TALENT COURSE 11
LEARNING FROM DATA: BAYESIAN METHODS AND MACHINE LEARNING

Lecture 3: Bayesian updating, Bayesian convergence, and The Lighthouse

Daniel Phillips
Ohio University
TU Darmstadt
ExtreMe Matter Institute

TALENT Course 11 is possible thanks to funding from the STFC

BAYES' THEOREM

Thomas Bayes (1701?-1761)

http://www.bayesian-inference.com

$$\mathrm{pr}(A \mid B, I) = \frac{\mathrm{pr}(B \mid A, I)\mathrm{pr}(A \mid I)}{\mathrm{pr}(B \mid I)}$$

Likelihood — Prior

$$\mathrm{pr}(\mathrm{model} \mid \mathrm{data}, I) = \frac{\mathrm{pr}(\mathrm{data} \mid \mathrm{model}, I)\mathrm{pr}(\mathrm{model} \mid I)}{\mathrm{pr}(\mathrm{data} \mid I)}$$

Posterior — Normalization

Probability as degree of belief cf. frequentist view

COIN TOSSING REVIEW

- Is this a fair coin?
- Deductive argument:

$$\text{Fair} \Rightarrow \text{pr}(\text{Heads}) = \text{pr}(\text{Tails}) = 0.5$$

$$\Rightarrow \text{pr}(R \text{ heads out of } N \text{ tosses} \,|\, \text{fair coin}) = \binom{N}{R} (0.5)^R (0.5)^{N-R}$$

- Is sum rule obeyed here?
- In general: $\text{pr}(R \text{ heads out of } N \text{ tosses} \,|\, p_H) = \binom{N}{R} p_H^R (1 - p_H)^{N-R}$
- So:

$$\text{pr}(p_H \,|\, R \text{ heads out of } N \text{ tosses}, I) \propto \binom{N}{R} p_H^R (1 - p_H)^{N-R} \text{pr}(p_H \,|\, I)$$

NORMALIZATION

Flat prior⇒

$$\mathrm{pr}(p_H \,|\, R \text{ heads out of } N \text{ tosses}, I) = \mathcal{N} p_H^R (1 - p_H)^{N-R}$$

But we want:

$$\int dp_H \,\mathrm{pr}(p_H \,|\, R \text{ heads out of } N \text{ tosses}, I) = 1$$

$$\Rightarrow \mathcal{N} \frac{\Gamma(1 + N - R)\,\Gamma(1 + R)}{\Gamma(2 + N)} = 1, \text{ p . v . } N > R - 1 \text{ and } R > -1$$

$$\Rightarrow \mathcal{N} = \frac{\Gamma(2 + N)}{\Gamma(1 + N - R)\,\Gamma(1 + R)}$$

FREQUENTIST ESTIMATORS

$$p_{H\mathrm{ML}} = \frac{R}{N}$$

$$\sigma_{\mathrm{ML}} = \sqrt{\frac{(R-N)R}{N^3}} = \sqrt{\frac{p_{H\mathrm{ML}}(1-p_{H\mathrm{ML}})}{N}}$$

$$L(p_H) = L(p_{H\mathrm{ML}})\exp\left[-\frac{1}{2}\left(\frac{p_H - p_{HML}}{\sigma}\right)^2 + \text{higher orders}\right]$$

THE JOYS OF CONJUGATE PRIORS

$$\mathrm{pr}(p_H|R \text{ heads out of } N \text{ tosses}, I) \propto \binom{N}{R} p_H^R (1-p_H)^{N-R} \mathrm{pr}(p_H|I)$$

- So pick

$$\mathrm{pr}(p_H|I) = \frac{\Gamma(\alpha+\beta+2)}{\Gamma(\alpha+1)\Gamma(\beta+1)} p_H^\alpha (1-p_H)^\beta$$

- To go from prior to posterior requires two steps:

  a) Update $\alpha \rightarrow R+\alpha$; $\beta \rightarrow N-R+\beta$;

  b) Adjust normalization.

