9/6/19

Physics 8805: Learning from Data: Lecture 5

Before class: Set up laptop with Jupyter notebooks:

list on board:
- (A) radioactive_lighthouse_exercise_key.ipynb
- (B) parameter_estimation_Gaussian_noise.ipynb
- (C) parameter_estimation_fitting_straight_line_I.ipynb
- (D) assignment_01.ipynb

On board:

- Office hours next week: ?? (What works?)
- Assignment 1, due next Friday, is a follow-up to the radioactive lighthouse exercise using MCMC.

  Due: Complete notebook (D) by combining and extending results and code from (A) and (B) [or any other of our notebooks]

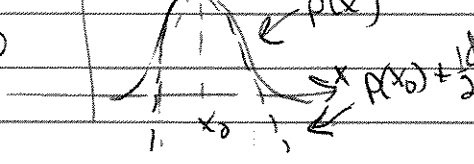
  Learning goals: (remember: toss is water to learn how to swim)
  - Be able to re-use md and Python from existing notebooks to perform similar tasks (even if not understanding all details) ⇒ generating data, sampling via MCMC, making plots.
  - Apply basic ideas of Bayesian statistics ⇒ Bayes theorem, priors, sampling posterior
  - Analyze results (with hints)
  - Try out markdown.

  note: shouldn't need to recalculate anything ⇒ focus on notebook ingredients

- Follow-up to discussion of Gaussian approximation to a pdf (see (27))

- Critical problem (not yet mentioned) with doing a straight Taylor series approximation to a pdf: doesn't satisfy conditions for pdf (normalizable, positive definite).

⇒



- In contrast $L(x) = \log p(x) \doteq L(x_0) + \frac{1}{2}\frac{d^2L}{dx^2}\Big|_{x_0}(x-x_0)^2 \Rightarrow p(x) \approx Ae^{\frac{1}{2}\frac{d^2L}{dx^2}\Big|_{x_0}(x-x_0)^2}$ ($\frac{d^2L}{dx^2}\Big|_{x_0} < 0$) is positive definite, renormalizable, and a higher order approximation to $p(x)$ because it includes all orders in $(x-x_0)^2$.

9/6/19

Let's recap notebook (B) in light of Assignment 1.

- Jump to Bayesian approach and come back to contrast with frequentist approach. Goal: sample a posterior

two random variables; data (x points); data (where flashes hit x_k)

$$p(\underbrace{\mu,\sigma}_{\theta}|D,I) \longrightarrow p(\underbrace{x_0,y_0}_{\theta}|D,I) \quad \text{same general problem!}$$

$\Rightarrow p(\theta|D,I)$

- What do we need? From Bayes theorem, need

likelihood: $p(D|\mu,\sigma,I) \rightarrow p(D|x_0,y_0,I)$ } you figure out based on lighthouse exercise

priors $p(\mu,\sigma|I) \rightarrow p(x_0,y_0|I)$ }

- Note that you are generalizing the functions for log pdfs and the plotting of posteriors in notebook (A)

- Note in (B) the functions for log-prior and log-likelihood,
  - Here $\theta=[\mu,\sigma]$ is a vector of parameters,
  - $\Rightarrow \theta=[x_0,y_0]$
- Step through set up for emcee ← need to install emcee and corner. Best to use environment, but email if difficulties.
  - hint: nothing here will change!!
  - More next week on what is happening, but basically we are doing 50 random walks in parallel to explore the posterior. Where the walkers end up will define our samples of μ,σ
  - $\Rightarrow$ the histogram is an approximation to the joint posterior.
  - Plotting is also the same, once you change labels and mu_true, sigma_true to x0_true, y0_true. (And skip maxlike part!)
  - Analysis:
    - maximum likelihood here is frequentist estimate → optimization problem
    - Are μ and σ correlated or uncorrelated?
    - Read off marginalized estimates for μ and σ.

- Return as time permits to: Bayesian vs. frequentist probability on (24), (28) and confidence intervals
  denominator $p(D|I)$ on (25) and why it is a normalization

9/6/19

General form of the central limit theorem (CLT):

The sum of n random variables that are drawn from any pdf(s) & finite variance σ^2 tends as $n \to \infty$ to be Gaussian distributed about the expectation value of the sum, with variance $n\sigma^2$. (So we scale the sum by $\frac{1}{\sqrt{n}}$ — see next page.)

Consequences:

1. The mean of a large number of values becomes normally distributed regardless of the probability distribution from which the values are drawn. (This fails for lighthouse!)

2. Functions such as the Binomial and Poisson distributions all tend to look like Gaussian distributions in the limit of a large number of drawings

$$\text{E.g., } P_n = \frac{\lambda^n e^{\lambda}}{n!} \text{ (n integer)} \xrightarrow[n \to x \to \text{large}]{\lambda \to \text{large}} p(x) = \frac{e^{-(x-\lambda)^2/2\lambda}}{\sqrt{2\pi\lambda}} \leftarrow \begin{pmatrix} \mu = \lambda \\ \sigma^2 = \lambda \end{pmatrix}$$

(Class: How would you verify this in a Jupyter notebook? How would you prove it analytically?)

