Lecture 06 - Bayesian Fun Part II

Markov Chain Monte Carlo

Remember our Bayesian view:

$$p(\theta|D) = \frac{p(D|\theta) \cdot p(\theta)}{p(D)}$$

- $p(\theta|D)$: posterior: new updated belief
- $p(D|\theta)$: likelihood: where data comes in.
- $p(\theta)$: prior: our original belief
- $p(D)$: evidence

Maximum a posteriori estimate: (MAP):

$$\theta_{MAP} = \arg\max_{\theta} p(\theta|D) = \arg\max_{\theta} p(D|\theta) \cdot p(\theta)$$

Example: Small city with an infectious disease

tested: $n = 20$ infected: $k = 3$ emphasize test is perfect.

=> what is the rate of infected people

=> naive solution: $\frac{3}{20}$

Bayesian solution:

Likelihood: Binomial => test 20, get 3 infections with unknown rate of success θ

$$p(D|\theta) = \frac{n!}{k!(n-k)!}\,\theta^k(1-\theta)^{n-k}$$

Prior: • From other cities we know $\bar{\theta} = 0.1$

• choose conjugate prior

↳ posterior has same form as the prior

• conjugate prior to Binomial is Beta distribution

$P(\theta) = \text{Beta}(\theta, \alpha, \beta)$ we know mean should be 0.1

$\Rightarrow \frac{\alpha}{\alpha+\beta} \approx 0.1 \quad \alpha = 2\,,\ \beta = 20$

$$\text{Beta}(\theta, \alpha, \beta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)} \qquad B(x,y) = \int_0^1 t^{x-1}(1-t)^{y-1}\,dt$$

↳ Beta function

For our posterior this means:

$$p(\theta|D) \propto p(D|\theta)\cdot p(\theta) = \text{Binom}(\theta, n, k)\cdot \text{Beta}(\theta, \alpha, \beta)$$
$$= \text{Beta}(\theta, \alpha+k, \beta+n-k)$$
$$= \text{Beta}(\theta, 5, 37)$$

⇒ with conjugate priors you can just look up the result

⇒ For the posterior mean we get

$\frac{\alpha}{\alpha+\beta} = \frac{5}{5+37} = \frac{5}{42}$ compare to $\frac{3}{20}$ ⇒ prior makes a difference!

Conjugate priors : Normal distribution

Let's take a random variabe y with a normal distribution:

$$p(y|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{\frac{-(y-\mu)^2}{2\sigma^2}}$$

For our data $\{y_1,\dots,y_n\}$ we have

$$p(y_1,\dots,y_n|\mu,\sigma^2) = \prod_i p(y_i|\mu,\sigma^2)$$

$$\propto e^{-\frac{1}{2}\frac{\sum_i (y_i-\mu)^2}{\sigma^2}} \xrightarrow{\text{expand}} \sum_i \left(\frac{y_i-\mu}{\sigma}\right)^2$$

$$= \frac{1}{\sigma^2}\sum_i y_i^2 - \frac{2\mu}{\sigma^2}\sum y_i + \frac{n\mu^2}{\sigma^2}$$

$\Rightarrow$ The y values all sum up

$\Rightarrow$ We don't need to know the individual $y_1,\dots,y_n$

$\Rightarrow$ <u>Sufficient statistic</u> : $\left\{\sum_i y_i^2, \sum_i y_i\right\}$

assume we know σ^2 $\Rightarrow$ only need to estimate μ:

$$p(\mu|y_1,\dots,y_n,\sigma^2) \propto p(\mu|\sigma^2) \cdot \underbrace{e^{-\frac{1}{2\sigma^2}\sum_i (y_i-\mu)^2}}_{\text{likelihood}}$$

↳ prior $N(\hat{\mu},\tau^2)$

↳ we choose $\hat{\mu}$ and τ^2

likelihood (without normalization as σ is known)

⇒ posterior is normal as well

$$p(\mu | y_1, \dots, y_n, \sigma^2) \propto N\left(\frac{b}{a}, \underbrace{\left(\frac{1}{\tau^2} + \frac{n}{\sigma^2}\right)^{-1}}_{\frac{1}{a}}\right)$$

$$a = \frac{1}{\tau^2} + \frac{n}{\sigma^2}$$

$$b = \frac{\hat{\mu}}{\tau^2} + \frac{\sum y_i}{\sigma^2}$$

now we define $K = \frac{\sigma^2}{\tau^2}$: variance of my likelihood in units of the variance of the prior.

$$\mu_p = \frac{K}{K+n}\hat{\mu} + \frac{n}{K+n}\bar{y}; \qquad \bar{y} = \frac{1}{n}\sum_i y_i$$

⇒ weighted average of the prior mean and the mean of the data

⇒ as n becomes large, the influence of the prior becomes smaller and the data dominates.

Please look at lecture notes from last year for more examples.

We now are leaving the "nice" world of conjugate priors and learn how to sample from "nasty" posteriors.

