More Statistical Inference

Review

- Let event D = data we have observed.
- Let events $H_1, \ldots, H_k$ be events representing hypotheses we want to choose between.
- Use D to pick the "best" H.

- There are two "standard" ways to do this, depending on what information we have available.

Maximum likelihood hypothesis

- The maximum likelihood hypothesis (H^{ML}) is the hypothesis that maximizes the probability of the data given that hypothesis.

$$H^{\mathrm{ML}} = \underset{i}{\operatorname{argmax}}\, P(D \mid H_i)$$

- How to use it: compute $P(D \mid H_i)$ for each hypothesis and select the one with the greatest value.

Maximum a posteriori (MAP) hypothesis

- The MAP hypothesis is the hypothesis that maximizes the posterior probability:

Bayes

$$
\begin{aligned}
H^{\mathrm{MAP}} &= \operatorname*{argmax}_i P(H_i \mid D) \\
&= \operatorname*{argmax}_i \frac{P(D \mid H_i) P(H_i)}{P(D)} \\
&\propto \operatorname*{argmax}_i P(D \mid H_i) \boxed{P(H_i)}
\end{aligned}
$$

prior probability

- The $P(D \mid H_i)$ terms are now *weighted* by the hypothesis prior probabilities.

posterior prob. $P(H_i \mid D)$

One slide to rule them all

- The maximum likelihood hypothesis is the hypothesis that maximizes the probability of the observed data:

$$H^{\mathrm{ML}} = \operatorname*{argmax}_i P(D \mid H_i)$$

- The MAP hypothesis is the hypothesis that maximizes the posterior probability given D:

$$H^{\mathrm{MAP}} = \operatorname*{argmax}_i P(D \mid H_i) P(H_i)$$

- $P(H_i)$ is called the prior probability (or just prior).
- $P(H_i | D)$ is called the posterior probability.

- There are 3 robots.
- Robot 1 will hand you a snack drawn at random from 2 doughnuts and 7 carrots.
- Robot 2 will hand you a snack drawn at random from 4 apples and 3 carrots.
- Robot 3 will hand you a snack drawn at random from 7 burgers and 7 carrots.
- Suppose your friend goes up to a robot (you don't see this happen) and is given a carrot. Which robot did your friend probably approach?
- What if the prior probability of your friend approaching robots 1, 2, and 3 are 20%, 40%, and 40%, respectively?

- There are 3 robots.
- Robot 1 will hand you a snack drawn at random from 2 doughnuts and 7 carrots.
- Robot 2 will hand you a snack drawn at random from 4 apples and 3 carrots.
- Robot 3 will hand you a snack drawn at random from 7 burgers and 7 carrots.
- Suppose your friend goes up to a robot (you don't see this happen) and is given a carrot. Which robot did your friend probably approach?
- What if the prior probability of your friend approaching robots 1, 2, and 3 are 20%, 40%, and 40%, respectively?

Data / Hyp → R_1, R_2, R_3

↓

carrot

H^{ML} / H^{MAP}

$H^{ML} = \arg\max_i P(D|H_i)$

$H^{ML} = $ Robot 1

$P(D \mid H_i)$	$P(H_i)$
$P(C \mid R_1) = 7/9$	$P(R_1) = 0.2$
$P(C \mid R_2) = 3/7$	$P(R_2) = 0.4$
$P(C \mid R_3) = 1/2$	$P(R_3) = 0.4$

$H^{MAP} = \arg\max_i P(D|H_i)P(H_i)$

$= $ Robot 3

R_1	R_2	R_3
$(\frac{7}{9})(.2)$	$(\frac{3}{7})(.4)$	$(\frac{1}{2})(.4)$
$.1555$	$.1714..$	$\boxed{.2}$

Probability vs hypothesis

- Sometimes you only care about which hypothesis is more likely, and sometimes you need the actual probability.

$$P(H_i|D) = \frac{P(D|H_i)P(H_i)}{P(D)}$$

marginalization

$$= \frac{P(D \mid H_i)P(H_i)}{\sum_j P(D, H_j)}$$

def of cond. prob.

$$= \frac{P(D \mid H_i)P(H_i)}{\sum_j P(D \mid H_j)P(H_j)}$$

Formalization of Normalization

$$P(H_i|D) = \frac{P(D|H_i)P(H_i)}{P(D)} = \frac{P(D \mid H_i)P(H_i)}{\sum_j P(D \mid H_j)P(H_j)}$$

- In the robot problem, what is P(R3 | C)?

$$\frac{\frac{7}{9} \times .2 \quad \text{vs} \quad \frac{3}{7} \times .4 \quad \text{vs} \quad \frac{1}{2} \times .4}{.155 \qquad\qquad .1714 \qquad\qquad .2}$$

$$P(R_3|C) = \frac{.2}{.2 + .155 + .1714}$$

Probability vs hypothesis

- In the robot problem, what is P(R3 | C)?

