Left off with "Bayesian Learning"

eg. $P(h|D) = \frac{P(D|h)P(h)}{P(D)}$

but what does that mean? How to compute in practice?

Today: Sample-based approximate bayesian inference

running example: bayesian linear regression: height & weight

model: $w = b_0 + b_1 h + \epsilon$, $\epsilon \sim N(0, \sigma^2)$

$b_0, b_1 \sim$??

ols: $y = X\beta + \epsilon$

$\hat{\beta} = (X^T X)^{-1} X^T y$

$\hat{\beta} \sim N[\beta, \sigma^2 (X^T X)^{-1}]$

$D = \begin{bmatrix} h_1 & w_1 \\ h_2 & w_2 \\ \vdots & \\ h_n & w_n \end{bmatrix}$ $X = \begin{bmatrix} 1 & h_1 \\ 1 & h_2 \\ \vdots & \\ 1 & h_n \end{bmatrix}$

purpose: queries of form $P(w_{new} | h_{new}, D) = E[f(h)|D] = \int f(h;\theta) P(\theta|D) d\theta$

or more generally: $E[f(x)] = \int f(x) p(x) dx$

1) ordinary Monte-carlo:

$$E[f(x)] \approx \frac{1}{n} \sum_{i=1}^{n} f(x_i), \quad x_i \overset{iid}{\sim} p(x) \qquad \text{(general)} \qquad (1)$$

$$E[f(h|D)] \approx \frac{1}{n} \sum_{i=1}^{n} [b_0 + b_1 h], \quad b_0, b_1 \sim p(b_0, b_1 | D)$$

↖ how to get these samples??

ideas??

bias: $E[\theta] - E[\frac{1}{n}\sum f(x)] = \frac{1}{n}\sum$ (bias & var) over

ideas (cont'd)

→ OLS works in special cases where we have closed-form estimator

→ why hard? Recall: $P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)} = \frac{P(D|b_0,b_1)P(b_0,b_1)}{\iint_{b_0 b_1} P(D|b_0,b_1)P(b_0,b_1)}$ ← hard

↳ to do analytically, need tractable parametric form for product of both terms

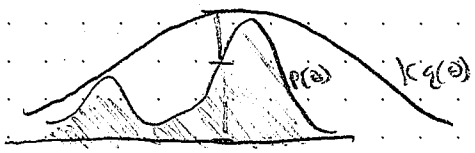
observe: $P(\theta|D) = \frac{1}{Z} P(D|\theta)P(\theta) \propto P(D|\theta)P(\theta)$

↑ proportional to

idea: ignore normalizing constant $\frac{1}{Z}$ somehow?

A) Rejection sampling

* ↳ introduce simple proposal distribution $q(\theta)$, & constant k s.t. $kq(\theta) \geq P(\theta)$



accept prob $a(\theta) = \frac{P(\theta)}{kq(\theta)}$ (2)

problems?
↳ obtaining k
↳ what if q is "bad" → rejects most proposals

B) importance sampling

idea: plug (2) directly into Prob (1)

$$\int f(x;\theta)p(\theta)d\theta = \int \frac{f(x;\theta)p(\theta)}{q(\theta)} q(\theta)\,d\theta$$

ie: sample from easy proposal dist., & then correct them using weights from (2)

problems? if "peaky", can be dominated by low-weight samples & be inefficient

3) MCMC* : Markov-chain Monte-Carlo

goal: more efficient method for getting samples from $P(\theta|D)$

↳ same assumptions of (1) & (2): only have proposal dist. & unnormalized prob

↳ intuitively, we'll get high-prob samples by searching for them

quick review of Markov Chains

↳ MC is an indexed sequence of RV's ; defined entirely by transition probs

$$P_{ij} = P(\theta_{t+1} = j \mid \theta_t = i) \quad \text{[transition dist]}$$

$$\begin{array}{c} \\ (i) \\ \\ \\ \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & \cdots & (j) \end{array} \\ \begin{bmatrix} & & & \\ & P_{ij} = P(j|i) & & \\ & & & \end{bmatrix} \end{array} \begin{array}{l} \\ \\ \\ \\ \rightarrow \Sigma = 1 \end{array}$$

↳ interested in equilibrium conds & convergence rate

key condition: reversibility ("detailed balance")

notation:

$\pi_i = P(\theta = i)$

π = eq. distribution

↳ in equilibrium if $\pi = \pi P \rightarrow \pi_j P_{ji} = \pi_i P_{ij}, \ \forall i,j$ (4)

↳ π is a distribution over Θ

* (b) ... build a sampling strategy which defines a MC, & *

therefore lets us guarantee convergence to a stationary (posterior) dist.

using a sequential random walk (ie search) algorithm

this is generic Metropolis-Hastings { simple algorithm:

