Module 1: Introduction to Bayesian Statistics

Rebecca C. Steorts

Agenda

- Motivations
- Traditional inference
- Bayesian inference
- Bernoulli, Beta
- Posterior of Bernoulli-Beta
- Conjugacy
- Example with 2012 election data
- No sleep example
- Marginal likelihood
- Posterior Prediction

What should you learn?

You should learn the main principles of Bayesian inference/prediction and how to apply these to real data analysis. You will continue with this in lab/homework to make sure that you understand these key principles.

Traditional inference

You are given **data** X and there is an **unknown parameter** you wish to estimate θ

How would you estimate θ?

- Find an unbiased estimator of θ.
- Find the maximum likelihood estimate (MLE) of θ by looking at the likelihood of the data.
- If you cannot remember the definition of an unbiased estimator or the MLE, review these before our next class.

Bayesian inference

Bayesian methods trace its origin to the 18th century and English Reverend Thomas Bayes, who along with Pierre-Simon Laplace discovered what we now call **Bayes' Theorem**

- $p(x \mid \theta)$ likelihood
- $p(\theta)$ prior
- $p(\theta \mid x)$ posterior
- $p(x)$ marginal distribution

Derive the posterior distribution of $p(\theta \mid x)$.

Bernoulli distribution

The Bernoulli distribution is very common due to binary outcomes.

- Consider flipping a coin (heads or tails).
- We can represent this a binary random variable where the probability of heads is θ and the probability of tails is $1-\theta$.

The write the random variable as $X \sim \text{Bernoulli}(\theta)\mathbb{1}(0 < \theta < 1)$

It follows that the likelihood is

$$p(x \mid \theta) = \theta^x (1-\theta)^{(1-x)} \mathbb{1}(0 < \theta < 1).$$

- Exercise: what is the mean and the variance of X?

Bernoulli distribution

- Suppose that $X_1, \dots, X_n \overset{iid}{\sim} \text{Bernoulli}(\theta)$. Then for $x_1, \dots, x_n \in \{0, 1\}$ what is the likelihood?

Notation

- $\propto$: means "proportional to"
- $x_{1:n}$ denotes $x_1, \ldots, x_n$

Likelihood

$$
\begin{aligned}
p(x_{1:n}|\theta) &= \mathbb{P}(X_1 = x_1, \ldots, X_n = x_n \mid \theta) \\
&= \prod_{i=1}^{n} \mathbb{P}(X_i = x_i \mid \theta) \\
&= \prod_{i=1}^{n} p(x_i|\theta) \\
&= \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i} \\
&= \theta^{\sum x_i}(1-\theta)^{n-\sum x_i}.
\end{aligned}
$$

Beta distribution

Given $a, b > 0$, we write $\theta \sim \text{Beta}(a, b)$ to mean that θ has pdf

$$p(\theta) = \text{Beta}(\theta|a, b) = \frac{1}{B(a, b)}\theta^{a-1}(1-\theta)^{b-1}\mathbb{1}(0 < \theta < 1),$$

i.e., $p(\theta) \propto \theta^{a-1}(1-\theta)^{b-1}$ on the interval from 0 to 1.

- Here,
$$B(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$$
.
- The mean is $E(\theta) = \int \theta\, p(\theta) d\theta = a/(a+b)$.

Posterior of Bernoulli-Beta

Lets derive the posterior of $\theta \mid x_{1:n}$

$$
\begin{aligned}
& p(\theta|x_{1:n}) \\
& \propto p(x_{1:n}|\theta)p(\theta) \\
& = \theta^{\sum x_i}(1-\theta)^{n-\sum x_i}\frac{1}{B(a,b)}\theta^{a-1}(1-\theta)^{b-1}I(0<\theta<1) \\
& \propto \theta^{a+\sum x_i - 1}(1-\theta)^{b+n-\sum x_i - 1}I(0<\theta<1) \\
& \propto Beta\big(\theta \mid a+\sum x_i,\, b+n-\sum x_i\big).
\end{aligned}
$$

Conjugacy

What do you notice about the prior and the posterior from the Bernoulli-Beta example that we just considered?

Conjugacy

If the posterior distribution comes from the same family of distributions as the prior, we say that the prior and posterior are conjugate distributions.

