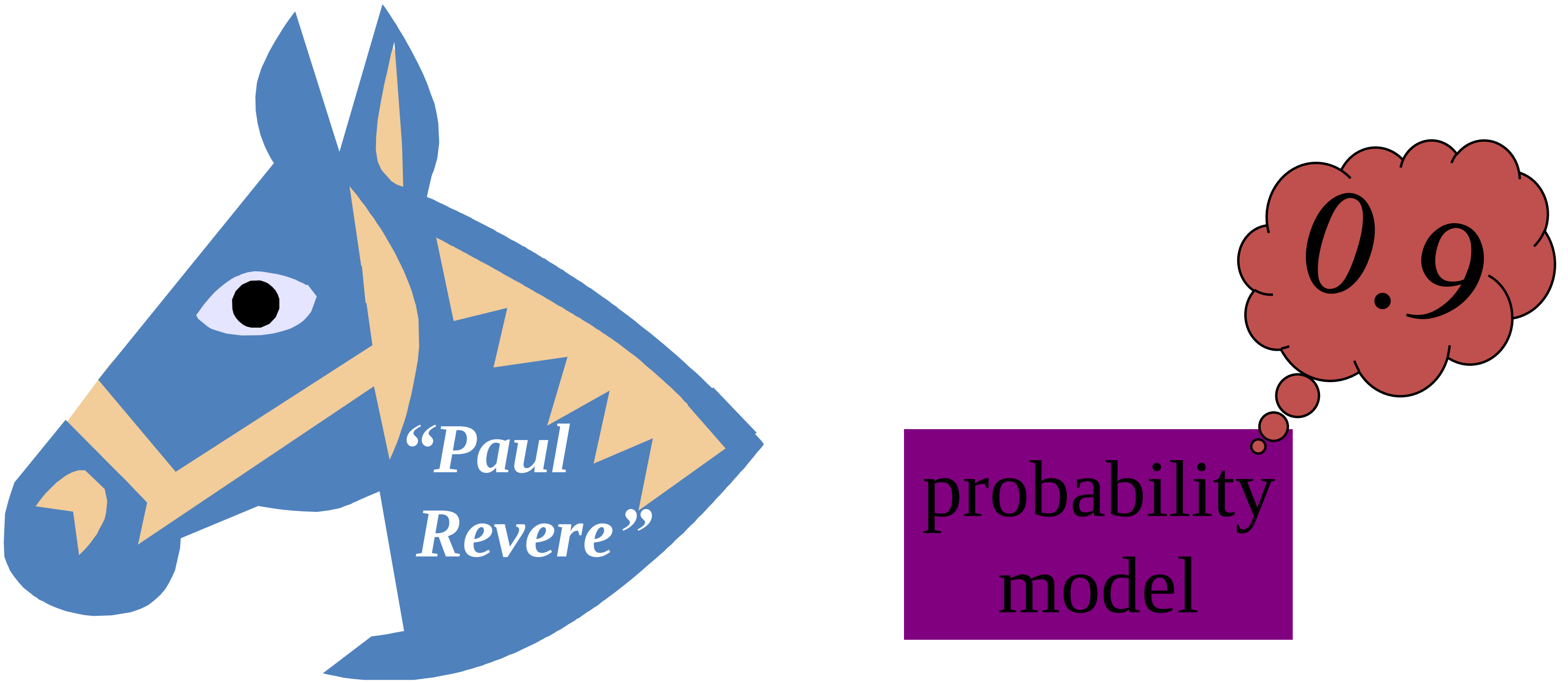
this course – why?

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=NxAX74gM8DY
- Lyrics: http://www.lyricsmania.com/fugue_for_tinhorns_lyrics_guys_and_dolls.html

Notation for Greenhorns



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

What does that really mean?

