Lec 6 2/20/18 Math 341

Previously

$\theta \sim P(\theta)$ the prior

X_1, X_2, X_3 observed

We want to know the distribution of X_4 given our (a) prior and (b) data

$$P(X_4 \mid X_1, X_2, X_3) = \int_{\Theta_0} P(X_4, \theta \mid X_1, X_2, X_3) = \sum_{\theta \in \Theta_0} P(X_4, \theta \mid X_1, X_2, X_3)$$

$P(A, B) = P(A \mid B)\, P(B)$

$P(A, B \mid C) = P(A \mid B, C)\, P(B \mid C)$

joint = cond $\times$ marginal

Looks messy but it's the same as

$$P(Y) = \sum_{X \in Supp(X)} P(X, Y) \quad \text{marginalizing out } X$$

$$= \sum_{\theta \in \Theta_0} \underbrace{P(X_4 \mid \theta, X_1, X_2, X_3)}_{\substack{= \\ P(X_4 \mid \theta)}} \underbrace{P(\theta \mid X_1, X_2, X_3)}_{\text{posterior}}$$

why?

Pf: since $X_1 \mid \theta, X_2 \mid \theta, X_3 \mid \theta, X_4 \mid \theta \overset{iid}{\sim} Bern(\theta)$

$$P(X_4 \mid \theta, X_1, X_2, X_3) \overset{\text{Def. cond prob.}}{=} \frac{P(\theta, X_4, X_1, X_2, X_3)}{P(\theta, X_1, X_2, X_3)} \overset{\text{def of cond. prob.}}{=} \frac{P(X_4, X_1, X_2, X_3 \mid \theta)\, \cancel{P(\theta)}}{P(X_1, X_2, X_3 \mid \theta)\, \cancel{P(\theta)}} = \frac{P(X_4 \mid \theta)\, \cancel{P(X_1 \mid \theta)\, P(X_2 \mid \theta)\, P(X_3 \mid \theta)}}{\cancel{P(X_1 \mid \theta)\, P(X_2 \mid \theta)\, P(X_3 \mid \theta)}} = P(X_4 \mid \theta)$$

$$P(X_4 \mid X_1, X_2, X_3) = \int_{\Theta_0} P(X_4 \mid \theta)\, P(\theta \mid X_1, X_2, X_3) = \sum_{\theta \in \Theta_0} P(X_4 \mid \theta)\, P(\theta \mid X_1, X_2, X_3)$$

How to parse this?

this is the same as $X_4 \mid \theta \sim Bern(\theta)$ but θ is a weighted averaged over all θ's in the posterior, weighted by the posterior.

$\neq P(X_4 \mid \hat{\theta}_{MLE})$ is the best done with the frequentist's MLE. $X_1 = 0, X_2 = 0, X_3 = 0$? $\hat{\theta}_{MLE} = 0 \Rightarrow X_4 \sim Deg(0)$ BAD!

Recall the iid Bernoulli model

with $\Theta_0 = \{0.25, 0.75\}$ $\quad P(\theta) = U(\Theta_0) = \begin{cases} 0.25 \text{ wp } 0.5 \\ 0.75 \text{ wp } 0.5 \end{cases}$

principle of indifference

DATA: $X_1 = 0,\ X_2 = 1,\ X_3 = 1$

What is Bayesian pt. estimate of θ? $\quad \hat{\theta}_{MAP} = \underset{\theta \in \Theta_0}{\text{argmax}} \{P(\theta \mid X)\} = \text{argmax}\{P(X \mid \theta) P(\theta)\}$

since $P(\theta)$ const. under princ. of indifference

$= \text{argmax}\{P(X \mid \theta)\}$

$= 0.75$

most probable θ after seeing data in the support of the prior, Θ_0

But $0.75 \neq 0.66 = \hat{\theta}_{MLE} = \underset{\theta \in \Theta}{\text{argmax}} \{\mathcal{L}(\theta; x)\} = \underset{\theta \in \Theta}{\text{argmax}} \{P(X; \theta)\}$

Why not? $\quad \underbrace{\Theta_0 \neq \Theta}$

likely a bad idea - why should you put zero prob on pieces of the parameter space?

