Bayesian Statistics Crash Course

Slack ID: @Philip_Thomas

Difference Between Frequentists and Bayesians (1)

- **Frequentist**
  - Data are IID random sample from data generating process
  - Parameters are fixed

- **Bayesian**
  - Data are fixed (we have the data we observed)
  - Parameters are unknown, described probabilistically

Difference Between Frequentists and Bayesians (2)

- **Frequentist**
  - Probability derived from relative frequency / counting procedure
  - Probabilistic quantity of interest → $P(data|H_0)$

- **Bayesian**
  - Probability represents "degree of belief"
    - Our belief is updated in the light of new information
  - Probabilistic quantity of interest → $P(\theta|data)$

Difference Between Frequentists and Bayesians (3)

- **Frequentist**
  - Probability derived from relative frequency / counting procedure
  - Probabilistic quantity of interest → $P(data|H_0)$

- **Bayesian**
  - Probability represents "degree of belief"
    - Our belief is updated in the light of new information
  - Probabilistic quantity of interest → $P(\theta|data)$

Difference Between Frequentists and Bayesians (4)

- **Frequentist**
  - Report point estimates and standard errors
  - Hypothesis testing with p-values and 95%
    - (or “fill your own” %) confidence intervals

- **Bayesian**
  - Report posterior distribution
  - Hypothesis testing is done with actual probability statements through posterior probabilities
    - Ex: What is $P(actual\ issued\ rate\ increases\ by\ 5\%|\ AB\ test\ data)$?

Question: 95% Confidence Interval

- What is the correct interpretation of confidence interval?
- Is it useful?
- What interpretation do people want?

- An interval that has 95% probability of containing the true parameter values
- An interval that if we repeated the sampling, and test *infinite (read: really really big)* number of times would contain the true parameter values 95% of the time

Bayes Theorem

$$\underbrace{P(A|B)}_{\text{Posterior}} = \frac{\overbrace{P(A)}^{\text{Prior}}\ \overbrace{P(B|A)}^{\text{Likelihood}}}{\underbrace{P(B)}_{\text{Preposterior}}}$$

- **A represents variable of interest**
- **B represents observation**
- Prior → Our belief of A before we see B
- Likelihood → Probability of observing B given A
- Posterior → Updated belief of A after observing B
- Preposterior → Probability of observing B

Example: Coin Flip

- I have a fair coin and a two-headed coin.
- I choose one of them with equal probability and I flip it.
- Given that I flipped a head, what is the probability that I chose the two-headed coin?

Example: Coin Flip – Solutions

1. Define the Bayesian updating objective:

Posterior → $P(choose\ two\ headed\ coints \mid H)$

2. State the Bayesian equation to determine the components needed for updating:

$$P(choose\ two\ headed\ coints \mid H) = \frac{P(choose\ two\ headed\ coins)\ P(H|choose\ two\ headed\ coins)}{P(H)}$$

Prior → $P(choose\ two\ headed\ coins) = 0.5$

Likelihood → $P(H|choose\ two\ headed\ coins) = 1$

Preposterior →$P(H) = P(H, choose\ two\ headed\ coins) + P(H, choose\ fair\ coins)$

$$P(H) = 1 * 0.5 + 0.5 * 0.5 = 0.75$$

Example: Coin Flip – Solutions

3. Use Bayesian Updating Equation

$$P(choose\ two\ headed\ coints \mid H) = \frac{P(choose\ two\ headed\ coins)\ P(H|choose\ two\ headed\ coins)}{P(H)}$$

$$P(choose\ two\ headed\ coints \mid H) = \frac{0.5 * 1}{0.75} = \frac{2}{3}$$

Bayesian in Continuous Space (1)

- Bayesian formula we discussed so far is for updating **discrete probability distribution**
- However, many real world problems are often **continuous**
  - Revenue % increase can take any values from -inf to inf
  - Conversion ratio can take any values from 0 to 1
- For continuous distribution, we need to update the probability distribution, not the probability estimate

$$P(\theta|Data) = \frac{P(\theta)\, P(Data|\theta)}{P(Data)} \longrightarrow f(\theta|Data) = \frac{f(\theta)\, f(Data|\theta)}{\int_\theta \, f(\theta)\, f(Data|\theta) d\theta}$$