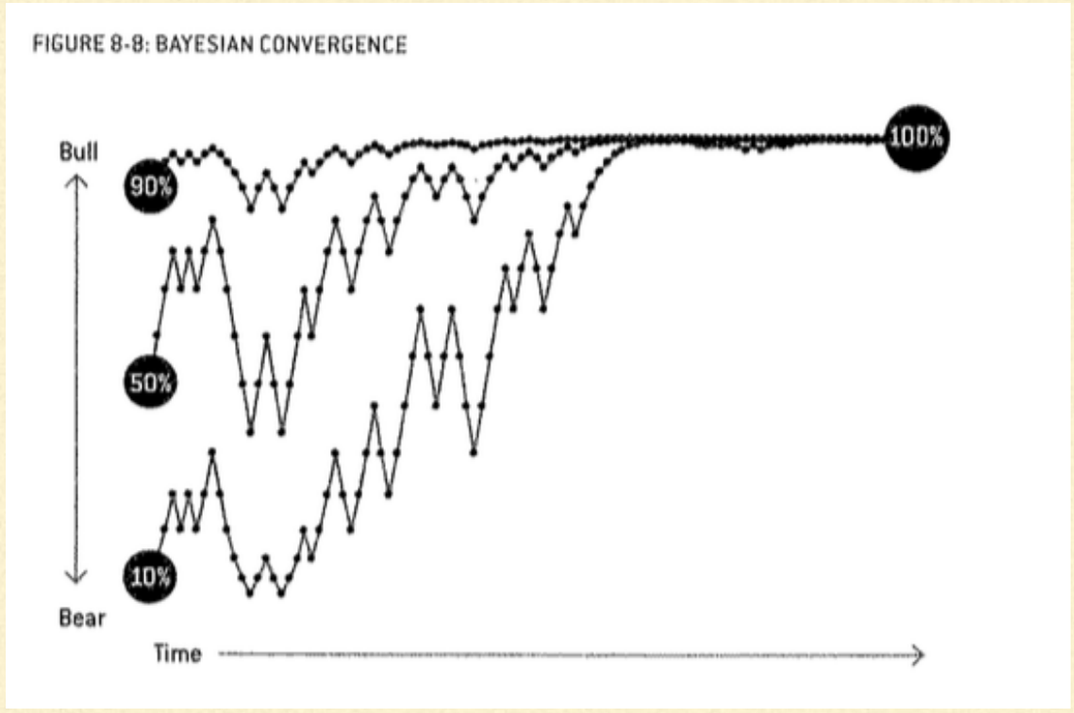
- No particular gain in this problem, but for problems where numerical integration is required this speeds things up markedly.

COIN TOSSING TAKE AWAYS

- pr(p_H|data,I) is the product of the binomial distribution and the prior
- Normalization set by sum rule/marginalization over p_H
- The frequentist result corresponds to a particular choice of prior
- For a conjugate prior: posterior is a beta function, just like the prior, but with different parameters. Fast updating!
- Can do analysis sequentially or all at once
- MUST NOT bootstrap

BAYESIAN CONVERGENCE

- In the short run, priors affect the strength of the conclusion that you draw
- BUT, you can apply Bayes' theorem over and over
- So, if you keep getting data, in the long run all priors lead to the same posterior
- "Long run" can take a while, depending on the accuracy you want



"Bayesian convergence"

Nate Silver: "The Signal and The Noise"

- 1000th coin toss has less impact on posterior than the first one, i.e., if you already have a lot of data, one more piece of information will shift your assessment less

BAYES' THEOREM FOR YOUR LIFE

pr(hypothesis|new data, other stuff you knew before) ∝
pr(new data|hypothesis) pr(hypothesis|other stuff you knew before)

What you learnt

Prior

- Degree of belief in a hypothesis is modified by new data
- But impact of data on belief in the hypothesis depends on previous degree of belief regarding hypothesis
- "Flat prior" versus prior of probability 1 or 0.