So let's say $\Theta_0 = \Theta = (0,1)$, the parameter space of the Bernoulli

What prior should we put on this? Via the principle of indifference, we use the

$\theta \sim \underbrace{U(0,1)}_{\text{std. uniform}}$ i.e. the "uniform prior". It is now continuous

Recall its density $\quad P(\theta) = \begin{cases} 1 & \text{if } \theta \in (0,1) \\ 0 & \text{o/t} \end{cases}$

no special privilege to any $\theta \in \Theta$

$X = \langle 0,1,1 \rangle$. Let's get best guess

constant as function of θ

$$\hat{\theta}_{MAP} = \underset{\theta \in (0,1)}{\text{argmax}} \{P(\theta \mid X)\} = \text{argmax}\left\{\frac{P(X \mid \theta) P(\theta)}{P(X)}\right\} = \text{argmax}\{P(X \mid \theta)\} = \text{argmax}\{\theta^2(1-\theta)\}$$

constant as function of θ

Same problem as solved before:

$$\frac{d}{d\theta}\left[\theta^2(1-\theta)\right] = \frac{d}{d\theta}\left(\theta^2-\theta^3\right) = 2\theta - 3\theta^2 = 0 \Rightarrow \hat{\theta}_{MAP} = \frac{2}{3} = \hat{\theta}_{MLE}$$

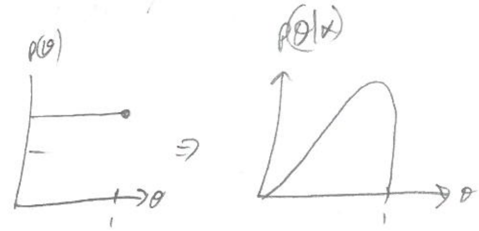
New Question.... What if we're interested in

$$P\left(\theta \in [0.6, 0.7] \mid X = \langle 0,1,1\rangle\right)$$

How likely is the true θ between 0.6 and 0.7 after we see the data.

Frequentist answer: 0 or 1. Not possible!

Bayesian answer: $= \int_{0.6}^{0.7} P(\theta \mid X = \langle 0,1,1\rangle)\, d\theta$



Let's try to solve for the posterior now.

$$P(\theta|X) = \frac{P(X|\theta)\,\overbrace{P(\theta)}^{=1}}{P(X)} = \frac{P(X|\theta)}{\underset{\overset{\|}{?}}{P(X)}} = \frac{P(X|\theta)}{\underbrace{\int_{\Theta} P(X|\theta)\overbrace{P(\theta)}^{=1}\, d\theta}_{\text{Bayes thm (marginalizing over } \theta)}} = \frac{P(X|\theta)}{\int_0^1 P(X|\theta)\, d\theta}$$

$$= \frac{\theta^2(1-\theta)}{\int_0^1 \theta^2(1-\theta)\, d\theta} = \frac{\theta^2(1-\theta)}{\int_0^1 (\theta^2-\theta^3)\, d\theta} = \frac{\theta^2(1-\theta)}{\left[\frac{\theta^3}{3}\right]_0^1 - \left[\frac{\theta^4}{4}\right]_0^1} = \frac{\theta^2(1-\theta)}{\underbrace{\frac{1}{3}-\frac{1}{4}}_{\frac{1}{12}}} = 12\,\theta^2(1-\theta)$$

$$\Rightarrow \int_{0.6}^{0.7} 12\,\theta^2(1-\theta)\, d\theta = 12\left[\frac{\theta^3}{3} - \frac{\theta^4}{4}\right]_{0.6}^{0.7} = 0.1765$$ assuming the prior of indifference

---

Let's solve for the posterior generally $P(\theta|X)$ where $X = \langle x_1, \ldots, x_n\rangle$ see n data pts.

and $\theta \sim U(0,1)$

$$P(\theta|X) = \frac{P(X|\theta)\,P(\theta)}{P(X)} \overset{\text{"1"}}{} = \frac{\prod_{i=1}^{n} P(X_i|\theta)}{\int_{\Theta} P(X|\theta)\,P(\theta)\,d\theta} = \frac{\prod_{i=1}^{n} P(X_i|\theta)}{\int_0^1 \prod_{i=1}^{n} P(X_i|\theta)\,d\theta}$$

