9/27/19

Physics 8805: Learning from Data: Lecture 11

On board: Reminder: all handwritten lecture notes are available on Carmen in scanned pdf form.

Notebooks for today:
- MCMC-diagnostics.ipynb

Great reference: "The Matrix Cookbook". Full of useful identities.

Quick follow-up to last time:

Suppose you have sampled the posterior $p(\vec{\theta}|D)$ where $\vec{\theta}$ is a vector of 8 parameters. You have N samples, each with 8 numbers. Now you want the joint posterior for θ_0 and θ_1, with θ_2-θ_7 marginalized. What do you do?

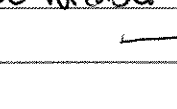
answer: Take θ_0 and θ_1 only from the N samples.

If you want a 2-d posterior, histogram them.

If you want $\langle g(\theta_0,\theta_1)\rangle \approx \frac{1}{N}\sum_{i=1}^{N} g(\theta_0^{(i)},\theta_1^{(i)})$, ignoring θ_2-θ_7!

- The confusion might be that you know $p(\theta_0,\theta_1)=\int d\theta_2 \cdots d\theta_7\, p(\theta_0,\theta_1,\ldots\theta_7)$ but don't realize that to do that integral in this context is just to ignore those variables (because projecting = marginalizing).

$p(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu_x)^2}{2\sigma_x^2}}$



- Recap of combining random variables:

$X\sim N(\mu_x,\sigma_x^2)$ and $Y\sim N(\mu_y,\sigma_y^2) \Rightarrow aX+bY \sim N(a\mu_x+b\mu_y, a^2\sigma_x^2+b^2\sigma_y^2)$

- generalize to $X_1+X_2+X_3+\ldots X_m \sim N(\mu_1+\ldots+\mu_m, \sigma_1^2+\cdots+\sigma_m^2)$

- Now continue for case of correlated $X,Y \Rightarrow$ (63), (64)

9/27/19

MCMC Sampling II: Assessing Convergence

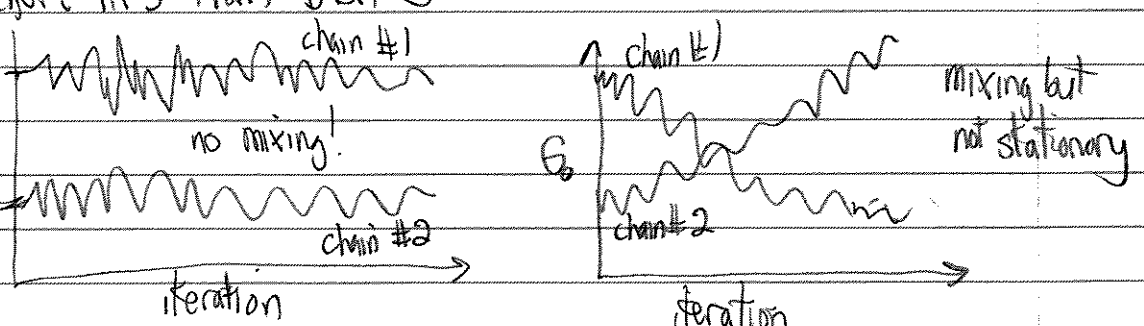
- We've seen that using MCMC with the Metropolis-Hastings algorithm can lead to a Markov chain — a set of configurations of the parameters we are sampling — that enables inference ⇒ they are samples of the posterior of interest.

- But how do we know the chain is converged — that is, that it is providing samples of the stationary distribution?
  ⇒ we need diagnostics of convergence.

- What can go wrong?
  - Could fail to have converged
  - Could exhibit apparent convergence but full posterior has not been sampled
  - Convergence could be only approximate because the samples are correlated (need to run longer).

- Strategies to monitor convergence:
  1. Run multiple chains with distributed starting points
  2. Compute variation between and within chains
     ⇒ look for mixing and stationarity
  3. Make sure acceptance rate for MC steps is not too low or too high.

Figure 11.3 from BDA-3:



Future: parallel tempering

9/27/19

Step through MCMC-diagnostics.ipynb notebook ...

[Note: this notebook needs to be rewritten! Here: laundry list of diagnostics ⇒ return when we do pymc3.]