Start with independent random variables $x_1, \ldots, x_n$ drawn from a distribution with mean $\langle x \rangle = \int x\, p(x)\, dx = 0$ ⇐ generalize later and $\langle x^2 \rangle = \sigma$ $\left[\langle x^n \rangle = \int x^n p(x)\, dx\right]$

Let $X = \frac{1}{\sqrt{n}}(x_1 + x_2 + \cdots + x_n) = \sum_{j=1}^{n} \frac{x_j}{\sqrt{n}}$ [need to scale by $\frac{1}{\sqrt{n}}$ for finite X in $n \to \infty$ limit]

What is the distribution of X?

⇒ call it $p(X|I)$ where I is the information about how X is drawn.

Plan: Use the sum and product rule and their consequences to relate $p(X)$ to what we know of $p(x_j)$. [Be careful of large X vs. small x!]

9/6/19

Class: fill in the rule used to justify the following steps:
(Note: we'll suppress I to keep from getting too cluttered.)

introduce x_i into problem

$$p(X) = \int_{-\infty}^{\infty} dx_1 \cdots dx_n \, p(X, x_1, \ldots, x_n)$$ why? marginalization

$$= \int_{-\infty}^{\infty} dx_2 \cdots dx_n \, p(X|x_1,\ldots,x_n) p(x_1,\ldots,x_n)$$ product rule

$$= \int_{-\infty}^{\infty} dx_2 \cdots dx_n \, p(X|x_1,\ldots,x_n) p(x_2)\cdots p(x_n)$$ independence

What is $p(X|x_1,\ldots,x_n)$? $\Rightarrow \delta\left(X - \frac{1}{\sqrt{n}}(x_1+\ldots+x_n)\right)$

Rather than use the δ-function to evaluate one integral, use a Fourier representation:

$$\delta\left(X - \frac{1}{\sqrt{n}}(x_1+\ldots x_n)\right) = \frac{1}{2\pi}\int_{-\infty}^{\infty} dw \, e^{iw\left(X - \frac{1}{\sqrt{n}}\sum_{j=1}^{n} x_j\right)} = \frac{1}{2\pi}\int_{-\infty}^{\infty} dw \, e^{-iwX} \prod_{j=1}^{n}\left[\int_{-\infty}^{\infty} dx_j \, e^{\frac{iwx_j}{\sqrt{n}}} p(x_j)\right]$$

Observe that the terms in [ ] are all the same and they have factorized!
Suppose we Taylor expand $e^{\frac{iwx_j}{\sqrt{n}}}$, assuming that the Fourier integral is dominated by small x as $n\to\infty$: (when does this fail?)

$$e^{\frac{iwx}{\sqrt{n}}} = 1 + \frac{iwx}{\sqrt{n}} + \frac{(iw)^2x^2}{2n} + O\left(\frac{w^3x^3}{n^{3/2}}\right) \Rightarrow \int_{-\infty}^{\infty} dx\, p(x)\left[1 + \frac{iwx}{\sqrt{n}} + \frac{(iw)^2x^2}{2n} + \ldots\right]$$

$\Rightarrow = 1 + \frac{iw}{\sqrt{n}}\underbrace{\langle x\rangle}_{0 \text{ by assumption}} - \frac{w^2}{2n}\underbrace{\langle x^2\rangle}_{\sigma^2} + \langle x^3\rangle O\left(\frac{w^3}{n^{3/2}}\right)$ ← σ^2 assumed to be finite

$$\Rightarrow p(X) = \frac{1}{2\pi}\int_{-\infty}^{\infty} dw \, e^{-iwX}\left[1 - \frac{w^2}{n}\sigma^2 + O\left(\frac{w^3}{n^{3/2}}\right)\right]^n \quad \text{but } \lim_{n\to\infty}\left(1+\frac{a}{n}\right)^n = e^a$$

$$\Rightarrow \frac{1}{2\pi}\int_{-\infty}^{\infty} dw \, e^{-iwX} e^{-w^2\sigma^2} = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{X^2}{2\sigma^2}} \quad \text{(QED)}$$

- To generalize to $\langle x\rangle \neq 0$, consider $X = [(x_1+\ldots x_n) - n\mu]/\sqrt{n}$ and change to $y_i = x_i - \mu$ ⇒ sum of y_i's applies.

- Why does this work?

9/6/19

The pdfs we've seen for $p(\mu, \sigma | D, I)$ were characterized by elliptical contours of equal probability density whose major axes are aligned with the μ and σ axes.
- We have commented that this is a signal of independent random variables.
- Let's look at a case where this is not true and then look analytically at what we should expect with correlations.