Bayesian in Continuous Space (2)

$$f(\theta|Data) = \frac{f(\theta)\, f(Data|\theta)}{\int_{\theta}\, f(\theta)\, f(Data|\theta) d\theta}$$

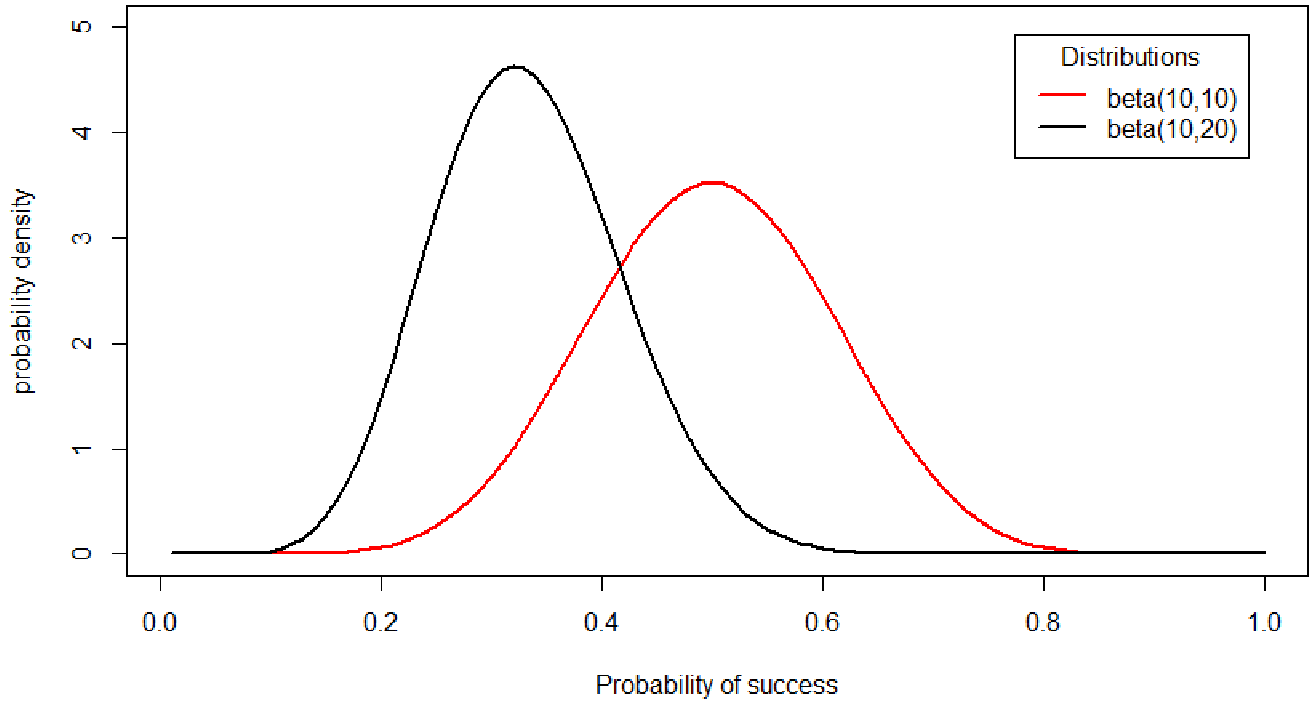
- The denominator of Bayes' Theorem (the integral part) makes the Bayesian updating difficult.
  - Intractable solutions
  - Computationally expensive
- For this reason, many clever algorithms are developed to efficiently for Bayesian updating
  - **Metropolis → Discussed briefly today**
  - Metropolis-Hastings
  - Gibbs Sampling
  - Hamiltonian Monte Carlo
- Whenever we can: Find tractable (or analytical) posterior
  - **Conjugate Prior → Discussed briefly today**

Conjugate Prior (1)

There are several prior-likelihood pairs that play nice to each other. This property is called ***conjugacy***.