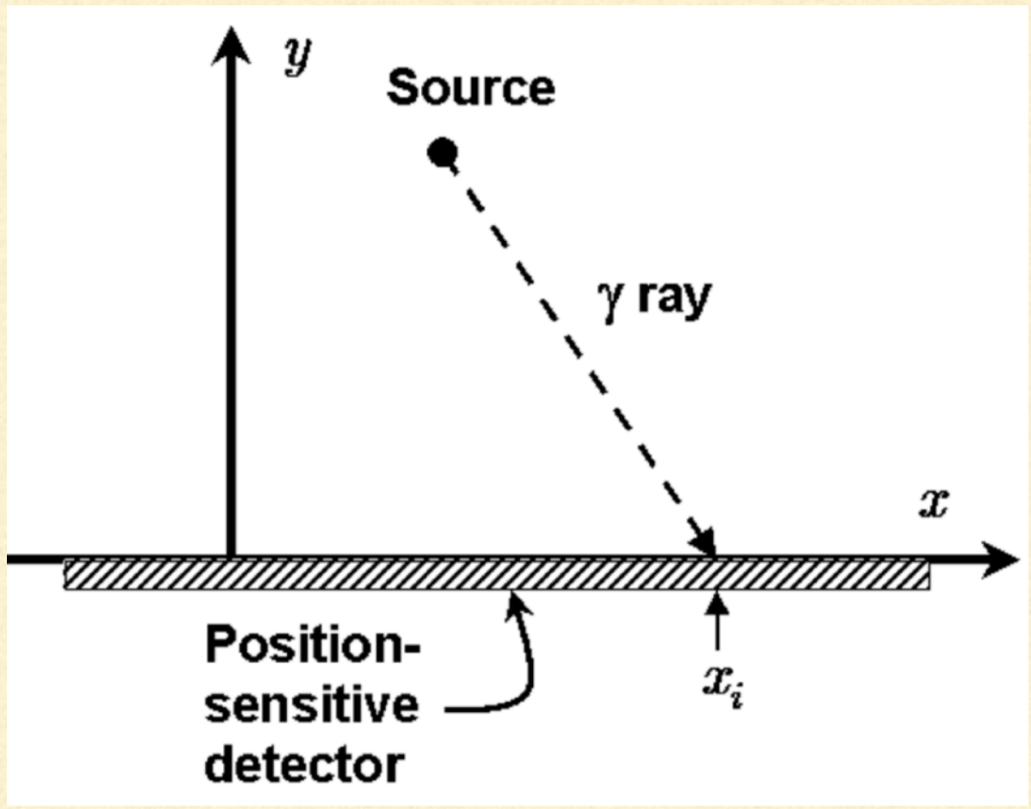
  Things I know nothing about vs. Things I will never change my mind about

- It's okay to have a prior
- It's not okay to not assess it honestly. It's not okay not to state it.

TO THE LIGHTHOUSE!

Gull, updated by Sivia

You are lost on a beach and it's dark. But there's a lighthouse, a distance β off shore. The lighthouse emits pulses of light at regular intervals. It is rotating (as lighthouses do) at constant angular velocity. The pulses are therefore received at different positions along the shore. You are lost and it's dark. You do have a map that shows the lighthouse is1 km away from the shore. You also just happen to have a huge, linear CCD with you. So you set it up the shore and record the positions where the lighthouse pulses hit the shore. Given a set of pulses $\{x_k:k=1,..,N\}$ what is your best estimate for the lighthouse's position, α, along the shore relative to you?



PRIOR FIRST

- pr(β|map, I'm on the shore)=?
- pr(α|β,I)=?