$\Theta = (0,1)$, $P(\theta) = 1$

$$= \frac{\theta^{\Sigma x_i}(1-\theta)^{n-\Sigma x_i}}{\underbrace{\int_0^1 \theta^{\Sigma x_i}(1-\theta)^{n-\Sigma x_i}\,d\theta}_{\text{Famous Integral, a "special function"}}} = \underbrace{\frac{\theta^{\Sigma x_i}(1-\theta)^{n-\Sigma x_i}}{B(\Sigma x_i+1,\ n-\Sigma x_i+1)}}_{\text{this is a PDF of a Beta random r.v}} = \text{Beta}(\Sigma x_i+1,\ n-\Sigma x_i+1)$$

the Beta is the posterior of the iid Bernoulli likelihood and the uniform prior

$$B(\alpha,\beta) := \int_0^1 t^{\alpha-1}(1-t)^{\beta-1}\,dt$$

$$Y \sim \text{Beta}(\alpha,\beta) = \frac{1}{B(\alpha,\beta)}\, y^{\alpha-1}(1-y)^{\beta-1} \qquad \text{Supp}[Y] = (0,1)$$

Is this a valid PDF?

$$\int_0^1 \frac{1}{B(\alpha,\beta)}\, y^{\alpha-1}(1-y)^{\beta-1}\,dy = \frac{1}{B(\alpha,\beta)}\underbrace{\int_0^1 y^{\alpha-1}(1-y)^{\beta-1}\,dy}_{B(\alpha,\beta)} = \frac{B(\alpha,\beta)}{B(\alpha,\beta)} = 1$$

Param space:

when is $\int_0^1 y^{\alpha-1}(1-y)^{\beta-1}\,dy$ finite?

If $\alpha = 0$ $\quad y^{\alpha-1} = y^{-1}$ its a hyperbola around $y=0 \Rightarrow$ diverge!

If $\beta = 0$ $\quad (1-y)^{\beta-1} = y^{-1}$ " " " " " " " " " $y=1 \Rightarrow$ diverge

the divergence ~~is~~ worse as α, β grow negative

$$\Rightarrow \boxed{\alpha > 0,\ \beta > 0}$$

$$E(Y) = \int_0^1 y \frac{1}{B(\alpha,\beta)} y^{\alpha-1}(1-y)^{\beta-1}\,dy = \frac{1}{B(\alpha,\beta)}\int_0^1 y^{\alpha}(1-y)^{\beta-1}\,dy$$

$$= \frac{B(\alpha+1,\beta)}{B(\alpha,\beta)}$$

To simplify this, we need to introduce another famous integral:

$$\Gamma(\alpha) = \int_0^\infty t^{\alpha-1}e^{-t}\,dt \qquad \text{"the gamma function"} \quad \forall \alpha > 0$$

we need two facts

① By integration by parts, you can show that

$$\Gamma(\alpha+1) = \alpha\,\Gamma(\alpha)$$ the gamma function is the generalization of the factorial function, $\alpha!$

② $B(\alpha,\beta) = \dfrac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$ which can be proven using ch. of var's in mult. var. calc.

$$E(Y) = \frac{B(\alpha+1,\beta)}{B(\alpha,\beta)} = \frac{\dfrac{\Gamma(\alpha+1)\cancel{\Gamma(\beta)}}{\Gamma(\alpha+1+\beta)}}{\dfrac{\Gamma(\alpha)\cancel{\Gamma(\beta)}}{\Gamma(\alpha+\beta)}} = \frac{\dfrac{\alpha\,\cancel{\Gamma(\alpha)}}{(\alpha+\beta)\cancel{\Gamma(\alpha+\beta)}}}{\dfrac{\cancel{\Gamma(\alpha)}}{\cancel{\Gamma(\alpha+\beta)}}} = \boxed{\frac{\alpha}{\alpha+\beta}}$$

$Var(Y) = \ldots$ similar tricks employed (on HW)

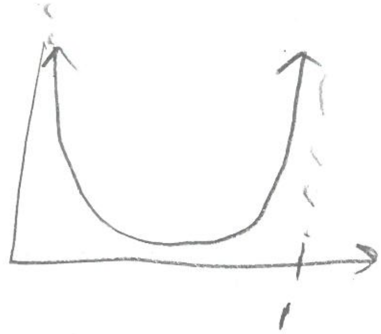
$$Mode(Y) = \underset{y\in(0,1)}{\operatorname{argmax}}\left\{\frac{1}{B(\alpha,\beta)}y^{\alpha-1}(1-y)^{\beta-1}\right\} = \underset{y\in(0,1)}{\operatorname{argmax}}\left\{y^{\alpha-1}(1-y)^{\beta-1}\right\} = \text{take log}\ldots$$

$$= \underset{y\in(0,1)}{\operatorname{argmax}}\left\{(\alpha-1)\ln(y) + (\beta-1)\ln(1-y)\right\} \Rightarrow \frac{\alpha-1}{y} - \frac{\beta-1}{1-y} = 0$$

$$\Rightarrow \frac{1-y}{y} = \frac{\beta-1}{\alpha-1} \Rightarrow \frac{1}{y} - 1 = \frac{\beta-1}{\alpha-1} \Rightarrow \frac{1}{y} = \frac{\beta-1+\alpha-1}{\alpha-1} \Rightarrow \boxed{Mode(Y) = \frac{\alpha-1}{\alpha+\beta-2}}$$