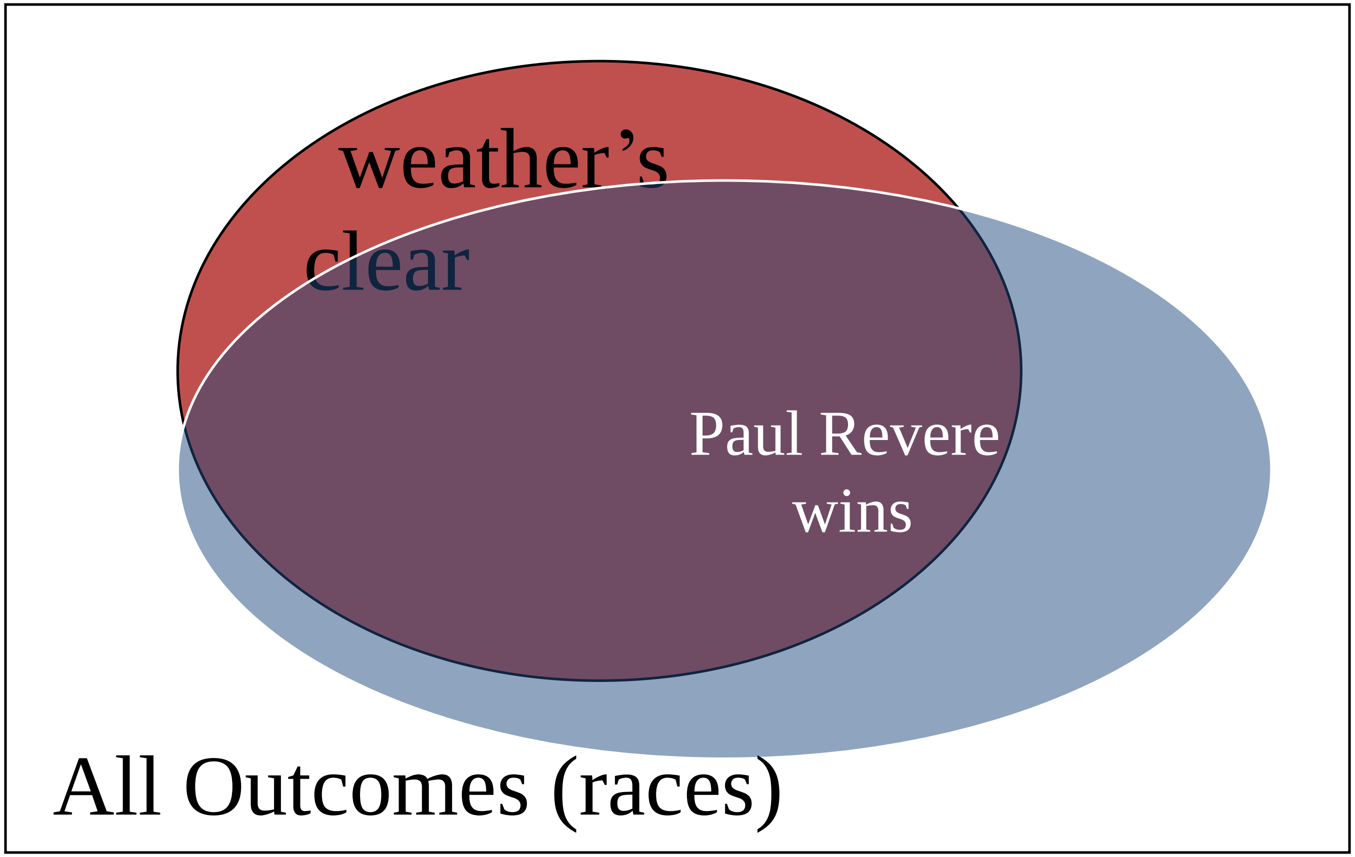
p(Paul Revere wins | weather’s clear) = 0.9

- Past performance?
  - Revere’s won 90% of races with clear weather
- Output of some computable formula?
  - Ok, but then which formulas should we trust?