- BDA-3 Figure 11.1
  - a) not converged
  - b) 1000 iterations ⇒ possibly converged
  - c) (correlated) draws from target distribution

- Straight line fitting again...
  - emcee sampler — but with Metropolis-Hasting ⇒ stepsize specified
  - get chains: sampler.flatchain flattens the chain (all walkers)

don't use the code here as an example ⇒ use other notebooks

- How do we know this chain has converged to the posterior?
  - Standard error of the mean
    - How does mean of $\bar{\theta}$ deviate in the chain ⇒ simulation error of the mean, not the underlying uncertainty of $\bar{\theta}$.

$$SE(\bar{\theta}) = \frac{\text{Posterior Standard Deviation}}{\sqrt{N}}$$

← cf. notebook on CLT and expected distribution of means.

  - Visualization with moving average ⇒ check for stability

- Autocorrelation: do you recognize the formula
- Acceptance rate. Usually autotuned in packaged MCMC software.

- Assessing mixing: Gelman Rubin diagnostic (pass 1)
  - Basic idea: multiple chains from different walkers (after warmup) are split up and one looks at the variance within a chain and between chains.
  - See BDA-3 pp 284-5 for details. We'll come back to this.
  - Now documented internally in notebook, but we'll see it again.

Try changing stepsize from .002 to .005 or higher.

(68)

9/27/19

Maximum Likelihood for Least-Squares Fitting

- Here we consider the problem of fitting a polynomial to correlated data. We do this here with a frequentist approach and come back later to the Bayesian way.

- Our underlying motivation is to get familiar with the linear algebra manipulations.

- We'll use the notation and some of the discussion from Hogg, Bovy, Lang, arXiv: 1008.4686 called "Data analysis recipes: Fitting a model to data".

- This paper has 41 extended reference notes, worth looking at. Provocative statements about fitting straight lines:
  "Let us break with tradition and observe that in almost all cases in which scientists fit a straight line to their data, they are doing something simultaneously wrong and unnecessary,"
  - wrong because it is very rare that a set of two-dimensional measurements are truly drawn from a narrow, linear relationship
  - probably not linear in detail
  - unnecessary because communicated slope and intercept are much less informative than the full distribution of data.

- Let's consider fitting N data points to a straight line $y = mx + b$.

$$Y = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix} \Rightarrow \underset{\text{matrix}}{N\times 1} \qquad A = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_N \end{bmatrix} \Rightarrow \underset{\text{matrix}}{N\times 2} \qquad \Sigma = \begin{bmatrix} \sigma_{y_1}^2 & \rho_{12}\sigma_{y_1}\sigma_{y_2} & \cdots & \rho_{1N}\sigma_{y_1}\sigma_{y_N} \\ & \sigma_{y_2}^2 & & \\ & & \ddots & \\ \text{symmetric} & & & \sigma_{y_N}^2 \end{bmatrix}$$

- Goal: find $\theta = \begin{bmatrix} b \\ m \end{bmatrix}$ where $Y = A\theta$ ($N\times N$ covariance matrix)

$\underset{\text{matrix}}{2\times 1}$

$(N\times 1) \quad (N\times 2)\cdot(2\times 1) = (N\times 1)$ ✓ (equation has consistent matrices)

9/27/19 (69)

Check the $N=2$ case:

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \end{bmatrix} \begin{bmatrix} b \\ m \end{bmatrix} = \begin{pmatrix} b+mx_1 \\ b+mx_2 \end{pmatrix}$$ so each row is $y_i = b+mx_i$ ✓

- Reader: convince yourself that $N=3$ and higher is correct.

Aside: Why don't we just solve the matrix equation?

$$Y = A\theta \overset{?}{\Rightarrow} \theta = A^{-1}Y \; ??$$

Because the equation is overconstrained for $N>2$ (ok for $N=2$!)

Frequentist answer: maximize the likelihood $\propto e^{-\chi^2}$

familiar, uncorrelated case: $\chi^2 = \sum_{i=1}^{N} \frac{[y_i - f(x_i)]^2}{\sigma_{y_i}^2}$ (χ^2 is a scalar.)

generalized, correlated case: $\chi^2 = \underbrace{[Y-A\theta]^T}_{(1\times N)} \underbrace{\Sigma^{-1}}_{(N\times N)} \underbrace{[Y-A\theta]}_{(N\times 2)-(N\times 2)(2\times 1)} \to 1\times 1$

maximum likelihood estimate, MLE

Claim: $\hat{\theta} = [A^T\Sigma^{-1}A]^{-1}[A^T\Sigma^{-1}Y]$

Aside: why can't I say $[A^T\Sigma^{-1}A]^{-1}[A^T\Sigma^{-1}Y] = A^{-1}\Sigma(A^T)^{-1}A^T\Sigma^{-1}Y = A^{-1}Y$?
⇒ not square, invertible matrices