↳ define transition rule as $a_{ij} = \min\left[1, \underbrace{\frac{\pi_j q_{ji}}{\pi_i q_{ij}}}_{\text{Hitchhiker ratio}} \underbrace{\frac{}{}}_{\text{proposal ratio}}\right]$

Metropolis 1953
Hastings 1970

proof: $\pi_i q_{ij} a_{ij} = \pi_i q_{ij} \frac{\pi_j q_{ji}}{\pi_i q_{ij}}$

$= \pi_j q_{ji}$

$= \pi_j q_{ji} a_{ji}$ // $\therefore$ reversible

key idea: moves based on relative probability of different setting of Θ

Metropolis-Hastings MCMC (cont'd)

Basic Convergence ideas

1) Mixing : how well is chain moving around parameter space?

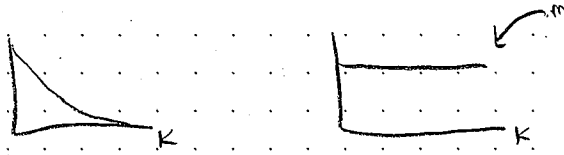
good bad

2) Auto-correlation : a statistic for examining mixing

- lag-k auto-corr := ρ_k

$$\rho_k = \frac{\sum_{i=1}^{n-k}(x_i - \bar{x})(x_{i+k} - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

- IE the average corr of points k steps apart



good bad

← means samples are highly correlated over many steps, thus not mixing

↳ eg. MH may be rejecting too much

Example: non-linear physics inference

- harping on "allows inference when joint posterior only known to a constant"
- how often does this really come up?? All the time!

eg. physics ← eg game engine

↳ $x_{t+1} = F_{\Phi}(x_t, u_t) + \epsilon$, $F = \langle$ some black box dynamics $\rangle$

↳ generative model : $\Phi, \sigma \sim P(\Phi, \sigma | h)$ likelihood: $L(\Phi | \sigma, x, u) = N(F_{\Phi}(x,u), \sigma^2)$

$\epsilon \sim N(0, \sigma_0^2)$

want: $P(\Phi, \sigma | h) \propto P(h | \Phi, \sigma) P(\Phi) P(\sigma)$

notation:
- $\theta = \{\text{model params}\}$
- $\pi = P(\theta)$
- $\pi_i = P(\theta = i)$
- $m = |\theta|$

$i = (i_1, i_2, \ldots i_m)$

i_c = randomly chosen component

$i_{\bar{c}}$ = remaining components $= (i_1, \ldots i_{c-1}, i_{c+1}, \ldots i_m)$

j is a neighbor of i if:
- $j_{\bar{c}} = i_{\bar{c}}$
  - ↳ IE can move from i to j by changing i_c

Gibbs sampling: Special case of M+H where q is a proper (normalized) conditional distribution

$$q_{ij} = \frac{1}{m} \frac{\pi_j}{\sum_{k: k_{\bar{c}} = i_{\bar{c}}} \pi_k}$$

} proposes some neighbor j with prob. normalized by all other possible neighbors

why better?
- ↳ have that $a_{ij} \equiv 1$

requires:
- ↳ deriving conditional expression

also:
- ↳ can only make steps along axes in which conditionals are defined

Proof that $a_{ij} = 1$ for Gibbs Proposals:

$$\pi_i q_{ij} = \pi_i q_{ij} \frac{\pi_j}{\pi_j}$$

$$= \pi_j \frac{\pi_i q_{ij}}{\pi_j}$$

$$= \pi_j \left[\frac{\pi_i}{\pi_j} \cdot \frac{\pi_j}{m \sum_{k: k_{\bar{c}} = i_{\bar{c}}} \pi_k} \right]$$

$$= \pi_j \left[\frac{\pi_i}{m \sum_{k: k_{\bar{c}} = i_{\bar{c}}} \pi_k} \right]$$

$$= \pi_j q_{ji}$$

B/c π_i & π_j are neighbors, we have that $i_{\bar{c}} = j_{\bar{c}} = k_{\bar{c}}$ (IE the share the same normalizing constant)

$\therefore \pi_i q_{ij} = \pi_j q_{ji} \Rightarrow$ Gibbs sampling is reversible, so no acceptance needed $[\text{IE } a_{ij} = a_{ji} = 1]$

Gibbs Sampling (cont'd)