    $p(Y \mid X)$ versus $q(Y \mid X)$

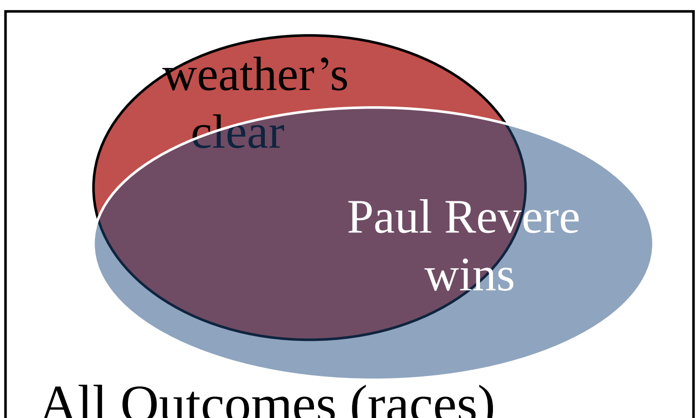
p is a function on sets of "outcomes"

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



p is a function on sets of “outcomes”

$$\underbrace{p(\text{win} \mid \text{clear})}_{\text{syntactic sugar}} \equiv p(\underbrace{\text{win, clear}}_{\substack{\text{logical conjunction}\\ \text{of predicates}}}) \,/\, p(\underbrace{\text{clear}}_{\substack{\text{predicate selecting}\\ \text{races where}\\ \text{weather's clear}}})$$

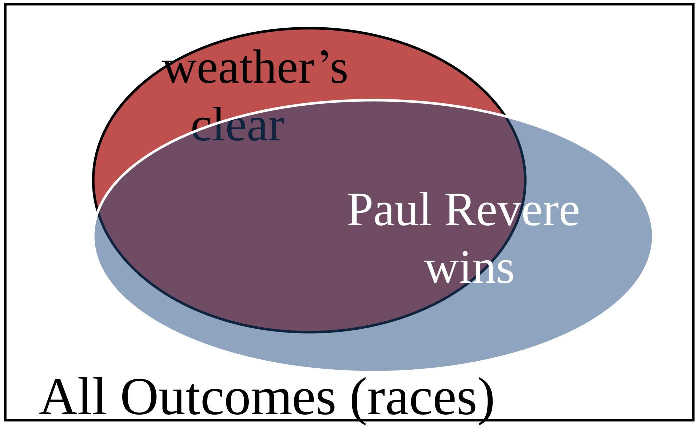


p measures total probability of a set of outcomes (an “event”)

Required Properties of p (axioms)

- $p(\emptyset) = 0$ $\qquad$ $p(\text{all outcomes}) = 1$
- $p(X) \leq p(Y)$ for any $X \subseteq Y$
- $p(X) + p(Y) = p(X \cup Y)$ provided $X \cap Y = \emptyset$

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



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 Left 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

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

If this prob is, 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 outcomes will be not races but character sequences $(x_1, x_2, x_3, \ldots)$ 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.`"

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

Apply the Chain Rule

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

$= p(x_1=\texttt{h})$ 4470/ 52108

$* \; p(x_2=\texttt{o} \mid x_1=\texttt{h})$ 395/ 4470

$* \; p(x_3=\texttt{r} \mid x_1=\texttt{h}, x_2=\texttt{o})$ 5/ 395

$* \; p(x_4=\texttt{s} \mid x_1=\texttt{h}, x_2=\texttt{o}, x_3=\texttt{r})$ 3/ 5

$* \; p(x_5=\texttt{e} \mid x_1=\texttt{h}, x_2=\texttt{o}, x_3=\texttt{r}, x_4=\texttt{s})$ 3/ 3

$* \; p(x_6=\texttt{s} \mid x_1=\texttt{h}, x_2=\texttt{o}, x_3=\texttt{r}, x_4=\texttt{s}, x_5=\texttt{e})$ 0/ 3