Plausibility argument for $\hat{\theta}$ result. We need square, invertible matrices

$$Y = A\theta \qquad N\times 1 = (N\times 2)\cdot(2\times 1) \qquad \text{for } \theta = \begin{bmatrix} b \\ m \end{bmatrix}$$

$$\Sigma^{-1}Y = \Sigma^{-1}A\theta \qquad (N\times N)\cdot(N\times 1) = N\times 1$$

$$A^T\Sigma^{-1}Y = A^T\Sigma^{-1}A\theta \qquad (2\times N)\cdot(N\times 1) = 2\times 1$$

$$\underbrace{(2\times N)\cdot(N\times N)\cdot(N\times 2)}_{2\times 2} = 2\times 2$$

Now invert: $\boxed{\theta = (A^T\Sigma^{-1}A)^{-1}(A^T\Sigma^{-1}Y)}$

9/27/19

Before proving the result more carefully, let's generalize to a higher-order polynomial:

If $y_i = qx_i^2 + mx_i + b \Rightarrow \theta = \begin{bmatrix} b \\ m \\ q \end{bmatrix} \Rightarrow A = \begin{bmatrix} 1 & x_1 & x_1^2 \\ \vdots & \vdots & \vdots \\ 1 & x_N & x_N^2 \end{bmatrix}$

and similarly for higher order.

Prove using indices with summation convention:

$$\chi^2 = (Y - A\theta)^T \Sigma^{-1} (Y - A\theta) = (Y_i - A_{ij}\theta_j)\,\Sigma^{-1}_{ii'}\,(Y_{i'} - A_{i'j'}\theta_{j'})$$

where i, i' run from 1 to N and j, j' from 1 to p (highest power x^{p-1})

Find MLE from $\frac{\partial \chi^2}{\partial \theta_k} = 0$ for $k = 1$ to p

$$\Rightarrow \left.\frac{\partial \chi^2}{\partial \theta_k}\right|_{\theta \to \hat{\theta}} = -A_{i'j}\delta_{jk}\Sigma^{-1}_{ii'}(Y_{i'} - A_{i'j}\hat{\theta}_j) + (Y_i - A_{ij}\hat{\theta}_j)\Sigma^{-1}_{ii'}(-A_{i'j'}\delta_{j'k}) = 0$$

- move $\hat{\theta}$ terms to the opposite side, and show doubled terms are equal (overall multiply by -1)

$$\Rightarrow A_{i'k}\Sigma^{-1}_{ii'}Y_{i'} + Y_i\Sigma^{-1}_{ii'}A_{i'k} = A_{ik}\Sigma^{-1}_{ii'}A_{i'j}\hat{\theta}_{j'} + A_{ij}\hat{\theta}_j\Sigma^{-1}_{ii'}A_{i'k}$$

(take $j' \to j$; Σ^{-1} symmetric)

$$\underbrace{A_{i'k}}_{(A^T)_{ki'}}\Sigma^{-1}_{ii'}Y_{i'} + A_{ik}\Sigma^{-1}_{ii'}Y_i = A_{ik}\Sigma^{-1}_{ii'}A_{i'j}\hat{\theta}_j + A_{ik}\Sigma^{-1}_{ii'}A_{i'j}\hat{\theta}_j$$

($i \leftrightarrow i'$ and $\Sigma^{-1}_{ii'} = \Sigma^{-1}_{i'i}$)

$$\Rightarrow 2(A^T)_{ki}\Sigma^{-1}_{ii'}Y_i = 2(A^T)_{ki}(\Sigma^{-1})_{i'i}A_{i'j}\hat{\theta}_j$$

or $$(A^T\Sigma^{-1}Y) = (A^T\Sigma^{-1}A)\hat{\theta}$$

($A^T\Sigma^{-1}A$: square, invertible)

$$\Rightarrow \boxed{\hat{\theta} = (A^T\Sigma^{-1}A)^{-1}(A^T\Sigma^{-1}Y)} \quad \text{Q.E.D.}$$