8/28/19

Physics 8805: Learning from Data: Lecture 3

Before class: Set up projector with laptop and notebooks:
- medical example key
- flipping coins
- exploring pdfs
- Parameter estimation: Gaussian noise
- Parameter estimation: fitting straight line I

On board:
- I will be out of town on Friday but we will have a Jupyter notebook lab session facilitated by Xilin and Jordan. You'll work on:
  - * radiactive lighthouse (quick overview today)
  - exploring pdfs
  - parameter estimation (two notebooks available)

Step through medical example key.
Comments:
2. Why is it $p(H|D)$ and not $p(H,D)$?
Recall that $p(H,D) = p(H|D) \cdot p(D)$ ← If you know $p(D)=1$, then they are the same.
↖ generally what you are interested in.
3,4 straightforward
5. Emphasis on sum rule. Why didn't a column of the sum product rule notebook add to 1? Because we were looking at $p(\text{tall}, \text{blue}) + p(\text{short}, \text{blue}) \neq 1$ whereas $p(\text{tall}|\text{blue}) + p(\text{short}|\text{blue}) = 1$ ✓
6. Emphasize usefulness of Bayes' rule to express $p(H|D)$ in terms of $p(D|H)$.
7. straightforward
8. Standard, but not so obvious at first. After familiar, jump right to end.
9. Stress the analysis of the result.

8/28/19

Recap of coin flipping notebook.

- Recall the names of the pdfs in Bayes' rule: posterior, likelihood, prior, evidence; and Bayesian updating: prior + data → posterior → new prior

Take-aways from coin flipping. ⇒ Note: added "New data" button (show code)

(1) Different priors eventually give the same posterior with enough data ⇒ This is called Bayesian convergence.
- How many tosses are enough?
  - Hit "New data" multiple times to see fluctuations.
  - Clearly depends on p_h and how close you want the posteriors to be
  - $p_h = 0.4$ ⇒ ≈ 200 tosses get you most of the way
  - $p_h = 0.9$ ⇒ much longer for informative prior.

(2) Ask class: Why does the "anti-prior" work well even though its dominant assumptions are proven wrong early on?
⇒ "heavy tails" mean it is like uniform (renormalized!) after the ends are eliminated. A lesson for formulating priors: allow for deviations from your expectations.

(3) Return to pages (14) and (15) and go through them, code: `y_i = stats.beta.pdf(x, alpha_i + heads, beta_i + N - heads)`

(4) Is there a difference between updating sequentially or all at once?
Do simplest problem first: two tosses.
- Let results be $D = \{D_k\}$ (in practice tails → 0's and 1's ← heads ⇒ $R = \sum_k D_k$)

General $p(p_h | \{D_k\}, I) \propto p(\{D_k\} | p_h, I)\, p(p_h | I)$

$K=1$ ⇒ $p(p_h | D_1, I) \propto p(D_1 | p_h, I)\, p(p_h | I)$ ← flipped both at once

$K=2$ only ⇒

$p(p_h | D_2, D_1, I) \propto p(D_2, D_1 | p_h, I)\, p(p_h | I)$ — class: why? Bayes' rule

$\propto p(D_2 | p_h, D_1, I)\, p(p_h | D_1, I)$ — Bayes' rule ($D_1 \in I$!)

$\propto p(D_2 | p_h, I)\, p(p_h | D_1, I)$ — if tosses are independent

Last line is sequential! (Prior for 2nd flip is posterior from first flip)

So all at once is same as sequential as function of p_h when normalized.

8/30/19

To go to 3:

$$p(p_h|D_3,D_2,D_1,I) \propto p(D_3|p_h,I) \times p(p_h|D_2,D_1,I)$$
$$\propto p(D_3|p_h,I) \times p(D_2|p_h,I) \times p(D_1|p_h,I)\, p(p_h)$$

and so on.

(5) What about "bootstrapping"?

Why can't we use the data to improve the prior and apply it (repeatedly) for the same data. Consider an extreme case!

$$p_1(p_h|D_1,I) \propto p(D_1|p_h,I)\, p(p_h|I)$$

$$\Rightarrow p_2(p_h|D_2,I) \propto p(D_1|p_h,I)\, p_1(p_h|D_1,I)$$
$$\propto [p(D_1|p_h,I)]^2\, p(p_h|I)$$

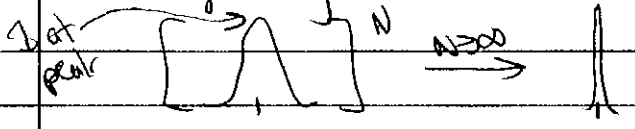
keep going?

$$p_N(p_h|D_1,I) \propto p(D_1|p_h,I)\, p_{N-1}(p_h|D_1,I)$$
$$\propto [p(D_1|p_h,I)]^N\, p(p_h|I)$$

Suppose D_1 was 0, then $[p(\text{tails}|p_h,I)]^N \propto (1-p_h)^N\, p(p_h|I) \rightarrow$ only $p_h=0$

If D_1 was 1, then $[p(\text{heads}|p_h,I)]^N \propto p_h^N\, p(p_h|I) \rightarrow$ only $p_h=1$

More generally it would cause the posterior to get narrower and narrower.



Moral: Don't do that!

Something to come back to: Frequentist point estimates.

Maximum-likelihood means: What value of p_h maximizes $\mathcal{L} = N_{p_h^R}^{\ }(1-p_h)^{N-R}$

$$\frac{d}{dp_h}(\ ) = N\left(R p_h^{R-1}(1-p_h)^{N-R} - (N-R)\, p_h^R (1-p_h)^{N-R-1}\right) = 0 \Rightarrow \boxed{p_h = \frac{R}{N}}$$

Similarly, $\sigma = \sqrt{\frac{p_h(1-p_h)}{N}}$ Revisit next week in context of Gaussian approximation

8/28/19

- Set up radioactive lighthouse problem as another Bayesian updating (and parameter estimation) problem. See page (16).
  - Go through notebook on Friday.

- Interlude: bring up exploring_pdfs.ipynb
  - Point out projected posterior plots and preview what we learn.
  - Look at multimodal plot and note that projections don't always show multiple modes.
  - Sampling from 2d pdfs ⇒ fluctuations.

Parameter estimation

Overview comments

- In general terms, "parameter estimation" in physics means obtaining values for parameters (constants) that appear in a theoretical model that describes data. (Exceptions exist, of course)
  - examples: - couplings in a Hamiltonian
    - coefficients of a polynomial or exponential model of data

- Conventionally this process is known as "fitting the parameters" and the goal is to find the "best fit" and maybe error bars.

- We will make particular interpretations of these phrases from our Bayesian point of view.

- Plan: set up the problem and look at how familiar ideas like "least-squares fitting" show up from a Bayesian perspective.

- As we proceed, we'll make the case that for physics a Bayesian approach works well.