Lec 5 2/17/16 Math 390

For Later

Previously

$$p(\theta|x_1,\dots,x_n) = \frac{p(x_1,\dots,x_n|\theta)\,p(\theta)}{p(x_1,\dots,x_n)}$$

Assumptions? None needed here... But...

we know $X_1, X_2, X_3 \overset{iid}{\not\sim}$ b/c $p(X_{n+1}|X_1,\dots,X_n) = p(X_{n+1})$ i.e. no learning could occur!

What assumption(s) did we need to do our calc's??

$$= \frac{\prod_{i=1}^{n} p(x_i|\theta)\; p(\theta)}{\sum_{\theta' \in \Theta_0} p(x_1,\dots,x_n|\theta')\,p(\theta')}$$

conditional independence assumption

not needed but reasonable

$$= \sum_{\theta' \in \Theta_0} \prod_{i=1}^{n} p(x_i|\theta')\,p(\theta')$$

turns out if $p(X_1,\dots,X_n|\theta) = \prod_{i=1}^{n} p(X_i|\theta)$

$$\iff p(X_1,\dots,X_n) = p(X_{\pi(1)},\dots,X_{\pi(n)})$$

for any permutation π

exchangeability $\to$ most done in real world scientific applications

"deFinetti's Thm" (ch 3)

Order of $X_1, \dots, X_n$ does not matter. If we observe 0,1,1 or 101 or 110 we still have same posterior $P(\theta \mid 0,1,1)$.

So in a setup, we need either

(a) $X_1|\theta, \dots, X_n|\theta \overset{ind}{\sim}$ usually $X_1|\theta, \dots, X_n|\theta \overset{iid}{\sim}$

(b) $X_1, \dots, X_n \overset{exch}{\sim}$ usually all $\sim \mathcal{F}$

when a,b equivalent If data is exch. $\Rightarrow$ Bayesian model is guaranteed to exist but you may not know $P(X|\theta) \in \mathcal{F}$, $P(\theta) \in \mathcal{F}_0$

So let's return $X_1, X_2, X_3 \overset{exch}{\sim} Bern(\theta)$ $X_1 = 0$, $X_2 = 1$, $X_3 = 1$

$\Theta_0 = \{0.25, 0.75\}$

$\hat{\theta}_{MAP} = \text{argmax}\{p(\theta|x)\}$

$\hat{\theta}_{MMSE} = \mathbb{E}[\theta|x] \leftarrow$ yields θ's possibly not in Θ_0

$\hat{\theta}_{MAE} = \text{Med}[\theta|x]$

$\Rightarrow$ Want a prior that puts mass all throughout $\Theta = [0,1]$

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

known as an "uninformative" or "objective" prior

$$P(\theta|X_1,X_2,X_3) = \frac{P(X_1,X_2,X_3|\theta)\,P(\theta)}{P(X_1,X_2,X_3)} \quad \text{if } X_1=0, X_2=1, X_3=1$$

$$\Rightarrow = 12(1-\theta)\theta^2$$

What does X_* look like? (future data)

$$P(X_*|X_1,X_2,X_3) = \int_{\theta\in\Theta_0} P(X_*|\theta)\,P(\theta|X_1\overset{=0}{,}X_2\overset{=1}{,}X_3\overset{=1})\,d\theta = \int_0^1 \theta^{X_*}(1-\theta)^{1-X_*}\,12(1-\theta)\theta^2\,d\theta$$

$$= 12\underbrace{\int_0^1 \theta^{X_*+2}(1-\theta)^{2-X_*}\,d\theta}_{B(X_*+3,\,3-X_*)}$$

Beta Function

$$B(\alpha,\beta) := \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} \quad \text{s.t.} \quad \Gamma(x) := \int_0^\infty x^t e^{-x}\,dx \overset{(x-1)!\ \text{if } x\in\mathbb{N}}{=}$$

extension of "!" to $\mathbb{R}^+$

$$= 12\, \frac{\Gamma(X_4+3)\,\Gamma(3-X_4)}{\Gamma(6) = 5! = 120} = \frac{1}{10}(X_4+2)!\,(2-X_4)!$$

$$= \begin{cases} 0 & \text{wp } \frac{1}{10}\, 2!\,2! = 0.4 \\ 1 & \text{wp } \frac{1}{10}\, 3!\,1! = 0.6 \end{cases} = \text{Bern}(0.6)$$

Let's generalize... $\underbrace{X_1|\theta, \ldots, X_n|\theta}\overset{iid}{\sim} \text{Bern}(\theta)$ w/ prior $\theta \sim U(0,1)$

$X := \sum X_i \longleftarrow X|\theta \sim \text{Bin}(n, \theta)$

n doesn't appear here

$$P(\theta|x) = \frac{P(x|\theta)\,\underbrace{P(\theta)}_{=1}}{P(x) = \int_\Theta P(x|\theta)\,P(\theta)\,d\theta}$$

$$= \frac{\cancel{\binom{n}{x}}\,\theta^x(1-\theta)^{n-x}}{\int_\Theta^1 \cancel{\binom{n}{x}}\,\theta^x(1-\theta)^{n-x}\,d\theta}$$

$$= \frac{1}{B(x+1, n-x+1)}\,\theta^x(1-\theta)^{n-x} = \text{Beta}(x+1,\, n-x+1)$$

the Beta Distribution changes as a normal distribution...