More formally, the prior is called a conjugate family for the likelihood funtion.

Example

- The Gaussian family is conjugate to itself with respect to a Gaussian likelihood.
- That is, if the likelihood function is Gaussian, choosing a Gaussian prior over the mean will ensure that the posterior distribution is also Gaussian.
- We will return to conjugacy later.
- First, we will look at a simple example to illustrate the idea of the Bernoulli-Beta.

Approval ratings of Obama

What is the proportion of people that approve of President Obama in PA?

- We take a random sample of 10 people in PA and find that 6 approve of President Obama.
- The national approval rating (Zogby poll) of President Obama in mid-September 2015 was 45%. We'll assume that in PA his approval rating is approximately 50%.
- Based on this prior information, we'll use a Beta prior for θ and we'll choose a and b.
- Let's first look at a simple choice of just setting the prior.

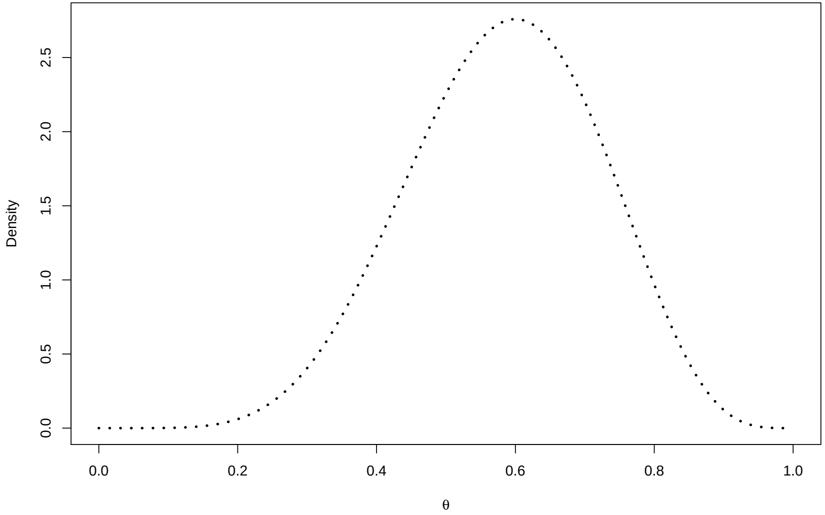
Obama Example

```r
n = 10
# Fixing values of a,b.
# I've chosen the prior on Beta to be skewed
a = 21/8
b = 0.04
th = seq(0,1, length=500)
x = 6

# we set the likelihood, prior, and posteriors with
# THETA as the sequence that we plot on the x-axis.
# Beta(c,d) refers to shape parameter
like = dbeta(th, x+1, n-x+1)
prior = dbeta(th, a, b)
post = dbeta(th, x+a, n-x+b)
```

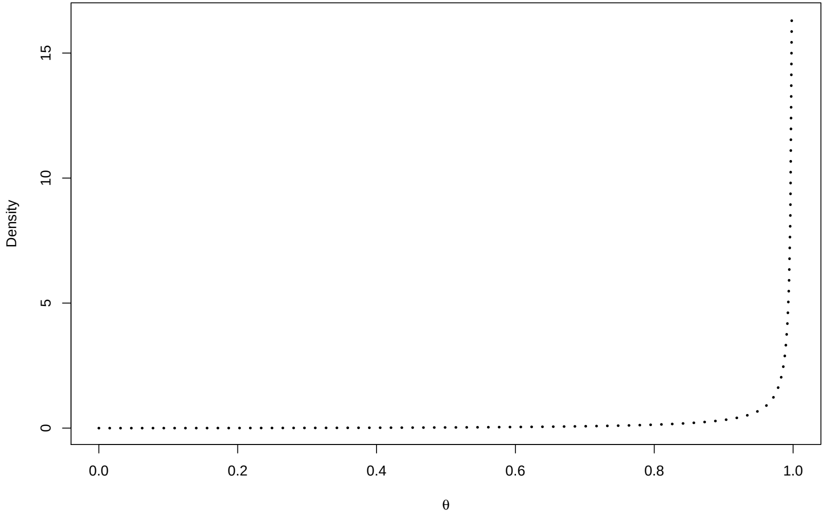
Likelihood

```r
plot(th, like, type='l', ylab = "Density",
     lty = 3, lwd = 3, xlab = expression(theta))
```