$$\mathrm{pr}(\beta \,|\, I) = \delta(\beta - 1 \text{ km})$$

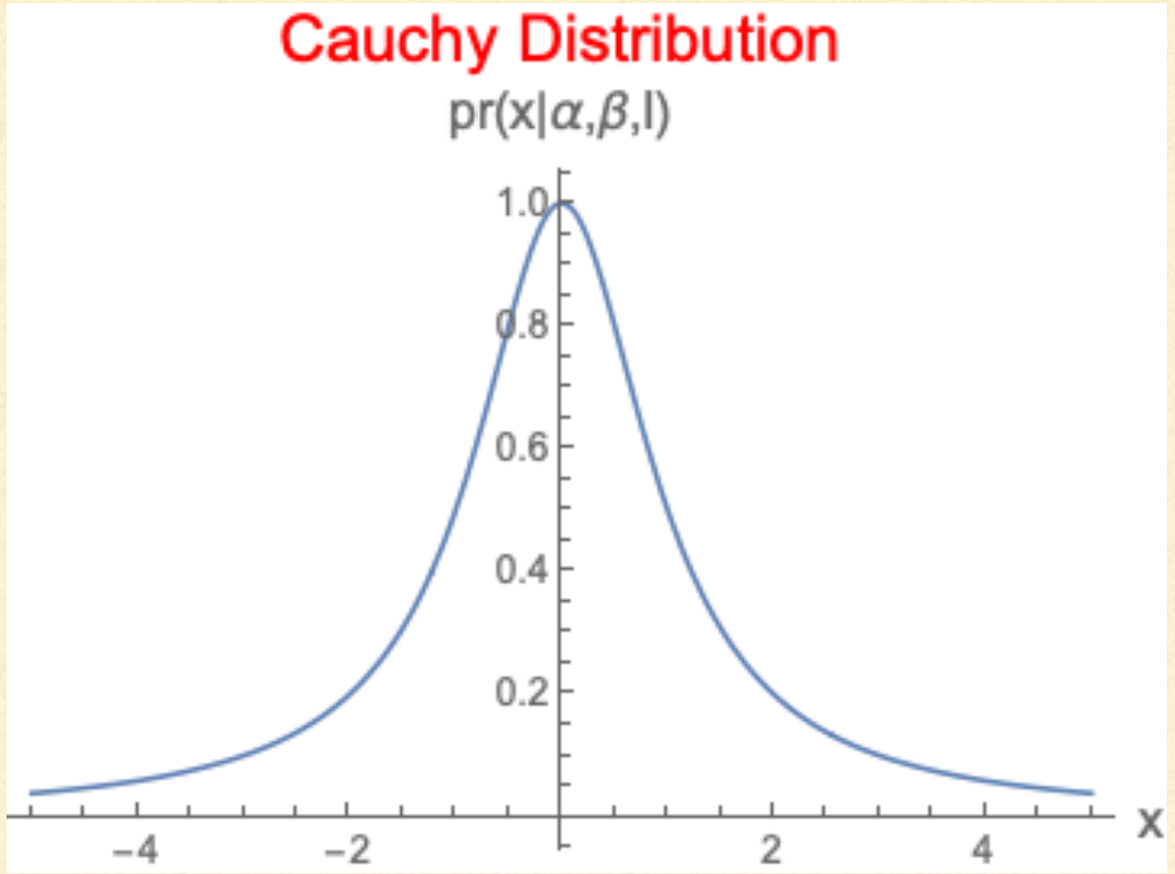
$$\mathrm{pr}(\alpha|\beta, I) = \mathrm{pr}(\alpha|I) = \begin{cases} A & \alpha_{\min} \leq \alpha \leq \alpha_{\max}, \\ 0 & \text{otherwise.} \end{cases}$$

What is joint pdf pr(α,β|I)?

What happens if I then integrate over β?

NOW THE LIKELIHOOD

- Bayes says: $pr(\alpha|\{x_k\},\beta,I) \propto pr(\{x_k\}|\alpha,\beta,I)pr(\alpha|\beta,I)$
- So can we figure out $pr(\{x_k\}|\alpha,\beta,I)$?
- How is it related to $pr(x_k|\alpha,\beta,I)$?
- What do we know about the x_k's?
- $pr(\theta_k|\alpha,\beta,I)=1/\pi$



$$\Rightarrow \mathrm{pr}(x_k \,|\, \alpha, \beta, I) = \frac{\beta}{\pi[\beta^2 + (x_k - \alpha)^2]}$$

LOG LIKELIHOOD

$$\mathrm{pr}(\{x_k\}\,|\,\alpha,\beta,I) = \prod_{k=1}^{N} \mathrm{pr}(x_k\,|\,\alpha,\beta,I) \propto \prod_{k=1}^{N} \frac{1}{\beta^2 + (x_k-\alpha)^2}$$

Let's take a log

$$L = \log[\mathrm{pr}(\alpha\,|\,\{x_k\},\beta,I)] = \mathrm{constant} - \sum_{k=1}^{N} \log[\beta^2 + (x_k-\alpha)^2]$$

$$\left.\frac{dL}{d\alpha}\right|_{\alpha=\alpha_0} = 2\sum_{k=1}^{N} \frac{x_k-\alpha_0}{\beta^2 + (x_k-\alpha_0)^2} = 0.$$

Not analytically soluble

So calculate pr(α|{x_k},β,I) numerically