$* \; \ldots \quad = 0$

counts from
Brown corpus

Back Off On Right Side

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

$\approx p(x_1=\texttt{h})$ — 4470/ 52108

$* \; p(x_2=\texttt{o} \mid x_1=\texttt{h})$ — 395/ 4470

$* \; p(x_3=\texttt{r} \mid x_1=\texttt{h}, x_2=\texttt{o})$ — 5/ 395

$* \; p(x_4=\texttt{s} \mid x_2=\texttt{o}, x_3=\texttt{r})$ — 12/ 919

$* \; p(x_5=\texttt{e} \mid x_3=\texttt{r}, x_4=\texttt{s})$ — 12/ 126

$* \; p(x_6=\texttt{s} \mid x_4=\texttt{s}, x_5=\texttt{e})$ — 3/ 485

$* \; \ldots \; = 7.3\text{e-}10 * \ldots$

counts from Brown corpus

Change the Notation

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

$\approx p(x_1=\texttt{h})$	4470/ 52108
$* \; p(x_2=\texttt{o} \mid x_1=\texttt{h})$	395/ 4470
$* \; p(x_i=\texttt{r} \mid x_{i-2}=\texttt{h}, x_{i-1}=\texttt{o}, i=\texttt{3})$	5/ 395
$* \; p(x_i=\texttt{s} \mid x_{i-2}=\texttt{o}, x_{i-1}=\texttt{r}, i=\texttt{4})$	12/ 919
$* \; p(x_i=\texttt{e} \mid x_{i-2}=\texttt{r}, x_{i-1}=\texttt{s}, i=\texttt{5})$	12/ 126
$* \; p(x_i=\texttt{s} \mid x_{i-2}=\texttt{s}, x_{i-1}=\texttt{e}, i=\texttt{6})$	3/ 485
$* \; \ldots = 7.3\text{e-}10 * \ldots$	

counts from Brown corpus

Another Independence Assumption

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

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

$* \; p(x_2=\texttt{o} \mid x_1=\texttt{h})$ 395/ 4470

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

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

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

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

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

counts from Brown corpus

Simplify the Notation

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

$\approx p(x_1=\mathtt{h})$ — 4470/ 52108

$* \ p(x_2=\mathtt{o} \mid x_1=\mathtt{h})$ — 395/ 4470

$* \ p(\mathtt{r} \mid \mathtt{h}, \mathtt{o})$ — 1417/ 14765

$* \ p(\mathtt{s} \mid \mathtt{o}, \mathtt{r})$ — 1573/ 26412

$* \ p(\mathtt{e} \mid \mathtt{r}, \mathtt{s})$ — 1610/ 12253

$* \ p(\mathtt{s} \mid \mathtt{s}, \mathtt{e})$ — 2044/ 21250

$* \ \ldots$

counts from
Brown corpus

Simplify the Notation

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

$\approx p(\texttt{h} \mid \text{BOS}, \text{BOS})$

$* \; p(\texttt{o} \mid \text{BOS}, \texttt{h})$

$* \; p(\texttt{r} \mid \texttt{h}, \texttt{o})$

$* \; p(\texttt{s} \mid \texttt{o}, \texttt{r})$

$* \; p(\texttt{e} \mid \texttt{r}, \texttt{s})$

$* \; p(\texttt{s} \mid \texttt{s}, \texttt{e})$

$* \; \ldots$

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

counts from Brown corpus

These basic probabilities are used to define p(horses)

