# Bayes' Theorem

# *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)

Let's revisit this

`"Horses and Lukasiewicz are on the curriculum."`

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

# Bayes' Theorem

- p(A | B) = p(B | A) * p(A) / p(B)

- Easy to check by removing syntactic sugar
- Use 1: Converts p(B | A) to p(A | B)
- Use 2: Updates p(A) to p(A | B)

- Stare at it so you'll recognize it later

# Language ID

- Given a sentence x, I suggested comparing its prob in different languages:
  - p(SENT=x | LANG=english) (i.e., $p_{english}$(SENT=x))
  - p(SENT=x | LANG=polish) (i.e., $p_{polish}$(SENT=x))
  - p(SENT=x | LANG=xhosa) (i.e., $p_{xhosa}$(SENT=x))
- But surely for language ID we should compare
  - p(LANG=english | SENT=x)
  - p(LANG=polish | SENT=x)
  - p(LANG=xhosa | SENT=x)

# Language ID

- For language ID we should compare
  - p(LANG=english | SENT=x)
  - p(LANG=polish | SENT=x)  *a posteriori*
  - p(LANG=xhosa | SENT=x)
- For ease, multiply by p(SENT=x) and compare
  - p(LANG=english, SENT=x)
  - p(LANG=polish, SENT=x)
  - p(LANG=xhosa, SENT=x)

  sum of these is a way to find p(SENT=x); can divide back by that to get posterior probs
- Must know prior probabilities; then rewrite as
  - p(LANG=english) * p(SENT=x | LANG=english)
  - p(LANG=polish) * p(SENT=x | LANG=polish)
  - p(LANG=xhosa) * p(SENT=x | LANG=xhosa)

  *a priori* likelihood (what we had before)

# Let's try it!

"First we pick a random LANG,
then we roll a random SENT
with the LANG dice."

| | | prior prob | | likelihood | | |
|---|---|---|---|---|---|---|
| best | 0.7 | p(LANG=english) | * | p(SENT=x \| LANG=english) | 0.00001 | |
| | 0.2 | p(LANG=polish) | * | p(SENT=x \| LANG=polish) | 0.00004 | |
| | 0.1 | p(LANG=xhosa) | * | p(SENT=x \| LANG=xhosa) | 0.00005 | best |

↑ from a very simple model: a single die whose sides are the languages of the world

↑ from a set of trigram dice (actually 3 sets, one per language)

| | joint probability | | |
|---|---|---|---|
| = | p(LANG=english, SENT=x) | 0.000007 | |
| = | p(LANG=polish, SENT=x) | 0.000008 | best compromise |
| = | p(LANG=xhosa, SENT=x) | 0.000005 | |

| probability of evidence | | |
|---|---|---|
| p(SENT=x) | 0.000020 | total over all ways of getting SENT=x |

# Let's try it!

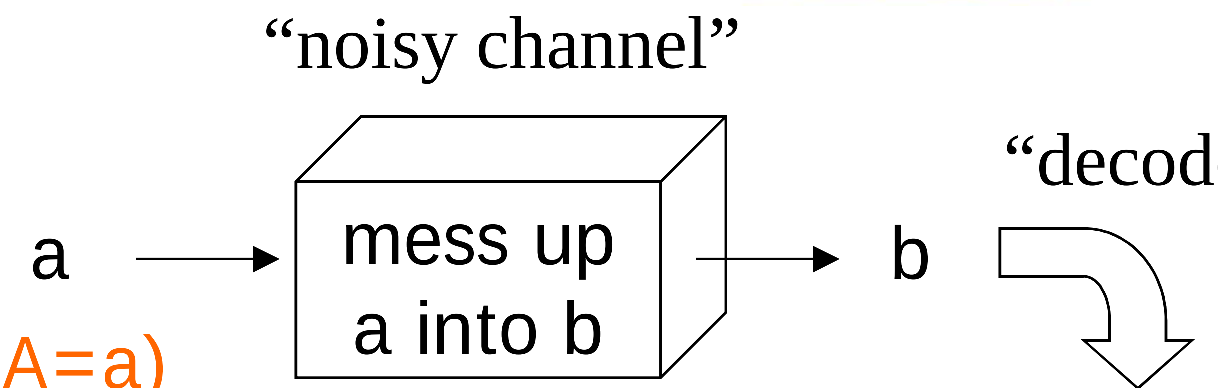
"First we pick a random LANG,
then we roll a random SENT
with the LANG dice."

| | | | | |
|---|---|---|---|---|
| ... | = | p(LANG=english, SENT=x) | 0.000007 | |
| | = | p(LANG=polish, SENT=x) | 0.000008 | best compromise |
| | = | p(LANG=xhosa, SENT=x) | 0.000005 | |
| | | joint probability | | |
| add up | | p(SENT=x) | 0.000020 | total probability of getting SENT=x one way or another! |
| | | probability of evidence | | |

| | | | |
|---|---|---|---|
| normalize (divide by a constant so they'll sum to 1) | p(LANG=english \| SENT=x) | 0.000007/0.000020 = 7/20 | |
| | p(LANG=polish \| SENT=x) | 0.000008/0.000020 = 8/20 | best |
| | p(LANG=xhosa \| SENT=x) | 0.000005/0.000020 = 5/20 | |
| | posterior probability | | |

given the evidence SENT=x,
the possible languages sum to 1

# General Case ("noisy channel")




$p(A=a)$

$p(B=b \mid A=a)$

most likely
reconstruction of a

language → text
text → speech
spelled → misspelled
English → French

maximize $p(A=a \mid B=b)$

$= p(A=a)\ p(B=b \mid A=a)\ /\ (B=b)$

$= p(A=a)\ p(B=b \mid A=a)$

$/\ \Sigma_{a'}\ p(A=a')\ p(B=b \mid A=a')$

# Language ID

- For language ID we should compare
  - p(LANG=english | SENT=x)
  - p(LANG=polish | SENT=x)
  - p(LANG=xhosa | SENT=x)