Example:

- Beta distribution is a commonly used probability distribution for describing probability of an event.
- $Beta(\alpha, \beta)$



Conjugate Prior (2)

- Let's imagine we have a coin
- What is the probability that the coin is fair?
  - The coin looks just like other coin
  - Fair coin generally seems to have probability around 0.5 of coming up heads
  - However, we are not quite certain that P(head) is 0.5

- Now, let's flip the coin 10 times.
- we observed 9 heads and 1 tail
- What is the probability that the coin is fair?

- Posterior = $P(heads\ |9\ heads\ and\ 1\ tail)$ = ?

Conjugate Prior (3)

- Let coming up head as π
- Describe prior → $f(\pi)$

$$f(\pi) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} \pi^{\alpha-1}(1-\pi)^{\beta-1}$$

- Describe likelihood → n flips, y heads → binomial $p(y|\pi)$~$f(y|\pi)$

$$f(y|n,\pi) = \binom{n}{y} \pi^{y}(1-\pi)^{n-y}$$

Conjugate Prior (4)

Side note:
Kernel of a PDF or PMF is a function that proportional to the distribution itself

- Posterior through Bayes' Theorem

$$\left.\begin{aligned} f(\pi|y) &= \frac{\boldsymbol{f(\pi)f(y|\pi)}}{f(y)} \\ f(\pi|y) &\propto \boldsymbol{f(\pi)f(y|\pi)} \end{aligned}\right\}$$

This is called kernel

$$\left.\begin{aligned} f(\pi|y) &= \boldsymbol{\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}\pi^{\alpha-1}(1-\pi)^{\beta-1}\binom{n}{y}\pi^{y}(1-\pi)^{n-y}} \\ f(\pi|y) &\propto \boldsymbol{\pi^{\alpha-1}(1-\pi)^{\beta-1}\pi^{y}(1-\pi)^{n-y}} \end{aligned}\right\}$$