$$P(R_3|C) = \frac{P(C|R_3)P(R_3)}{P(C)}$$

$$P(R_3|C) = \frac{P(C|R_3)P(R_3)}{\sum_{i=1}^{3} P(C, R_i)}$$

$$P(R_3|C) = \frac{P(C|R_3)P(R_3)}{\sum_{i=1}^{3} P(C|R_i)P(R_i)}$$

= (7/9 * 2/10) / (7/9 * 2/10 + 3/7 * 4/10 + 1/2 * 4/10) =~ 0.2952

- Suppose I work in FJ in a windowless office. I want to know whether it's raining outside. The chance of rain is 70%. My colleague walks in wearing his raincoat. If it’s raining, there’s a 65% chance he’ll be wearing a raincoat. Since he’s very unfashionable, there’s a 45% chance he’ll be wearing his raincoat even if it’s not raining. My other colleague walks in with wet hair. When it’s raining there’s a 90% chance her hair will be wet. However, since she sometimes goes to the gym before work, there’s a 40% chance her hair will be wet even if it’s not raining.
- What’s the posterior probability that it’s raining?

- Suppose I work in FJ in a windowless office. I want to know whether it's raining outside. The chance of rain is 70%. My colleague walks in wearing his raincoat. If it's raining, there's a 65% chance he'll be wearing a raincoat. Since he's very unfashionable, there's a 45% chance he'll be wearing his raincoat even if it's not raining. My other colleague walks in with wet hair. When it's raining there's a 90% chance her hair will be wet. However, since she sometimes goes to the gym before work, there's a 40% chance her hair will be wet even if it's not raining.
- What's the posterior probability that it's raining?

Hyps = RAIN, ¬RAIN

data = wearing a raincoat

wet hair

$$P(\text{Rain} \mid \text{Coat}, \text{WetHair}) = \frac{P(\text{Coat}, \text{WetHair} \mid \text{Rain})\, P(\text{Rain})}{P(\text{Coat}, \text{WetHair})}$$

$P(\text{Rain}) = 0.7$

$P(\neg\text{Rain}) = 0.3$

$P(\text{Coat} \mid \text{Rain}) = 0.65$

$P(\text{Coat} \mid \neg\text{Rain}) = 0.45$

$P(\text{Wet} \mid \text{Rain}) = 0.9$

$P(\text{Wet} \mid \neg\text{Rain}) = 0.4$

$P(\neg\text{Coat} \mid \text{Rain}) = 1 - 0.65 = .35$

$P(\text{Coat}, \text{WetHair} \mid \text{Rain})$???

$\neq P(\text{Coat} \mid \text{Rain})\, P(\text{WetHair} \mid \text{Rain})$

- We can't solve this problem because we don’t have any information about the probability of Colleague 1 wearing a raincoat and Colleague 2 having wet hair occurring *simultaneously*.
- We don't know P(C, W | R).
- Let's make an *assumption* that C and W are conditionally independent given that it is raining (or not raining).
- P(C, W | R) = P(C | R) * P(W | R)
  - (and similarly for given ~R)

Combining evidence

- It is very common to make this independence assumption for multiple pieces of evidence (data).

BAYES RULE

$$P(H_i \mid D_1, \ldots, D_m) = \frac{P(D_1, \ldots, D_m \mid H_i)P(H_i)}{P(D_1, \ldots, D_m)}$$

conditional indep assumption

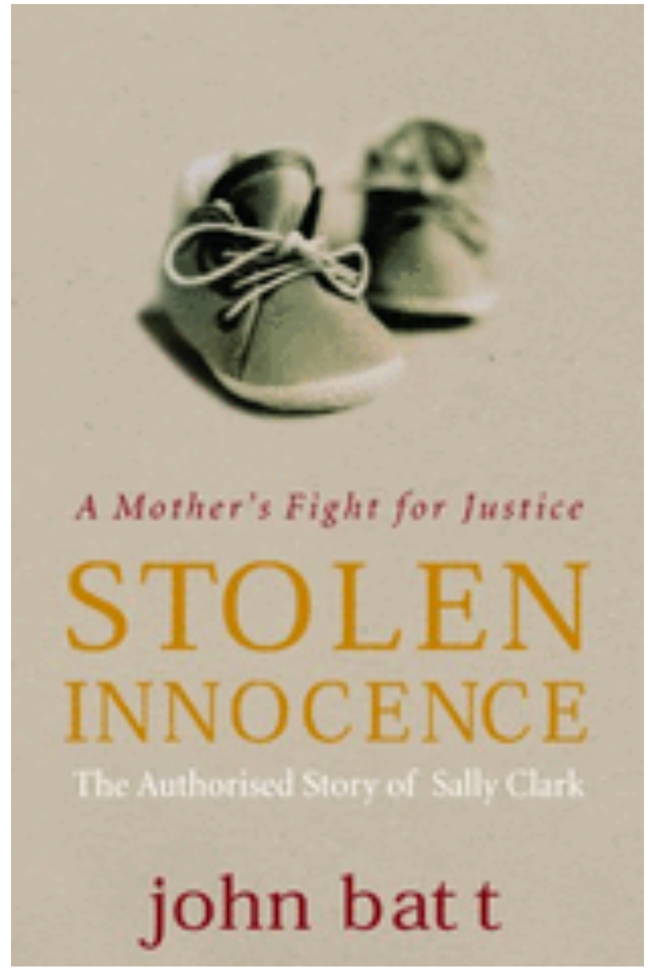
$$= \frac{\big(P(D_1 \mid H_i) \cdots P(D_m \mid H_i)\big)P(H_i)}{P(D_1, \ldots, D_m)}$$