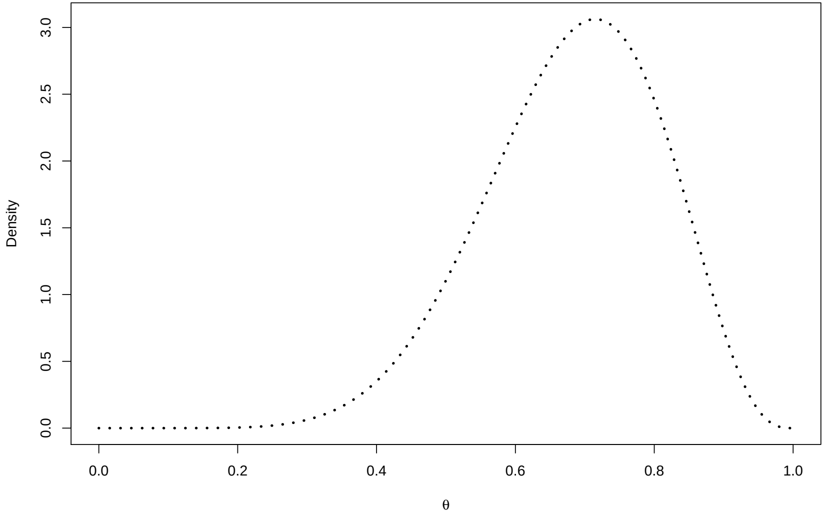
Prior

```r
plot(th, prior, type='l', ylab = "Density",
     lty = 3, lwd = 3, xlab = expression(theta))
```

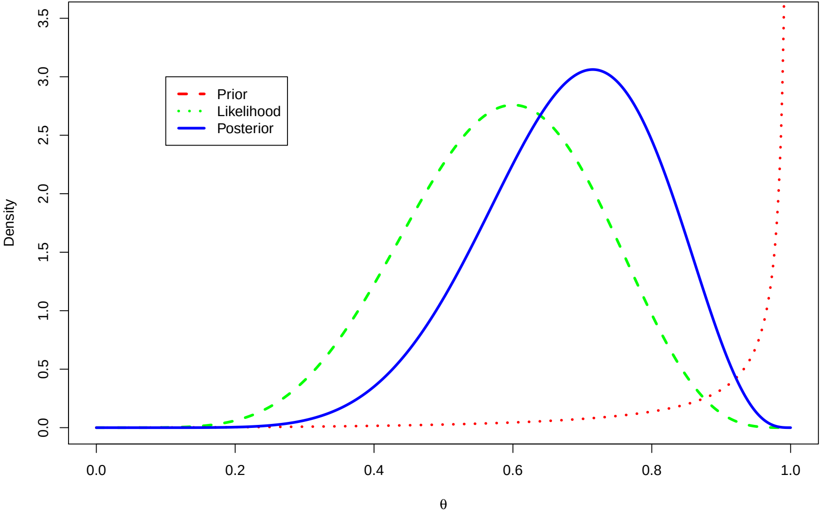


Posterior

```r
plot(th, post, type='l', ylab = "Density",
     lty = 3, lwd = 3, xlab = expression(theta))
```



Likelihood, Prior, and Posterior



Back to the Prior

- We choose the prior here very arbitrarly due to very little information of "subjective knowledge."
- Let's consider an example where we have more infomation and can set a, b from such subjective information.

How Much Do You Sleep Example

We are interested in a population of American college students and the proportion of the population that sleep at least eight hours a night, which we denote by θ.

How Much Do You Sleep Example

- *The Gamecock*, at the USC printed an internet article “College Students Don’t Get Enough Sleep" (2004).
  - Most students spend six hours sleeping each night.
- 2003: University of Notre Dame’s paper, *Fresh Writing*.
  - The article reported took random sample of 100 students:
  - “approximately 70% reported to receiving only five to six hours of sleep on the weekdays,
  - 28% receiving seven to eight,
  - and only 2% receiving the healthy nine hours for teenagers."

How Much Do You Sleep

- Have a random sample of 27 students is taken from UF.
- 11 students record that they sleep at least eight hours each night.
- Based on this information, we are interested in estimating θ.