This is called kernel

$$f(\pi|y) = \pi^{\alpha+y-1}(1-\pi)^{\beta+n-y-1}$$

$$f(\pi|y) = Beta(\alpha+y, \beta+n-y)$$

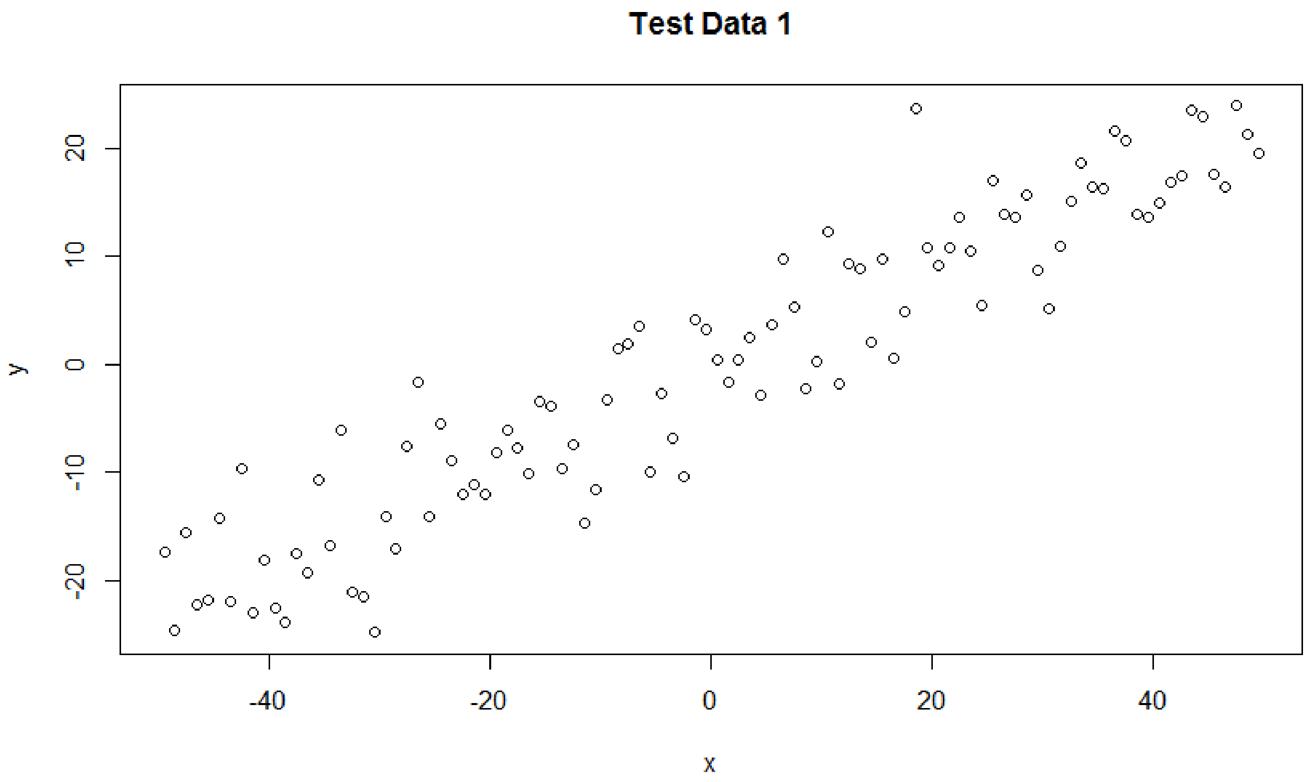
MCMC: Markov Chain Monte Carlo

- For cases where there is no conjugacy, we need to find a ways to directly sample the posterior
- The most common algorithm for this is using MCMC algorithm
  - Markov chain
    - Probability of state at t+1 only depend on t (memoryless)
  - Monte Carlo simulation
- The goal of MCMC is to produce dependent samples from posterior, $P(\theta|data)$
- This presentation will only discuss Metropolis/Metropolis-Hastings algorithm

MCMC - Linear Regression Example (1)

- Imagine data generating process from the following linear regression:

$$y = ax + b + N(0, \sigma)$$



MCMC - Linear Regression Example (2)

- We have the data.
- We want to generate the **posterior** for the regression constants (a, b) and the noise (standard deviation,σ)
- End result:
  - $P(a|data)$
  - $P(b|data)$
  - $P(\sigma|data)$

MCMC - Linear Regression Example (3)

Requirements:

- Define prior distribution for each parameter
- Define likelihood function given the proposed parameter
- Define the kernel of the posterior:
  - $posterior = prior * likelihood$
- Define proposal distribution (aka jumping distribution):
  - This is the distribution that proposed the new value for the posterior.
  - For Metropolis algorithm, this has to be symmetric (e.g. normal distribution)
- We are now ready to initiate Metropolis algorithm

MCMC - Linear Regression Example (4)

Steps:

- Define initial guess for each parameter
- Take one sample from proposed distribution
- Calculate the kernel of the posterior for the $posterior_{current}$ and $posterior_{proposed}$
- Calculate $\mathrm{r} = \min(posterior_{proposed}/posterior_{current}, 1)$
- If $r > U(0,1)$: set $posterior_{current} = posterior_{proposed}$
- Else: set $posterior_{current} = posterior_{current}$
- Repeat for N iterations

Remember to throw away the first Z iterations for burn in process.

Useful algorithms developed From Bayes' Theorem

- Hierarchical Modelling
- Kalman Filter / Ensemble Kalman Filter
- Bayesian Network
  - Directed acyclic graph
  - Influence diagram
- Bayesian A/B test
- Bayesian Deep Learning
  - Advantage of using Bayesian: Interpretability, cons: large computing requirements
  - Advantage of using deep learning: accurate, efficient learner, cons: low interpretability (statistical properties are not fully understood)
- And many more algorithms

Library / Tools for MCMC

- BUGS (Bayesian Inference using Gibbs Sampling) → http://www.mrc-bsu.cam.ac.uk/software/bugs/
  - First generation MCMC sampler
  - Tricky implementation in linux
  - Old and inefficient
- JAGS (Just Another Gibbs Sampler) → http://mcmc-jags.sourceforge.net/
  - Rebuild from scratch
  - Has many useful features that makes gibbs sampling easy to be implemented
  - Fast and efficient algorithm
  - Most commonly used library
- STAN (Named after Stanislaw Ulam) → http://mc-stan.org/
  - Bayesian figureheads starting to move to STAN
  - Use Hamiltonian Monte Carlo which purported to be more efficient compared to Gibbs sampling
  - Andrew Gelman, John Kruschke has recommended to use STAN