Simplify the Notation

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

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

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

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

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

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

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

$* \ \ldots$

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

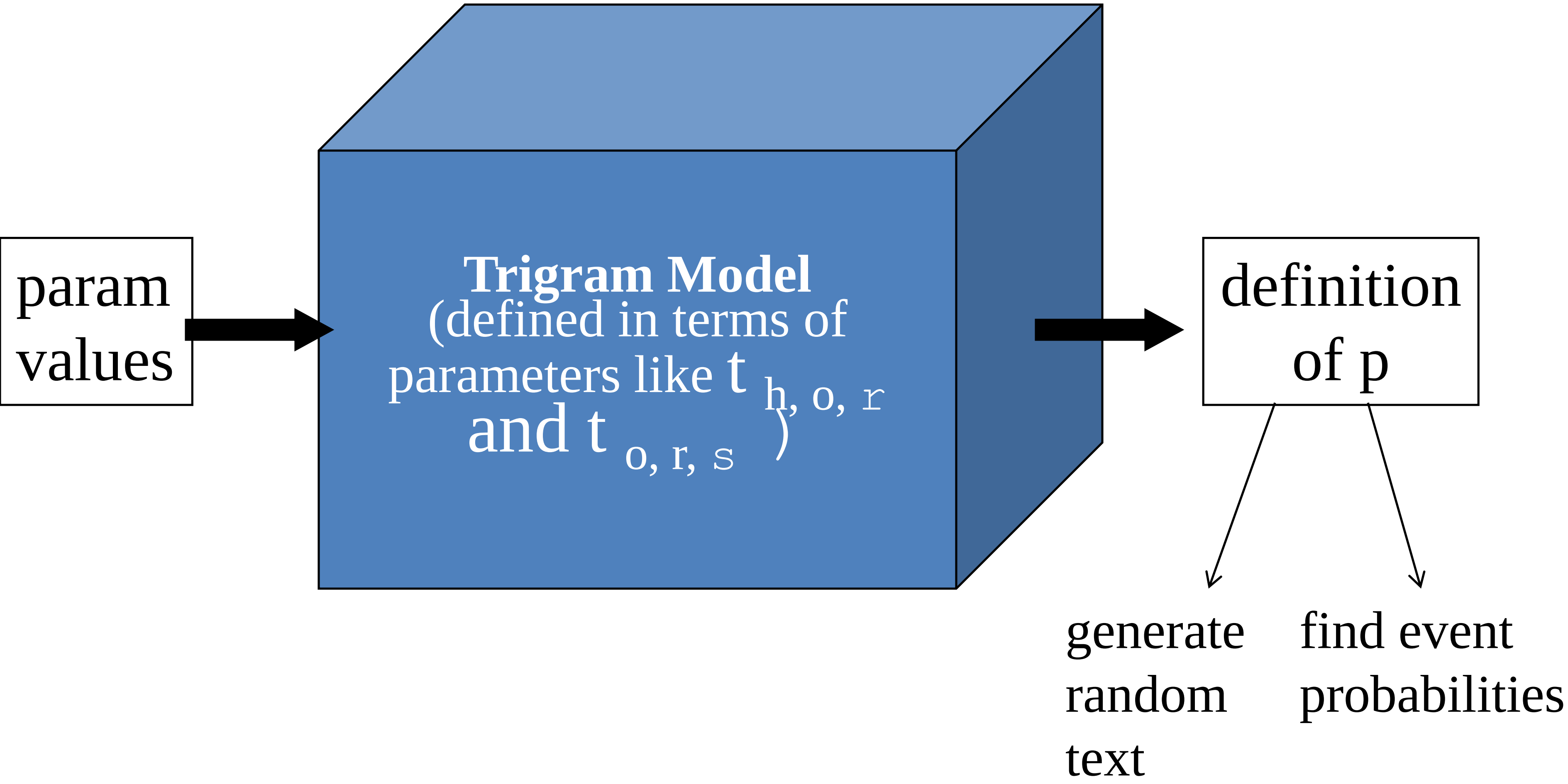
1610/ 12253

2044/ 21250

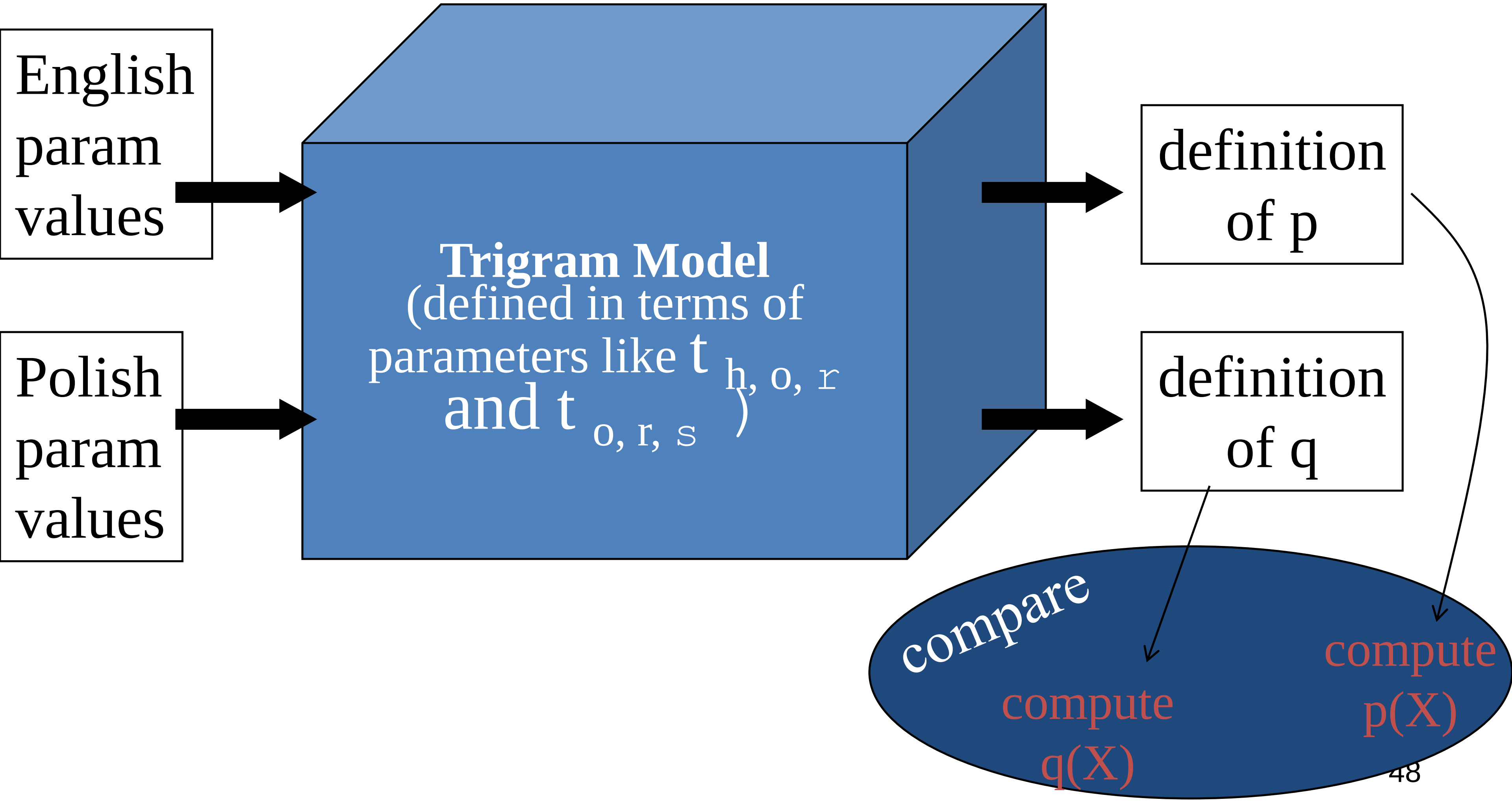
counts from Brown corpus

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

Definition: Probability Model

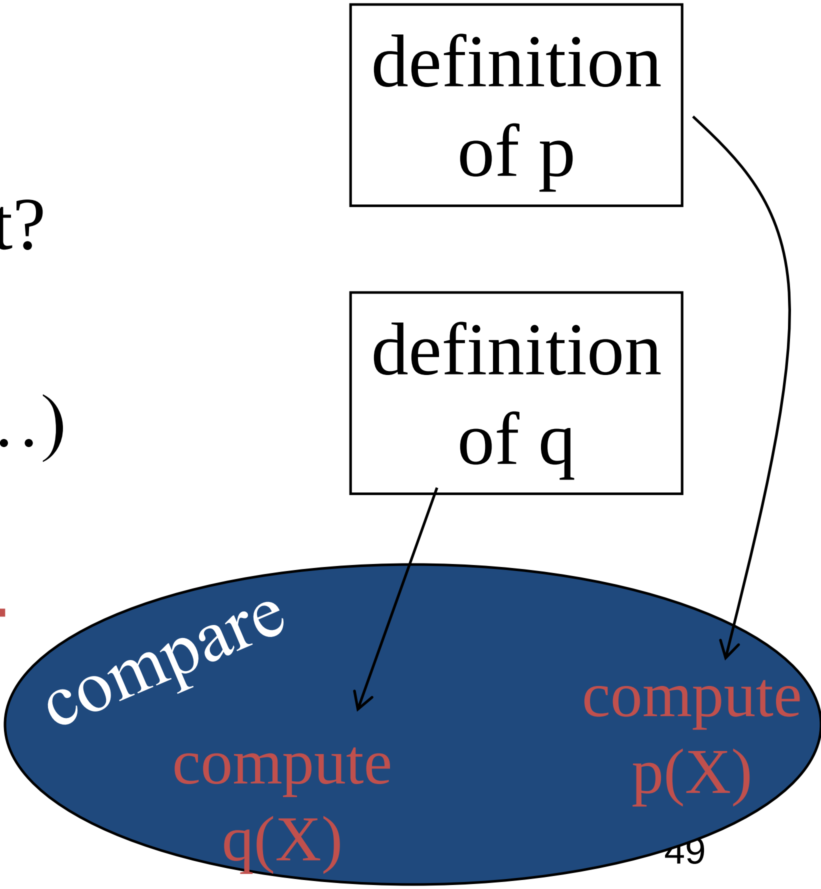


English vs. Polish



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)
    = p(sentence 1, sentence 2, …)
    = p(sentence 1)
    * p(sentence 2 | ~~sentence 1)~~
    * …



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(`horses`)

- But we rewrote p(`horses`) as

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

  $\approx p(x_1=\texttt{h}) * p(x_2=\texttt{o} \mid x_1=\texttt{h}) * \ldots$

- What does this variable=`value` notation mean?

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

Answer: variable is really a function of Outcome

- $p(x_1=\mathtt{h}) * p(x_2=\mathtt{o} \mid x_1=\mathtt{h}) * \ldots$
  - Outcome is a sequence of letters
  - x_2 is the second letter in the sequence
- p(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(weather's clear=`true`) or just p(weather's clear)