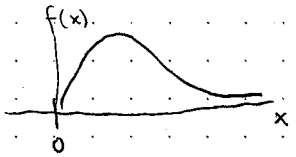
Example model: Simple linear regression

$$y = \beta_0 + \beta_1 x + \epsilon \quad , \quad \epsilon \sim N(0, \sigma^2)$$

note: $\tau = \frac{1}{\sigma^2}$

$X \sim \text{Gamma}(a,b) \Rightarrow f(x) = \frac{b^a}{\Gamma(a)} x^{a-1} e^{-bx}$

recall: $\Gamma(n) = (n-1)!$



↳ Bayesian version:

$$P(y_i \mid x_i, \beta_0, \beta_1, \tau) \sim N(\beta_0 + \beta_1 x_i, 1/\tau)$$

$$\beta_0 \sim N(\mu_0, 1/\tau_0)$$

$$\beta_1 \sim N(\mu_1, 1/\tau_1)$$

$$\tau \sim \text{Gamma}(a, b)$$

inference

↳ assume $[\mu_0, \mu_1, \tau_0, \tau_1, a, b]$ known constants (for now)

↳ model params $\Theta = [\beta_0 \ \beta_1 \ \tau]$

want: $P(\theta \mid x, y) = P(\beta_0, \beta_1, \tau \mid x, y) = P(y \mid \beta_0, \beta_1, \tau, x) \underbrace{P(\beta_0) P(\beta_1) P(\tau)}_{\text{why independent? model says so}} \underbrace{P(x)}_{\text{why ind? always observed}}$

joint posterior $\propto P(y \mid \beta_0, \beta_1, \tau, x) P(\beta_0) P(\beta_1) P(\tau)$

$$= N(\beta_0 + \beta_1 x, 1/\tau) \cdot N(\mu_0, 1/\tau_0) \cdot N(\mu_1, 1/\tau_1) \cdot \text{Gamma}(a, b)$$

$$= \left[\prod_{i=1}^{n} \frac{\sqrt{\tau}}{\sqrt{2\pi}} e^{\frac{-(y_i - \beta_0 + \beta_1 x_i)^2}{2(1/\tau)^2}}\right] \left[\frac{\sqrt{\tau_0}}{\sqrt{2\pi}} e^{\frac{-(\beta_0 - \mu_0)^2}{2(1/\tau_0)^2}}\right] \left[\frac{\sqrt{\tau_1}}{\sqrt{2\pi}} e^{\frac{(\beta_1 - \mu_1)^2}{2(1/\tau_1)^2}}\right] \left[\frac{b^a}{\Gamma(a)} \tau^{a-1} e^{-b\tau}\right] \propto \underline{\underline{\text{ugly!}}}$$

Gibbs for the save: consider each Θ_i in turn

$$P(\beta_0 \mid x, y, \beta_1, \tau) \sim N\left(\frac{\tau_0 \mu_0 + \tau \Sigma_i (y_i - \beta_1 x_i)}{\tau_0 + n\tau}, \frac{1}{\tau_0 + n\tau}\right)$$

$$P(\beta_1 \mid x, y, \beta_0, \tau) \sim N\left(\frac{\tau_1 \mu_1 + \tau \Sigma_i (y_i - \beta_0)}{\tau_1 + \tau \Sigma_i x_i^2}, \frac{1}{\tau_1 + \tau \Sigma_i x_i^2}\right)$$

$$P(\tau \mid x, y, \beta_0, \beta_1) \sim \text{Gamma}(a + n/2, \ b + \Sigma_i (y_i - \beta_0 - \beta_1 x_i)^2/2)$$

Gibbs sampling (cont'd)

So: have conditional distributions for each θ_i:

$\beta_0 \mid X, y, \beta_1, \tau$

$\beta_1 \mid X, y, \beta_0, \tau$

$\tau \mid X, y, \beta_0, \beta_1$

inference:

$i = 0$

while (not bored):

$$\beta_0^{(i)} \sim N\left(\frac{\tau_0 \mu_0 + \tau \Sigma_i (y_i - \beta_1 x_i)}{\tau_0 + n\tau}, \frac{1}{\tau_0 + n\tau}\right)$$

$$\beta_1^{(i)} \sim N\left(\frac{\tau_1 \mu_1 + \tau \Sigma_i (y_i - \beta_0)}{\tau_1 + \tau \Sigma_i x_i^2}, \frac{1}{\tau_1 + \tau \Sigma_i x_i^2}\right)$$

$$\tau^{(i)} \sim \text{Gamma}\left(a + n/2, b + \Sigma_i (y_i - \beta_0 - \beta_1 x_i)^2/2\right)$$

i++

⟨ inspect $\vec{\beta}_0, \vec{\beta}_1, \vec{\tau}$, eg $\hat{\beta}_0 = \overline{\beta_0}$ ⟩

take home:

- in general, MCMC does approximate bayesian inference for arbitrarily complex models
- standard Metropolis-Hastings is most general, but must be tuned (q can be slow)
- Gibbs sampling is special case of M-H when we have conditional distributions for some θ_i
  - ↳ strategy: write down joint posterior, isolate a single parameter, & plow through the horrible algebra