SPR day 03

Beta distributions and the Beta-Bernoulli joint model

Readings: PRML 1.2.3 Bayesian probabilities

2.1 Beta distribution
Gamma functions

Content:

- (1) Gamma Functions
- (2) Beta Distribution
- (3) Beta-Bernoulli joint model and its posterior
- (4) Maximum a-posteriori (MAP) Estimation
- (5) Posterior predictive

Gamma Function $\Gamma(x): \mathbb{R}_+ \to \mathbb{R}$

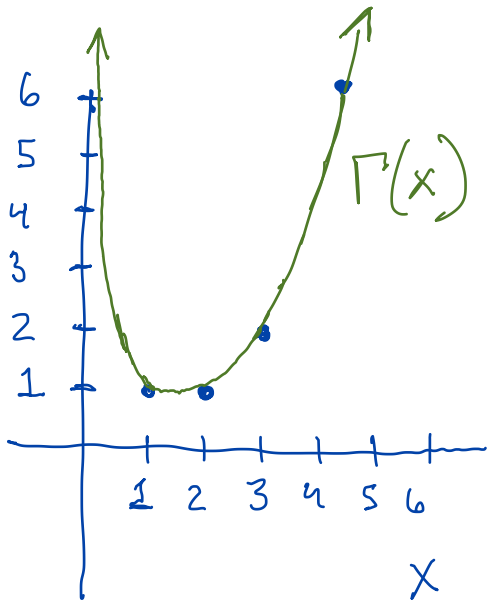
Recall the factorial function: $n! = n \cdot (n-1) \cdot (n-2) \cdots 1$

It only takes integer input.

Can we generalize it to take continuous real input?

Yes! We get a common function known as Gamma function

x	$(x-1)!$	$\Gamma(x)$
0.0		$+\infty$
0.5		0.57
1.0	1	1.0
1.5		0.88
2.0	1	1.0
3.0	2	2.0
4.0	6	6.0
5.0	24	24.0
6.0	120	120.0
7.0	720	720.0
⋮	⋮	⋮



See Jupyter notebook
link on Schedule

Formally:

$$\Gamma(a) = \int_0^\infty u^{a-1} e^{-u} du$$

Use standard implementation: SciPy

Numerical Stability:

Use $\log \Gamma(a)$ `scipy.special.gammaln`

Beta Distribution

Random variable μ

Sample space interval between 0 and 1

$0 \leq \mu \leq 1$

Parameters $a > 0$, $b > 0$

PDF function

$$\text{BetaPDF}(\mu | a, b) = \underbrace{\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}}_{\text{"normalizing constant," is constant wrt r.v. } \mu} \underbrace{\mu^{a-1}(1-\mu)^{b-1}}_{\text{interesting part}}$$

$$= c(a,b)\, \mu^{a-1}(1-\mu)^{b-1}$$

Useful facts:

$$\int_0^1 \text{BetaPDF}(\mu | a, b)\, d\mu = 1$$

because pdf must integrate to one over sample space

$$\int_0^1 c(a,b)\, \mu^{a-1}(1-\mu)^{b-1}\, d\mu = 1$$

by definition of the BetaPDF

$$\int_0^1 \mu^{a-1}(1-\mu)^{b-1}\, d\mu = \frac{1}{c(a,b)} = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$$

moving term const wrt μ outside integral

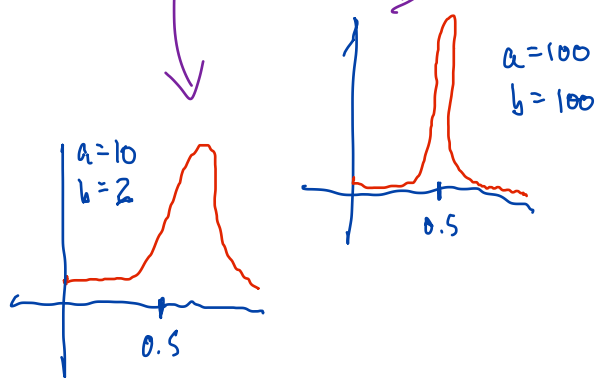
Beta Distribution: Visuals

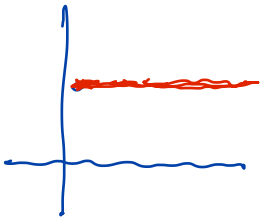
See Jupyter Notebook link on Schedule

How does shape of BetaPDF depend on $a>0$ $b>0$

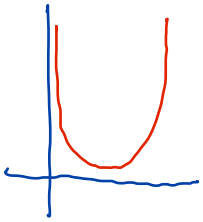
When $a>1$, $b<1$, mode of $\mu \to 1$

When $a>1$ and $b>1$, a unique mode exists: $\hat{\mu} = \frac{a-1}{a+b-2}$



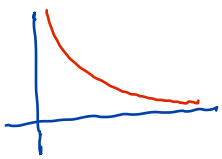


Special Case
$a=1$
$b=1$
uniform



When $a<1$, $b<1$
multiple modes

When $a<1$, $b>1$
mode at $\mu \to 0$



$+\infty$

heads very likely

a 1.0

heads unlikely

0.0

0.0 ← 1.0 → $+\infty$

b

tails unlikely

tails very likely

Beta-Bernoulli Joint Model

Define model over two random variables:

$\mu \in [0, 1]$ coin frequency of heads

$x_1, x_2, \cdots x_N$ with $x_n \in \{0, 1\}$ flip outcomes of N tosses

Joint distribution:

$$p(x_1, \cdots x_N, \mu) = \left[\prod_{n=1}^{N} \text{BernPMF}(x_n | \mu) \right] \cdot \text{BetaPDF}(\mu | a, b)$$

"likelihood" generates data

"prior" generate parameter

Given joint (our model), we can ask about probabilities implied by model:

			how to calculate from joint?
Posterior	$p(\mu \mid x_1 \cdots x_N)$	after seeing N flip outcomes, what is likely value of μ?	via Bayes rule
Evidence	$p(x_1 \cdots x_N)$	what is probability of seeing these N flip outcomes?	via sum rule
Predictive Posterior	$p(x_N \mid x_1 \cdots x_{N-1})$	after seeing N-1 flips, will next coin be heads?	via Bayes rule and sum rule

Posterior of μ for Beta-Bernoulli

$$p(\mu \mid x_1 \cdots x_N) = \frac{1}{p(x_1 \cdots x_N)} \left[\prod_{n=1}^{N} \text{BernPMF}(x_n \mid \mu) \right] \text{BetaPDF}(\mu \mid a, b)$$

by Bayes rule

$$= \frac{1}{p(x_1 \cdots x_N)} \cdot \left[\prod_{n=1}^{N} \mu^{x_n} (1-\mu)^{1-x_n} \right] \cdot c(a,b) \cdot \mu^{a-1} (1-\mu)^{b-1}$$

by substitution of definitions of PMFs

$$= \frac{c(a,b)}{p(x_1 \cdots x_N)} \mu^{\sum_n x_n} (1-\mu)^{\sum_n (1-x_n)} \cdot \mu^{a-1} (1-\mu)^{b-1}$$

because $\mu^s \cdot \mu^t = \mu^{s+t}$ for any s, t

$$= \text{const} \cdot \mu^{(H(x)+a-1)} (1-\mu)^{(T(x)+b-1)}$$

where $H(x) = \sum_n x_n$ Count #heads

$T(x) = \sum_n (1-x_n)$ #tails

we know total tosses is fixed $N = H(x) + T(x)$

We have written the posterior PDF so it looks like

$$\text{const} \cdot \mu^{\hat{a}-1} (1-\mu)^{\hat{b}-1}$$

Must be Beta distribution

$\hat{a} = a + H(x)$

$\hat{b} = b + T(x)$

Posterior over μ for Beta-Bern is Beta

How do we know?

Posterior PDF must integrate to one.

$\hat{a} > 0$, $\hat{b} > 0$ defined on prev. page

$$\int \underbrace{c(\hat{a}, \hat{b})}_{\text{unknown const}} \mu^{\hat{a}-1} (1-\mu)^{\hat{b}-1} d\mu = 1$$

to be a valid pdf

$$\int \mu^{\hat{a}-1} (1-\mu)^{\hat{b}-1} d\mu = \frac{1}{c(\hat{a}, \hat{b})}$$

multiply both sides by $\frac{1}{c(a,b)}$

$$\frac{\Gamma(\hat{a})\Gamma(\hat{b})}{\Gamma(\hat{a}+\hat{b})} = \frac{1}{c(\hat{a}, \hat{b})}$$

by the "useful" fact we proved on page 2

Thus, by inspection, our pdf is a Beta PDF

$$p(\mu | x_1 \cdots x_N) = \text{Beta}(\mu | \hat{a}, \hat{b})$$

$\hat{a} = a + \#\text{heads}$

$\hat{b} = b + \#\text{tails}$

A general pattern:

If you have an unknown distribution with pdf function up to a constant:

$$p(\theta | \lambda) = \text{const} \cdot f(\theta, \lambda)$$

and you know a distribution family D with pdf

$$p_D(\theta | \lambda) = c_D(\lambda) \cdot f(\theta, \lambda)$$

f same as above

c_D is a known function

then your distribution must belong to family D

MAP Estimator for μ

Suppose we observe N flip outcomes, $x_1 \cdots x_N$.

1) How to estimate a value for μ?
2) How to predict the next flip x_{N+1}?

One answer: MAP estimation.