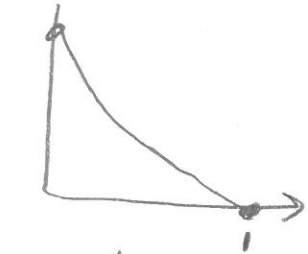
2nd deriv. check on HW

Shapes of the Beta r.v. density:

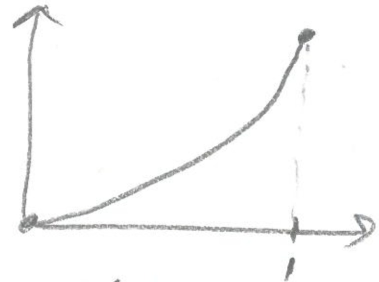


$\alpha = \beta = 0.5$

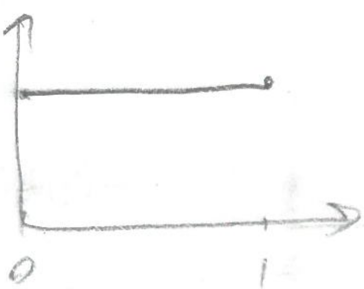
"arcsin distr"



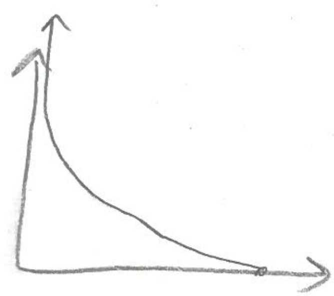
$\alpha = 1, \beta = 3$



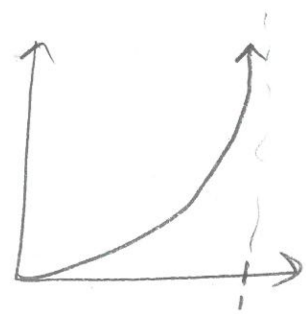
$\alpha = 5, \beta = 1$



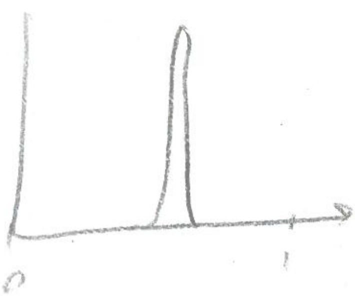
$\alpha = \beta = 1$



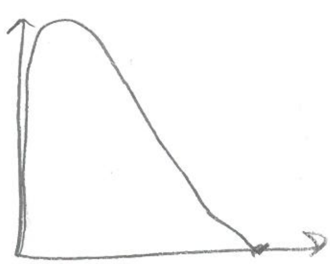
$\alpha = 0.99, \beta = 3$



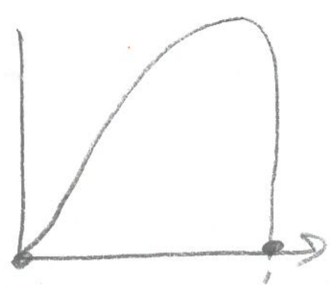
$\alpha = 5, \beta = 0.99$



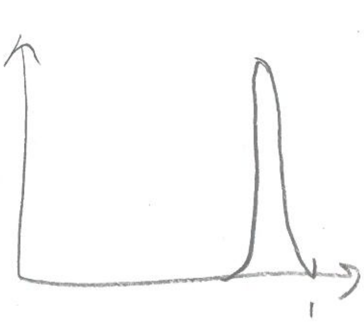
$\alpha = \beta = 100$



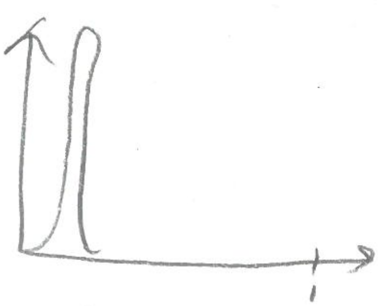
$\alpha = 1.01, \beta = 3$



$\alpha = 5, \beta = 1.01$



$\alpha = 100, \beta = 10$



$\alpha = 10, \beta = 100$