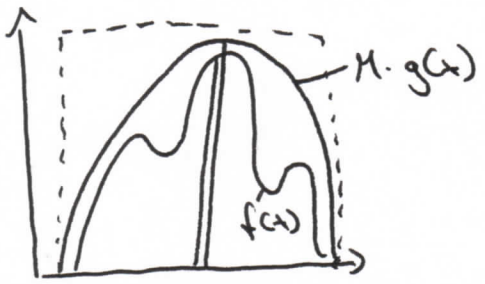
Markov-Chain-Monte-Carlo

We are back to sampling from a distribution, e.g. a posterior.

So far we have learned:

- invers transform
- rejection sampling
- rejection sampling on steroids.

Short recap for rejection sampling on steroids:



$f(x) < M \cdot g(x)$

$\Leftrightarrow M > \frac{f(x)}{g(x)}$, $M \geq 1$

if $M < 1$ you are in trouble because $M \cdot g(x) < f(x)$ could happen.

- Sample x from $g(x)$
- Sample u from $U(0,1) \leftarrow \frac{f(x)}{M \cdot g(x)} \leq 1$ always.
- check $u < \frac{f(x)}{M g(x)}$ $\Rightarrow$ yes: accept

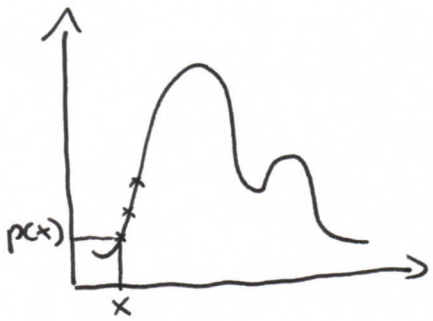
  otherwise: reject

$\Rightarrow$ we always accept if $f(x) = g(x)$, but that means we can sample from $f(x)$ anyways.

- It can be difficult to find $g(x)$
- Especially not obvious for multidimensional problems.

=> Learn about MCMC!

Intuition:



"Feel" your way along the function and sample more if you are on higher terrain.

=> We will assume that we can evaluate $p(x)$, this does NOT mean we know how to integrate or sample from it.

=> $p(x)$ could be a function, but also a look up table.

=> our random walk is "blind" => cannot see ahead to determine where I should go.

=> Compare height after taking a step to the height of the previous location => "feel" along the landscape.

Now all we need to do is formalize this intuition.

First: Markov Chains

This part of the lecture is based on Joe Blikstein "Introduction to probability", chapter 11. I highly recommend this chapter for a detailed introduction.

Markov Chain: A sequence of random variables taking values in a state space is called a Markov chain if the probability of the next step only depends on the current state.

⇒ We now are dealing with sequences of random variables.

⇒ Each random variable can have values from a state space: $x^{(i)} = [x_1^{(i)}, x_2^{(i)}, x_3^{(i)}, \dots, x_n^{(i)}]$

n: dimensions in space
i: step / iteration / time

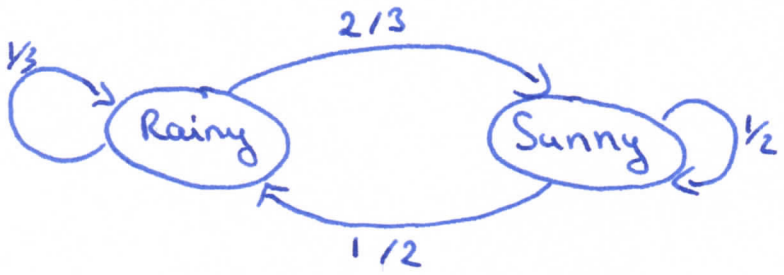
⇒ We generate a sequence of consecutive states with our chain: $x^{(1)} \to x^{(2)} \to x^{(3)} \dots$

Transitioning from one state to the other is described by the transition probability $T(x^{(i)} | x^{(i-1)})$

Note that it only depends on the previous state:

$$T(x^{(i)} | x^{(i-1)}, x^{(i-2)}, \dots, x^{(0)}) = T(x^{(i)} | x^{(i-1)})$$

Example: Rainy - Sunny



This Markov chain has two states {Rainy, Sunny}.
We can use the diagram to generate sequences like

Rainy, Rainy, Sunny, Rainy, Sunny, Sunny, Sunny, ...

The transition probabilities can be described by a transition matrix:

	Rainy	Sunny
Rainy	$\frac{1}{3}$	$\frac{2}{3}$
Sunny	$\frac{1}{2}$	$\frac{1}{2}$

Note that each row has to sum up to 1 (we have to go somewhere).

n-step transition probability: probability to go from one state to another in exactly n steps:

probability to go from Rainy to Rainy in 2 steps:

$$\underbrace{\frac{1}{3} \cdot \frac{1}{3}}_{\text{stay}} + \underbrace{\frac{2}{3} \cdot \frac{1}{2}}_{\substack{\text{Sunny,} \\ \text{rainy}}} \approx 0.444$$

The n-step transition probabilities are given by the n^{th} power of the transition matrix

If some conditions are met, the n-step transition probabilities will become stable for some large n, and it will be independent of the starting state.