MAP = maximum a posteriori = most likely value of μ by posterior density

1) Find $\hat{\mu}$ that is most likely under posterior $p(\mu | x_1 \cdots x_N)$
2) Predict next flip using likelihood with $\mu = \hat{\mu}$

$$p(x_{N+1}) = \text{BernPMF}(x_{N+1} | \hat{\mu})$$

Optimization problem: MAP

$$\hat{\mu} = \arg\max_{\mu \in [0,1]} p(\mu | x_1 \cdots x_N)$$

$$= \arg\max_{\mu \in [0,1]} \log \left[\frac{p(\mu)\, p(x_1 \cdots x_N | \mu)}{p(x_1 \cdots x_N)} \right]$$

$$= \arg\max_{\mu \in [0,1]} \log p(\mu) + \sum_{n=1}^{N} \log \text{BernPMF}(x_n | \mu)$$

$$= \arg\max_{\mu \in [0,1]} \; (a-1)\log\mu + H(x)\log\mu + (b-1)\log(1-\mu) + T(x)\log(1-\mu)$$

$$= \arg\max_{\mu \in [0,1]} \; s\log\mu + r\log(1-\mu)$$

with $s = H(x) + a - 1$

$r = T(x) + b - 1$

We've solved this before! See ML estimator derivation, which showed by taking derivative, setting to zero, & solving we get

$$\hat{\mu} = \frac{s}{s+r} \quad \text{when} \quad \begin{matrix} s \geq 0 \\ r \geq 0 \end{matrix}$$

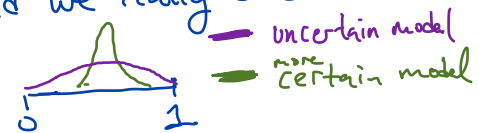
$$\text{thus } \hat{\mu} = \frac{H(x)+a-1}{\underbrace{H(x)+T(x)}_{=N \text{ \#flips}}+a+b-2}$$

requires
$H + a > 1$
$T + b > 1$
otherwise MAP does not exist

Problems with MAP

(1) does not always exist (e.g. if $N=0$, $a=\frac{1}{2}$, $b=\frac{1}{2}$)

(2) should we really condense whole distribution to one value?

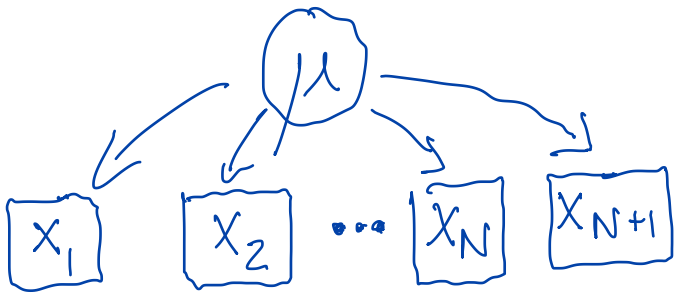


these have same MAP! but one is less confident.

should we act same in both cases?

Posterior Predictive to predict next coin

Recall model:



$$p(\mu) = \text{Beta}(\mu \mid a, b)$$

$$p(x \mid \mu) = \prod_n \text{Bern}(x_n \mid \mu)$$

Write down what we want conditioned on what's known

$$p(X_{N+1} = 1 \mid x_1 \cdots x_N)$$

Then rewrite in terms of all r.v.s

$$= \int p(X_{N+1} = 1, \mu \mid x_1 \cdots x_N) \, d\mu \quad \text{by sum rule}$$

$$= \int \text{Bern}(1 \mid \mu) \, \text{Beta}(\mu \mid \hat{a}, \hat{b}) \, d\mu$$

$$= c(\hat{a}, \hat{b}) \int \mu^1 \mu^{\hat{a}-1} (1-\mu)^{\hat{b}-1} \, d\mu$$

use Useful Fact ★ from page 3

$$= c_{\text{Beta}}(\hat{a}, \hat{b}) \frac{1}{c_{\text{Beta}}(\hat{a}+1, \hat{b})}$$

$$= \frac{\Gamma(\hat{a}+\hat{b})}{\Gamma(\hat{a})\Gamma(\hat{b})} \cdot \frac{\Gamma(\hat{a}+1)\Gamma(\hat{b})}{\Gamma(\hat{a}+\hat{b}+1)}$$

$$= \frac{\Gamma(\hat{a}+\hat{b})}{\Gamma(\hat{a})} \cdot \frac{\hat{a}\,\Gamma(\hat{a})}{(\hat{a}+\hat{b})\,\Gamma(\hat{a}+\hat{b})}$$

use identity $\Gamma(x+1) = x\,\Gamma(x)$ for all $x > 0$

$$= \frac{\hat{a}}{\hat{a}+\hat{b}}$$

Thus, by averaging over each possible μ, weighted by its posterior density,

$$p(X_{N+1}=1 \mid X_1, \cdots X_N) = \frac{\#\text{heads} + a}{N + a + b}$$

valid for all $a > 0$, $b > 0$.