So return to notebook (C) on fitting a straight line.
- Step through (6), renewing the statistical model.
- What are we trying to find? $p(\theta | D, I)$, just as for (5) and (D), with now $\theta = [b, m]$.
  - intercept, slope

Comments on notebook:
- note that x_i is also randomly distributed uniformly
- log likelihood gives fluctuating results whose size depend on # of data points N and standard deviation of noise σ_y.
  ⇒ if time, explore in exercise session how size varies with these.

- Compare priors on slope ⇒ uniform in m vs. uniform in angle
- implementation of plots comparing priors – class comments
  - with first set of data with N=20 points, does prior matter? No!
  - with second set " " " N=3 points, " " " ? Yes!
- note log posterior = log(likelihood) + log(prior)
  - maximum taken to be 1 for plotting
  - exponentiate: posterior = exp(log posterior)

intercept $b = 25.0$
slope $m = 0.5$
$\theta = 5.0$

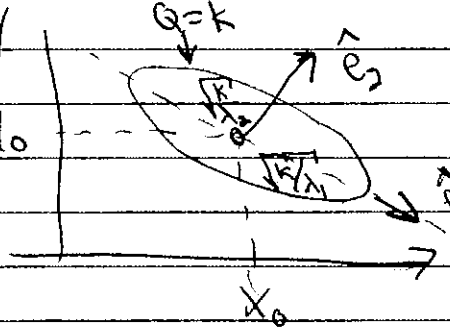
- What does it mean that the ellipses are slanted? (coming soon!)

- 2nd set of data: flat gives $b = -50 \pm 75$, $m = 1.5 \pm 1$ so barely in 3σ
  symmetric gives $b = 25 \pm 50$, $m = .5 \pm .75$ so much better!

Switch to corner!

9/6/19

Likelihoods with two variables (or posteriors) with quadratic approximation



Find X_0, Y_0 (best estimate) by differentiating

$L(X,Y) = \log p(X, Y \mid \{data\}, I)$

$$\Rightarrow \left.\frac{\partial L}{\partial X}\right|_{X_0,Y_0} = 0 \qquad \left.\frac{\partial L}{\partial Y}\right|_{X_0,Y_0} = 0$$

- To check reliability, Taylor expand around $L(X_0, Y_0)$:

$$L = L(X_0,Y_0) + \frac{1}{2}\left[\left.\frac{\partial^2 L}{\partial X^2}\right|_{X_0,Y_0}(X-X_0)^2 + \left.\frac{\partial^2 L}{\partial Y^2}\right|_{X_0,Y_0}(Y-Y_0)^2 + 2\left.\frac{\partial^2 L}{\partial X \partial Y}\right|_{X_0,Y_0}(X-X_0)(Y-Y_0) + \ldots\right] + \ldots \equiv L(X_0,Y_0) + \tfrac{1}{2}Q + \ldots$$

Makes sense to do this in matrix notation

$$Q = (X-X_0 \;\; Y-Y_0)\begin{pmatrix} A & C \\ C & B \end{pmatrix}\begin{pmatrix} X-X_0 \\ Y-Y_0 \end{pmatrix} \quad \text{symmetric}$$

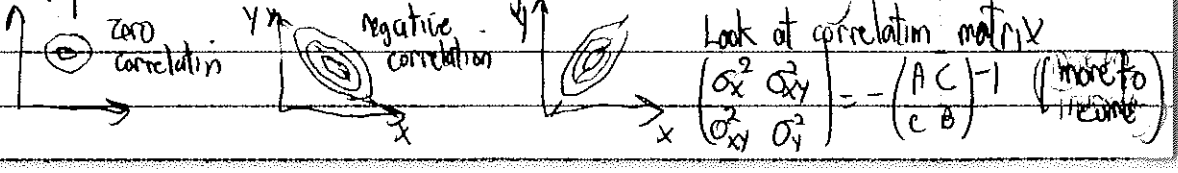
$$A = \left.\frac{\partial^2 L}{\partial X^2}\right|_{X_0,Y_0} \qquad B = \left.\frac{\partial^2 L}{\partial Y^2}\right|_{X_0,Y_0} \qquad C = \left.\frac{\partial^2 L}{\partial X \partial Y}\right|_{X_0,Y_0}$$

- So in quadratic approximation, the contour $Q=k$ for some k is an ellipse centered at X_0, Y_0. Orientation and eccentricity determined by A, B, and C.

- Principal axes found from eigenvectors of $\begin{pmatrix} A & C \\ C & B \end{pmatrix}$ (Hessian matrix)

$$\begin{pmatrix} A & C \\ C & B \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} = \lambda\begin{pmatrix} x \\ y \end{pmatrix} \Rightarrow \lambda_1, \lambda_2 < 0 \text{ (so } X_0, Y_0 \text{ is a maximum)}$$

What if ellipse is skewed? We can marginalize (discussed in future lecture)



Look at correlation matrix

$$\begin{pmatrix} \sigma_X^2 & \sigma_{XY}^2 \\ \sigma_{XY}^2 & \sigma_Y^2 \end{pmatrix} = -\begin{pmatrix} A & C \\ C & B \end{pmatrix}^{-1}$$

(more to come)