$$= \frac{\big(\prod_{j=1}^{m} P(D_j \mid H_i)\big)P(H_i)}{P(D_1, \ldots, D_m)}$$

where $P(D_1 \ldots, D_m) = \sum_{i=1}^{k} \Big(\prod_{j=1}^{m} P(D_j \mid H_i)\Big) P(H_i)$ } Normalization

This can be dangerous!

Sally Clark



prob of SIDS $\approx \frac{1}{8500}$

prob of 2 SIDS deaths $= \left(\frac{1}{8500}\right)^2 = \frac{1}{73 \text{ million}}$

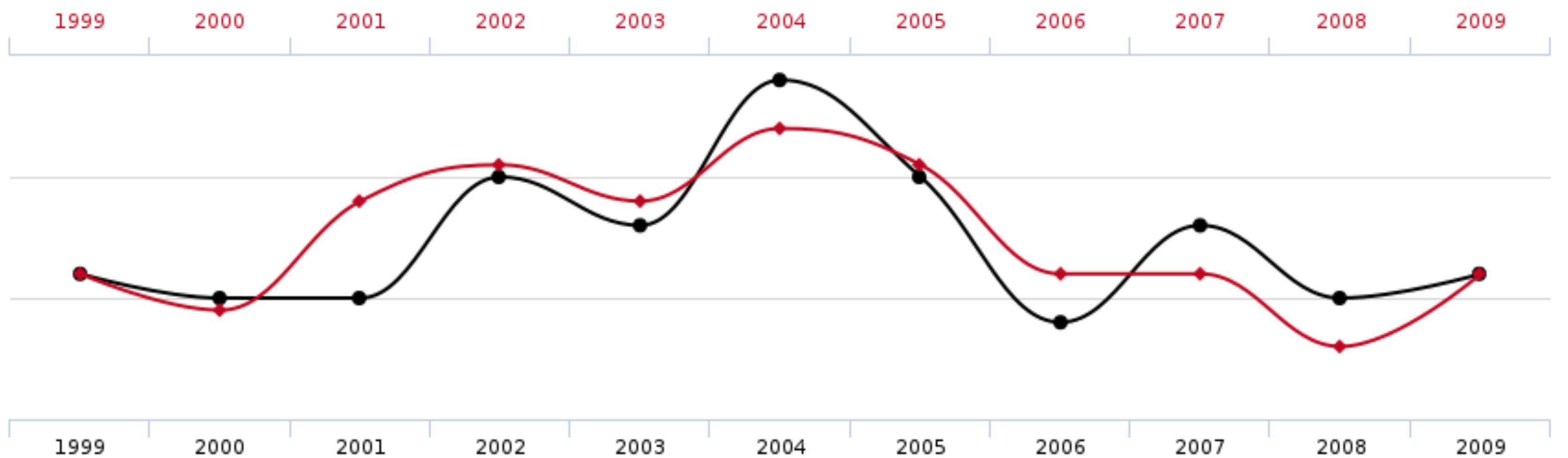
prob that Clark was innocent

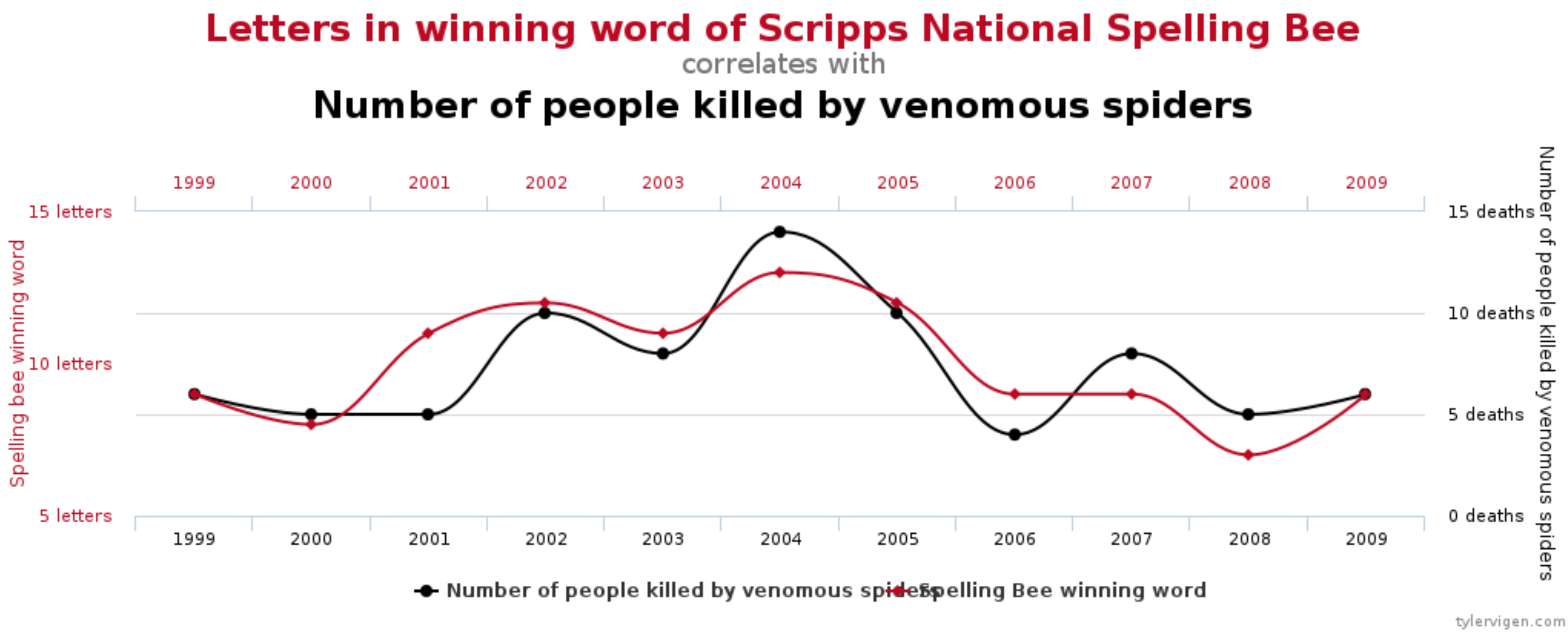
"Prosecutor's Fallacy"

prob of murder?

Hyps = SIDS / Murder

$\underline{P(D \mid SIDS)}\,\underbrace{P(SIDS)}_{\frac{1}{73 \text{ million}}}$ vs $\underline{P(D \mid Murder)}\,\underline{P(Murder)}$


1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009



Letters in winning word of Scripps National Spelling Bee
correlates with
Number of people killed by venomous spiders
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
15 letters
10 letters
5 letters
Spelling bee winning word
15 deaths
10 deaths
5 deaths
0 deaths
Number of people killed by venomous spiders
Number of people killed by venomous spiders
Spelling Bee winning word
tylervigen.com

Spam classification

- Suppose you have an email and you want to know if it's spam or not. hyps