How Much Do You Sleep

- From USC and UND, believe it's probably true that most college students get less than eight hours of sleep.
- Want our prior to assign most of the probability to values of $\theta < 0.5$.
- From the information given, we decide that our best guess for θ is 0.3, although we think it is very possible that θ could be any value in $[0, 0.5]$.

Our Model

Our model can be summarized by the Binomial-Beta distribution

$$X|\theta \sim \text{Binomial}(n, \theta) \tag{1}$$
$$\theta \sim \text{Beta}(a, b) \tag{2}$$

You can show that the posterior of

$$\theta \mid X \sim \text{Beta}(x + a, n - x + b)$$

Choice of a,b for Beta Prior

- Given this information, we believe that the median of θ is 0.3 and the 90th percentile is 0.5.
- Knowing this allows us to estimate the unknown values of a and b.
- How do we actually calculate a and b?

Choice of a,b for Beta Prior

We would need to solve the following equations:

$$\int_0^{0.3} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \theta^{a-1}(1-\theta)^{b-1}\, d\theta = 0.5$$
$$\int_0^{0.5} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \theta^{a-1}(1-\theta)^{b-1}\, d\theta = 0.9$$

In non-calculus language, this means the 0.5 quantile (50th percentile)$_=$0.3. The 0.9 quantile (90th percentile) = 0.5.

The equations are written as percentiles above!

- We can easily solve this numerically in `R` using a numerical solver `BBsolve` using the `BB` package. .
- The documentation for this package is not great, so beware.

How Much Do You Sleep

```
#load the BB package
library(BB)

## using percentiles
myfn <- function(shape){
    test <- pbeta(q = c(0.3, 0.5), shape1 = shape[1],
     shape2 = shape[2]) - c(0.5, 0.9)
    return(test)
    }
BBsolve(c(1,1), myfn)
```

```
##   Successful convergence.
## $par
## [1] 3.263743 7.185121
##
## $residual
## [1] 5.905161e-08
##
```

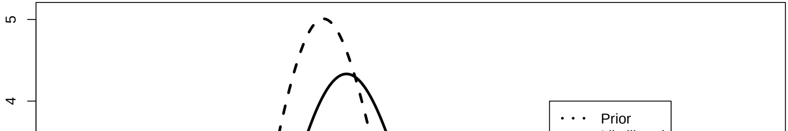
How Much Do You Sleep

Using our calculations from the Beta-Binomial our model is

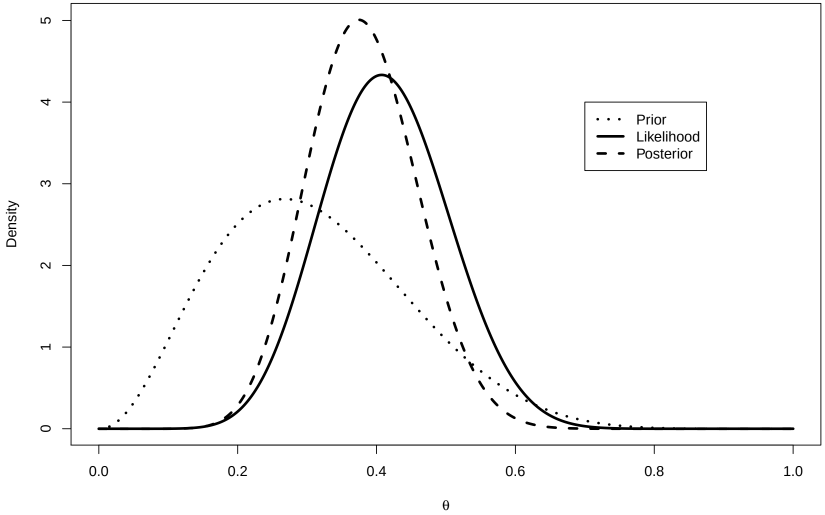
$$
\begin{aligned}
X \mid \theta &\sim \mathrm{Binomial}(27, \theta) \\
\theta &\sim \mathrm{Beta}(3.3, 7.2) \\
\theta \mid x &\sim \mathrm{Beta}(x + 3.3, 27 - x + 7.2) \\
\theta \mid 11 &\sim \mathrm{Beta}(14.3, 23.2)
\end{aligned}
$$

