More Statistical Inference

Review

- Let event D = data we have observed.
- Let events H_1, …, 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:

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

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

Posterior probability

- If you need the actual posterior probability:

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

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

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

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.

- A patient comes to visit Dr. Gregory House because they have a cough. After insulting and belittling the patient, House consults with his team of diagnosticians, who tell him that if a patient has a cold, then there's a 75% chance they will have a cough. But if a patient has the Ebola virus, there's a 80% chance they will have a cough.
- What is the maximum likelihood hypothesis for the diagnosis?

- After concluding the patient has Ebola, House fires all his diagnosticians for their poor hypothesis testing skills and hires new ones. This new team does some background research and discovers if they are only going to consider the common cold and Ebola, then before the symptoms are even considered, there’s a 1% chance the patient has Ebola and a 99% chance they have a cold.
- What is the MAP hypothesis for the diagnosis? What is the posterior probability the patient has Ebola?

- 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?

- 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 $P(C, W \mid R)$.
- Let's make an *assumption* that C and W are conditionally independent given that it is raining (or not raining).
- $P(C, W \mid R) = P(C \mid R) * P(W \mid R)$
  - (and similarly for given $\sim R$)

Combining evidence

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

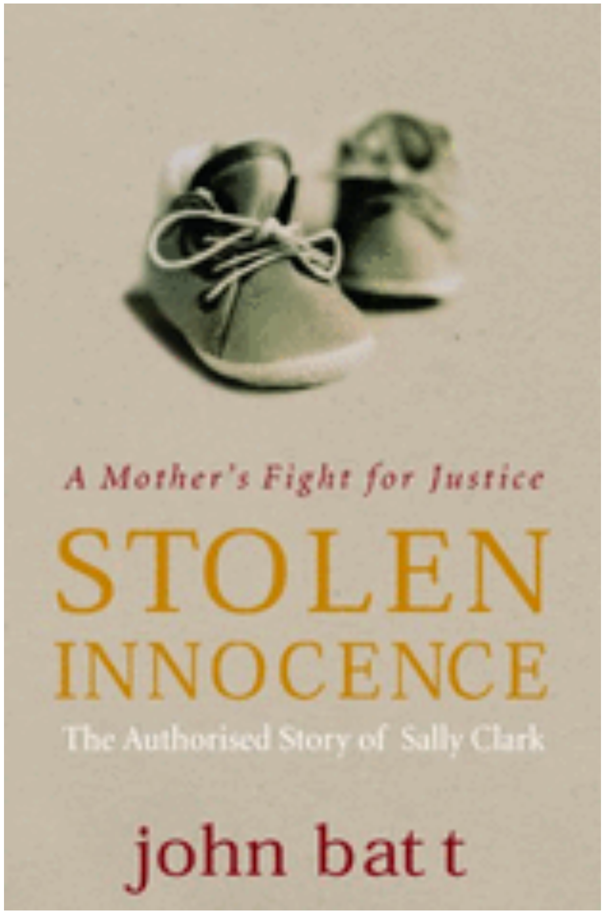
$$P(H_i \mid D_1, \ldots, D_m) = \frac{P(D_1, \ldots, D_m \mid H_i) P(H_i)}{P(D_1, \ldots, D_m)}$$

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

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

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

This can be dangerous!


A Mother's Fight for Justice
STOLEN
INNOCENCE
The Authorised Story of Sally Clark
john bat t

Spam classification

- Suppose you have an email and you want to know if it's spam or not.
- 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, ...
- You have access to a large number of old emails that are correctly categorized as spam or not-spam.
- How can you compute the probability that a new email is spam?

- 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, …, 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.*