- In general, the probability of an email being spam is 20%.
- Suppose you have a big list of words that "suggest" spam, like viagra, cialis, cash, ... data

$P(\text{viagra} \mid \text{spam}) P(\text{cialis} \mid \text{spam})$

$\neq P(\text{viagra}, \text{cialis} \mid \text{spam})$

Assume all our data is conditionally indep

- Two hypotheses: spam and not-spam.
- You know P(spam) and P(not-spam).
- Suppose your word list has *m* words in it.
- Our newly-observed email (our evidence/data) is the joint event $W_1, W_2, \ldots, W_m$ where each W_i is true or false if the word is in the email or not.
- Let's assume the words are all conditionally independent given the label (spam/not-spam), and that we can compute $P(W_i|\text{spam})$ and $P(W_i|\text{not-spam})$.

$$P(\text{spam} \mid W_1, \ldots, W_m) = \frac{P(W_1, \ldots, W_m \mid \text{spam})P(\text{spam})}{P(W_1, \ldots, W_m)}$$

$$= \frac{\big(P(W_1 \mid \text{spam}) \cdots P(W_m \mid \text{spam})P(\text{spam})\big)}{P(W_1, \ldots, W_m)}$$

$$= \frac{\big(\prod_{j=1}^{m} P(W_j \mid \text{spam})\big)P(\text{spam})\big)}{P(W_1, \ldots, W_m)}$$

The equation above is the basis for a probabilistic model called a *Naïve Bayes Classifier.*