How Much Do You Sleep

```
th = seq(0,1,length=500)
a = estimated$par[1]
b = estimated$par[2]
n = 27
x = 11
prior = dbeta(th,a,b)
like = dbeta(th,x+1,n-x+1)
post = dbeta(th,x+a,n-x+b)
plot(th,post,type="l",ylab="Density",lty=2,lwd=3,
xlab = expression(theta))
lines(th,like,lty=1,lwd=3)
lines(th,prior,lty=3,lwd=3)
legend(0.7,4,c("Prior","Likelihood","Posterior"),
lty=c(3,1,2),lwd=c(3,3,3))
```



How Much Do You Sleep



Cast of characters

- Observed data: x
- Note this could consist of many data points, e.g., $x = x_{1:n} = (x_1, \ldots, x_n)$.

likelihood	$p(x \mid \theta)$
prior	$p(\theta)$
posterior	$p(\theta \mid x)$
marginal likelihood	$p(x)$
posterior predictive	$p(x_{n+1} \mid x_{1:n})$
loss function	$\ell(s, a)$
posterior expected loss	$\rho(a, x)$
risk / frequentist risk	$R(\theta, \delta)$
integrated risk	$r(\delta)$

Marginal likelihood

The **marginal likelihood** is

$$p(x) = \int p(x|\theta)p(\theta)\, d\theta$$

- What is the marginal likelihood for the Bernoulli-Beta?

Example: Back to the Bernoulli-Beta

Suppose

$$\theta \sim Beta(a, b)$$

and

$$X_1, \ldots, X_n \mid \theta \overset{iid}{\sim} Bernoulli(\theta)$$

Example: Back to the Bernoulli-Beta

Then the marginal likelihood is

$$
\begin{aligned}
p(x_{1:n})
&= \int p(x_{1:n}|\theta)p(\theta)\,d\theta \\
&= \int_0^1 \theta^{\sum x_i}(1-\theta)^{n-\sum x_i}\frac{1}{B(a,b)}\theta^{a-1}(1-\theta)^{b-1}d\theta \\
&= \frac{1}{B(a,b)}\int_0^1 \theta^{\sum x_i+a-1}(1-\theta)^{n-\sum x_i+b-1}d\theta \\
&= \frac{B\left(a+\sum x_i,\, b+n-\sum x_i\right)}{B(a,b)}\int_0^1 \frac{\theta^{\sum x_i+a-1}(1-\theta)^{n-\sum x_i+b-1}}{B\left(a+\sum x_i,\, b+n-\sum x_i\right)}d\theta \\
&= \frac{B\left(a+\sum x_i,\, b+n-\sum x_i\right)}{B(a,b)},
\end{aligned}
$$

by the integral definition of the Beta function.

Posterior predictive distribution

Let $a_n = a + \sum x_i$ and $b_n = b + n - \sum x_i$.

Recall that the posterior distribution is $p(\theta|x_{1:n}) = \text{Beta}(\theta|a_n, b_n)$.

Let's derive the posterior preditive distribution.

Posterior predictive distribution

- We may wish to predict a new data point x_{n+1}
- Assumption 1: Assume that $x_{1:(n+1)}$ are independent given θ

$$
\begin{aligned}
p(x_{n+1}|x_{1:n}) &= \int p(x_{n+1},\theta|x_{1:n})\,d\theta \\
&\int \frac{p(x_{n+1},\theta,x_{1:n})}{p(x_{1:n})}\,d\theta \quad \text{(Conditional probability)} \\
&\int \frac{p(x_{n+1}|\theta,x_{1:n})p(\theta|x_{1:n})p(x_{1:n})}{p(x_{1:n})}\,d\theta \quad \text{(Product rule)} \\
&= \int p(x_{n+1}|\theta,x_{1:n})p(\theta|x_{1:n})\,d\theta \\
&= \int p(x_{n+1}|\theta)p(\theta|x_{1:n})\,d\theta \quad \text{By Assumption 1.}
\end{aligned}
$$

Posterior predictive distribution

The posterior predictive can be derived to be

$$\begin{aligned}\mathbb{P}(X_{n+1} = 1 \mid x_{1:n}) &= \int \mathbb{P}(X_{n+1} = 1 \mid \theta) p(\theta | x_{1:n}) d\theta \\ &= \int \theta \, \mathrm{Beta}(\theta | a_n, b_n) d\theta \\ &= \frac{a_n}{a_n + b_n} \quad \text{(Mean of Beta distribution)}.\end{aligned}$$