The 7th power of our Rainy-Sunny transition matrix:

(rounded values)

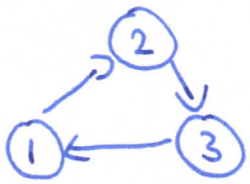
$$\begin{pmatrix} 0.4285 & 0.5714 \\ 0.4286 & 0.5714 \end{pmatrix}$$

Irreducible: Probability of reaching every state from any state is greater than zero.

irreducible

reducible

Aperiodic: No "deterministic" loops



irreducible, but period 3

Irreducible and aperiodic Markov chains reach a unique stationary distribution in the long run.

For a stationary distribution s (a row vector) and transition matrix T:

$$sT = s \quad \Rightarrow \text{ eigenvector}$$

$\Rightarrow$ A Markov chain whose initial distribution is the stationary distribution will stay in the stationary distribution forever.

For our example:

$$(s \quad 1-s) \begin{pmatrix} \frac{1}{3} & \frac{2}{3} \\ \frac{1}{2} & \frac{1}{2} \end{pmatrix} = (s \quad 1-s)$$

$$\Leftrightarrow \frac{1}{3}s + \frac{1}{2}(1-s) = s \qquad \Leftrightarrow \quad s = \frac{3}{7}$$

$$\frac{2}{3}s + \frac{1}{2}(1-s) = 1-s$$

$\Rightarrow$ The stationary distribution for our Rainy-Sunny example is $[\frac{3}{7} \; \frac{4}{7}]$.

Having an irreducible and aperiodic Markov Chain guarantees that a stationary distribution exists, but solving the eigenvector problem to find it can be intractable.

A stronger condition that can be easier to check is detailed balance:

$$P(x^{(i)}) \cdot T(x^{(i-1)} | x^{(i)}) = P(x^{(i-1)}) \cdot T(x^{(i)} | x^{(i-1)})$$

$\Rightarrow$ probability of going from one state to another and going backwards should balance out.

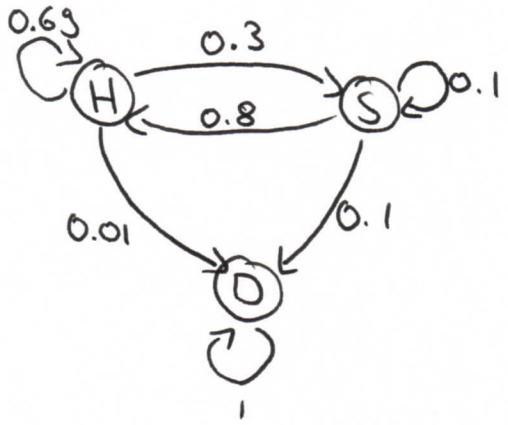
If $P(x)$ satisfies detailed balance for T, then $P(x)$ is a stationary distribution for T:

$$\sum_{x^{(i-1)}} P(x^{(i-1)}) \cdot T(x^{(i)} | x^{(i-1)}) = \sum_{x^{(i-1)}} P(x^{(i)}) \cdot T(x^{(i-1)} | x^{(i)})$$

$$= P(x^{(i)}) \cdot \underbrace{\sum_{x^{(i-1)}} T(x^{(i-1)} | x^{(i)})}_{=1} = P(x^{(i)})$$

Example: (numbers are made up)

H: healthy
S: sick
D: dead



transition probability:

	H	S	D
H	0.69	0.3	0.01
S	0.8	0.1	0.1
D	0	0	1

$T(x^{(i)} | x^{(i-1)})$

$\mu^{(1)} = (0.5, 0.2, 0.3)$

$\mu^{(2)} = \mu^{(1)} \cdot T$, $\mu^{(3)} = \mu^{(2)} \cdot T = \mu^{(1)} \cdot T \cdot T$ ...

At some point the chain will stabalize to a stable distribution.

Metropolis-Hastings:

Now we have everything together to define a Markov chain => Let's talk finally about sampling.

Goal: Sample from a target distribution $p(x)$

$$x^{(1)}, x^{(2)}, x^{(3)}, \ldots, x^{(t)} \sim p(x)$$

- random start $x^{(1)}$
- propose new state x^* according to

  $q(x^* | x^{(i)})$ => transition probability
- accept with probability $A(x^* | x^{(i)})$:

$$\min\left[1, \frac{p(x^*)}{p(x^{(i)})}\right] \Rightarrow \text{this is simplified and assumes } q(x^* | x^{(i)}) = q(x^{(i)} | x^*).$$

- If $p(x^*) > p(x^{(i)})$ we accept, because we make a step upwards
- If $p(x^*) < p(x^{(i)})$ we still need to accept it sometimes to keep exploring:

- Example: $p(x^{(i)}) = \frac{1}{3}$, $p(x^*) = \frac{1}{4}$ $\Rightarrow \frac{p(x^*)}{p(x^{(i)})} = \frac{3}{4}$

  how do I accept with $p = \frac{3}{4}$?

=> draw uniform $[0,1]$, if lower than $\frac{3}{4}$ we accept