$$X \sim \text{Beta}(\alpha, \beta) := \overbrace{\frac{1}{B(\alpha,\beta)}\, x^{\alpha-1}(1-x)^{\beta-1}}^{\text{PDF}} = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\, x^{\alpha-1}(1-x)^{\beta-1}$$

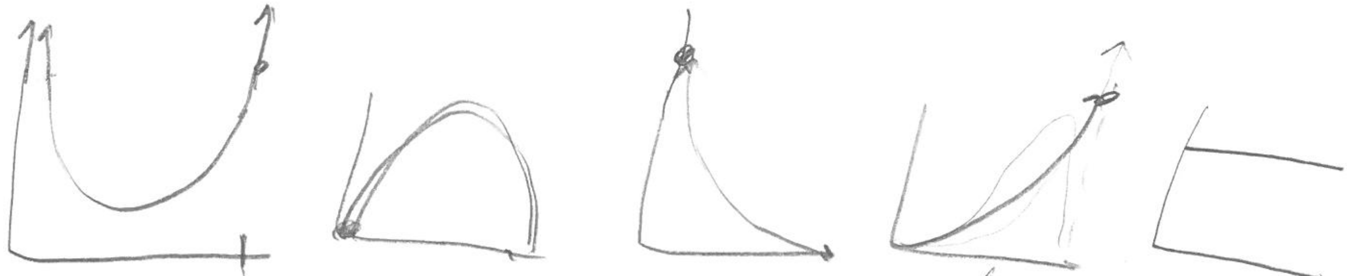
$\text{Supp}(X) = (0,1)$ Med(X) not available in closed form...

$$E(X) = \frac{\alpha}{\alpha+\beta} \quad \text{Var}(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} \quad \text{Mode}(X) = \frac{\alpha-1}{\alpha+\beta-2} \text{ if } \alpha,\beta > 1$$

$= 0.5$ if $\alpha = \beta$

Param Space

Beta distribution unbelievable... $\alpha \in (0,\infty)$, $\beta \in (0,\infty)$



"arcsin" $\alpha=\beta=0.5$ | $\alpha=\beta=2$ | $\alpha=1, \beta=3$ ↑ left | $\alpha=5, \beta=1$ ($\beta=0.99$, $\beta=1.01$) ↑ right | $\alpha=\beta=1$ uniform

(wikipedia)

Why not let $\theta \sim \text{Beta}(\alpha,\beta)$. The prior is now a beta.

$\alpha=\beta=1$ is called the "uninformative prior" i.e. $U(0,1)$

Let's see what happens...

$$P(\theta|x) = \frac{P(x|\theta)\,P(\theta)}{P(x)} = \frac{\cancel{\binom{n}{x}}\,\theta^x(1-\theta)^{n-x}\,\cancel{\frac{1}{B(\alpha,\beta)}}\,\theta^{\alpha-1}(1-\theta)^{\beta-1}}{\int_0^1 \cancel{\binom{n}{x}}\,\theta^x(1-\theta)^{n-x}\,\cancel{\frac{1}{B(\alpha,\beta)}}\,\theta^{\alpha-1}(1-\theta)^{\beta-1}\,d\theta}$$

$$= \frac{\theta^{x+\alpha-1}(1-\theta)^{n-x+\beta-1}}{\int_0^1 \theta^{x+\alpha-1}(1-\theta)^{n-x+\beta-1}\,d\theta}$$

$\alpha+\beta$ is "prior sample size" ⇑ $\frac{n}{n+\alpha+\beta}$???

data successes — data failures

$$= \frac{1}{B(x+\alpha, n-x+\beta)}\,\theta^{x+\alpha-1}(1-\theta)^{n-x+\beta-1} = \text{Beta}(\alpha+x, \beta+(n-x))$$

can be thought of as # prior successes — " failures

Prior is beta, posterior is beta!

$$\overset{\text{Beta}}{\downarrow} \qquad \overset{\text{Bin}}{\downarrow} \qquad \overset{\text{Beta}}{\downarrow}$$
$$p(\theta|x) \propto p(x|\theta)\, p(\theta)$$

"Beta" is the conjugate prior (i.e. posterior is in the same $\mathcal{F}$) for the Binomial likelihood. Normal! Warm, fuzzy feeling ... we will be looking at lots of conjugate priors this semester.