  *a posteriori*
- For ease, multiply by p(SENT=x) and compare
  - p(LANG=english, SENT=x)
  - p(LANG=polish, SENT=x)
  - p(LANG=xhosa, SENT=x)
- which we find as follows (we need prior probs!):
  - p(LANG=english) * p(SENT=x | LANG=english)
  - p(LANG=polish) * p(SENT=x | LANG=polish)
  - p(LANG=xhosa) * p(SENT=x | LANG=xhosa)

  *a priori* likelihood

# General Case ("noisy channel")

- Want most likely A to have generated evidence B
  - p(A = a1 | B = b)
  - p(A = a2 | B = b)
  - p(A = a3 | B = b)

  *a posteriori*
- For ease, multiply by p(B=b) and compare
  - p(A = a1, B = b)
  - p(A = a2, B = b)
  - p(A = a3, B = b)
- which we find as follows (we need prior probs!):
  - p(A = a1) * p(B = b | A = a1)
  - p(A = a2) * p(B = b | A = a2)
  - p(A = a3) * p(B = b | A = a3)

  *a priori* likelihood

# Speech Recognition

- For baby speech recognition we should compare
  - p(MEANING=gimme | SOUND=uhh)
  - p(MEANING=changeme | SOUND=uhh) *a posteriori*
  - p(MEANING=loveme | SOUND=uhh)
- For ease, multiply by p(SOUND=uhh) & compare
  - p(MEANING=gimme, SOUND=uhh)
  - p(MEANING=changeme, SOUND=uhh)
  - p(MEANING=loveme, SOUND=uhh)
- which we find as follows (we need prior probs!):
  - p(MEAN=gimme) * p(SOUND=uhh | MEAN=gimme)
  - p(MEAN=changeme) * p(SOUND=uhh | MEAN=changeme)
  - p(MEAN=loveme) * p(SOUND=uhh | MEAN=loveme)

*a priori* likelihood

# A simpler view? Odds Ratios

- What A values are probable, given that B=b?
- Bayes' Theorem says:
  - p(A=a1 | B=b) = p(A=a1) * p(B=b | A=a1) / p(B=b)
  - p(A=a2 | B=b) = p(A=a2) * p(B=b | A=a2) / p(B=b)
- Therefore
  - $$\frac{p(A=a1 \mid B=b)}{p(A=a2 \mid B=b)} = \frac{p(A=a1)}{p(A=a2)} * \frac{p(B=b \mid A=a1)}{p(B=b \mid A=a2)}$$

posterior odds ratio = prior odds ratio * likelihood ratio

# A simpler view? Odds Ratios

- $$\frac{p(A=a1 \mid B=b)}{p(A=a2 \mid B=b)} = \frac{p(A=a1)}{p(A=a2)} * \frac{p(B=b \mid A=a1)}{p(B=b \mid A=a2)}$$

posterior odds ratio | prior odds ratio | likelihood odds ratio

| | prior | likelihood | |
|---|---|---|---|
| 0.7 | p(LANG=english) | p(SENT=x \| LANG=english) | 0.00001 |
| 0.2 | p(LANG=polish) | p(SENT=x \| LANG=polish) | 0.00004 |
| 0.1 | p(LANG=xhosa) | p(SENT=x \| LANG=xhosa) | 0.00005 |

- A priori, English is 7 times as probable as Xhosa (7:1 odds)
- But likelihood of English is only 1/5 as large (1:5 odds)
- So a posteriori, English now 7 * 1/5 = 1.2 times as probable (7:5 odds)
  - That is: p(English) = 7/12, p(Xhosa) = 5/12 if no other options

# Growing evidence eventually overwhelms the prior

- We were expecting Polish text but actually it's English
- What happens as we read more & more words?
  - The prior odds ratio stays the same
  - But the likelihood odds ratio becomes extreme (much bigger or much smaller than 1, depending on which hypothesis is correct)
    - Suppose each trigram is 1.001 times more probable under the English model than the Polish model
    - Then after 700 trigrams, the likelihood ratio is > 2 in favor of English ($1.001^{700} > 2$)
    - And after 7000 trigrams, the likelihood ratio is > $2^{10}$ in favor of English!
  - As long as the prior p(English) > 0, eventually we come to believe it's English a posteriori. We get surer and surer with more evidence.

# Life or Death!

Does Epitaph have hoof-and-mouth disease?
He tested positive – oh no!
False positive rate only 5%

- $p(\text{hoof}) = 0.001$ so $p(\neg\text{hoof}) = 0.999$

---

- $p(\text{positive test} \mid \neg\text{hoof}) = 0.05$ "false pos"
- $p(\text{negative test} \mid \text{hoof}) = \varepsilon \geq 0$ "false neg"

  so $p(\text{positive test} \mid \text{hoof}) = 1-\varepsilon$

---

- What is p(hoof | positive test)?
  - Consider the hoof: ¬hoof odds ratio
  - Prior odds ratio 1:999 (improbable!)
  - Likelihood ratio at most 1:0.05, or equivalently 20:1
  - So posterior odds ratio at most 20:999, or about 1:50
    - That is, p(hoof | positive test) at most about 1/51