  - Outcome is a race
  - weather's clear is true or false

compare

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

Answer: variable is really a function of Outcome

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

compare

Random Variables:
What is “variable” in “p(variable=`value`)”?

- p(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(Outcome)=2) picks out *set* of outcomes w/2 heads
    = p({HHT,HTH,THH})
    = p(HHT)+p(HTH)+p(THH)

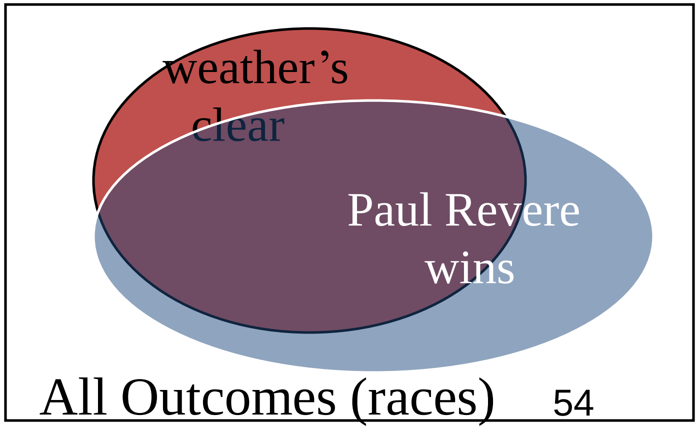
TTT	TTH	HTT	HTH
THT	THH	HHT	HHH

All Outcomes

Random Variables: What is “variable” in “p(variable=`value`)”?

- p(weather’s clear)
  - Outcome is a race
  - weather’s clear is true or false of the outcome
  - So p(weather’s clear) = p(weather’s clear(Outcome)=`true`)

picks out the *set* of outcomes with clear weather



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

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

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

    $= p(x_2(\text{Outcome})=\mathtt{o})$ picks out *set* of outcomes with …

    $= \Sigma$ p(Outcome) over all outcomes whose second letter …

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

Back to trigram model of p(horses)

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

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

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

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

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

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

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

$* \; \ldots$

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

$p(W_1 =$ `horses`$)$

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

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

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

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.

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

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

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

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

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

$\qquad + p(W_i=\texttt{horses} \mid T_i = \texttt{Verb}) * p(T_1 = \texttt{Verb}) + \ldots$

$= (72 / 55912) * (977 / 52108) + (\underline{0} / 15258) * (146 / 52108) + \ldots$

$= 2.4e\text{-}5 + 0 + \ldots + 0 \qquad = 2.4e\text{-}5$

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
  - 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

Cross-Entropy

- 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

- 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 …

Average?
Geometric average
of $1/2^3, 1/2^3, 1/2^3, 1/2^4$
$= 1/2^{3.25} \approx 1/9.5$

- high prob → low xent by 3 cosmetic improvements:
  - Take logarithm (base 2) to prevent underflow:

    log (1/8 * 1/8 * 1/8 * 1/16 …)
    = 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 → 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)