Similarly,

$$\mathbb{P}(X_{n+1} = 0 \mid x_{1:n}) = 1 - \mathbb{P}(X_{n+1} = 1 \mid x_{1:n}) = \frac{b_n}{a_n + b_n}.$$

Posterior predictive distribution (continued)

This implies that

$$p(x_{n+1}|x_{1:n}) = \begin{cases} \frac{a_n}{a_n+b_n} & \text{if } x_{n+1} = 1 \\ \frac{b_n}{a_n+b_n} & \text{if } x_{n+1} = 0 \end{cases}$$

Hence, the posterior predictive p.m.f. is

$$p(x_{n+1}|x_{1:n}) = \frac{a_n^{x_{n+1}} b_n^{1-x_{n+1}}}{a_n + b_n} \mathbb{1}(x_{n+1} \in \{0,1\}).$$

Overall Summary

- We covered the "cast of characters" needed to work with Bayesian models
- These include the likelihood, prior, posterior, marginal likelihood, and posterior predictive distribution
- We derived Bayes' Theorem
- Bernoulli-Beta
- Conjugacy

Background Knowledge

- Familiar with Discrete and Continuous Distributions
- Can calculate expectations and variances
- Change of variables
- Mean squared error
- Sufficiency
- Confident calculating the likelihood and log-likelihood
- Confident in working with partial derivatives
- Familiar maximizing or minimizing functions (and proving they are global max/min)

Detailed Summary for Exam

- Bayes Theorem
- Likelihood
- Prior
- Posterior derivation
- Marginal likelihood
- Posterior predictive distribution
- Conjugacy
- Proportionality
- Understanding when models are appropriate for data given to you (Ex: Approval ratings for Obama)
- What is an informative prior
- What is a non-informative prior
- Proper posterior
- How do you incorporate a pilot study into your posterior analysis (Ex: See sleep study)

Exercise

We write $X \sim \text{Poisson}(\theta)$ if X has the Poisson distribution with rate $\theta > 0$, that is, its p.m.f. is

$$p(x|\theta) = \text{Poisson}(x|\theta) = e^{-\theta}\theta^{x}/x!$$

for $x \in \{0, 1, 2, \dots\}$ (and is 0 otherwise). Suppose $X_1, \dots, X_n \overset{iid}{\sim} \text{Poisson}(\theta)$ given θ, and your prior is

$$p(\theta) = \text{Gamma}(\theta|a, b) = \frac{b^a}{\Gamma(a)}\theta^{a-1}e^{-b\theta}\mathbb{1}(\theta > 0).$$

What is the posterior distribution on θ?

Solution

Since the data is independent given θ, the likelihood factors and we get

$$
\begin{aligned}
p(x_{1:n}|\theta) &= \prod_{i=1}^{n} p(x_i|\theta) \\
&= \prod_{i=1}^{n} e^{-\theta}\theta^{x_i}/x_i! \\
&\underset{\theta}{\propto} e^{-n\theta}\theta^{\sum x_i}.
\end{aligned}
$$

Solution

Thus, using Bayes' theorem,

$$\begin{aligned} p(\theta|x_{1:n}) &\propto p(x_{1:n}|\theta)p(\theta) \\ &\propto e^{-n\theta}\theta^{\sum x_i}\theta^{a-1}e^{-b\theta}\mathbb{1}(\theta > 0) \\ &\propto e^{-(b+n)\theta}\theta^{a+\sum x_i - 1}\mathbb{1}(\theta > 0) \\ &\propto \text{Gamma}\left(\theta \mid a + \textstyle\sum x_i,\ b + n\right). \end{aligned}$$

Therefore, since the posterior density must integrate to 1, we have

$$p(\theta|x_{1:n}) = \text{Gamma}\left(\theta \mid a + \textstyle\sum x_i,\ b + n\right).$$

Module 1 Derivations

Class notes from Module 1 can be found below:

https://github.com/resteorts/modern-bayes/tree/master/lectures
ModernBayes20/lecture-1/01-class-notes