Predictive distribution

$$p(x_{n+1}|\underbrace{x_1,\ldots x_n}_{X}) = \int_{\Theta} p(x_{n+1}|\theta)\, p(\theta|\underbrace{x_1,\ldots x_n}_{X})\, d\theta$$

$$= \int_0^1 \theta^{x_{n+1}}(1-\theta)^{1-x_{n+1}} \frac{1}{B(x+\alpha, n-x+\beta)} \theta^{x+\alpha-1}(1-\theta)^{n-\alpha+\beta-1}\, d\theta$$

$$= \frac{1}{B(x+\alpha, n-x+\beta)} \int_0^1 \theta^{x_{n+1}+x+\alpha\cancel{+\alpha-1}}(1-\theta)^{n+x-\alpha+\beta\cancel{-1}-x_{n+1}}\, d\theta$$

$$= \frac{B(x_{n+1}+x+\alpha,\, n+1-x-x_{n+1}+\beta)}{B(x+\alpha,\, n-x+\beta)} = \frac{\dfrac{\Gamma(x_{n+1}+x+\alpha)\,\Gamma(n+1-x-x_{n+1}+\beta)}{\Gamma(n+1+\alpha+\beta)}}{\dfrac{\Gamma(x+\alpha)\,\Gamma(n-x+\beta)}{\Gamma(n+\alpha+\beta)}}$$

Use: $\Gamma(x+1) = \cancel{x! = \ldots} = x\Gamma(x)$

$$= \frac{\dfrac{\Gamma(x_{n+1}+x+\alpha)\,(n-x-x^*+\beta)\,\Gamma(n-x-x^*+\beta)}{(n+\alpha+\beta)\,\cancel{\Gamma(n+\alpha+\beta)}}}{\dfrac{\Gamma(x+\alpha)\,\Gamma(n-x+\beta)}{\cancel{\Gamma(n+\alpha+\beta)}}}$$

if $x_{n+1}=1$

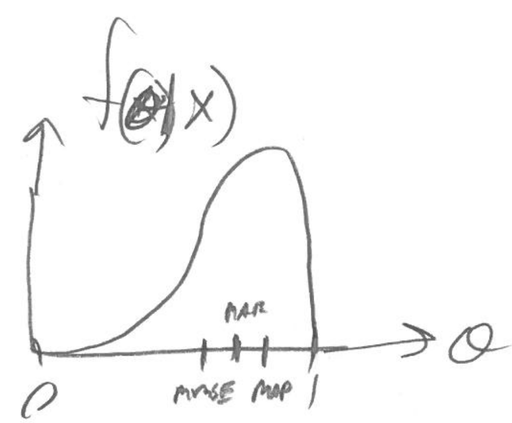
$$= \frac{\Gamma(1+x+\alpha)\,(n-x-1+\beta)\,\Gamma(n-x-1+\beta)}{(n+\alpha+\beta)\,\Gamma(x+\alpha)\,\Gamma(n-x+\beta)}$$

$$= \frac{(x+\alpha)\,\cancel{\Gamma(x+\alpha)}\,\cancel{(n-x-1+\beta)}\,\cancel{\Gamma(n-x-1+\beta)}}{(n+\alpha+\beta)\,\cancel{\Gamma(x+\alpha)}\,\cancel{(n-x-1+\beta)}\,\cancel{\Gamma(n-x-1+\beta)}}$$

$$\Rightarrow X^* \mid X \sim \text{Bern}\left(\frac{x+\alpha}{n+\alpha+\beta}\right)$$

Let's return to θ...

$$\theta \mid X \sim \text{Beta}(\alpha+x, \beta+n-x) \quad \text{e.g.}$$



$$\left.\begin{aligned} &\Rightarrow \hat{\theta}_{MAP} = \frac{\alpha+x-1}{\alpha+\beta+n-2} \\ &\Rightarrow \hat{\theta}_{MMSE} = \frac{\alpha+x}{\alpha+\beta+n} \\ &\Rightarrow \hat{\theta}_{MAE} \ \ldots \text{ need computer to get exactly} \end{aligned}\right\} \text{Answer the pt estimation question}$$

Pt estimation is less of an important goal. Because now we can make direct prob. statements on θ. E.g. Confidence Intervals

Recall frequentist CI's.. Pick α e.g. $\alpha = 5\%$ for a 95% CI

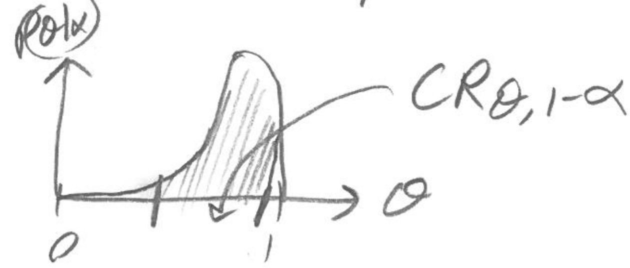
$X_1, \ldots, X_n \overset{iid}{\sim} \text{Bern}(\theta)$

Came from the fact

$$CI_{\theta, 1-\alpha} := \left[\hat{\theta} \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}}\right] \quad \text{that } \hat{\theta} = \hat{p} \sim N\left(\theta, \sqrt{\frac{\theta(1-\theta)}{n}}^{\,2}\right)$$

Recall that $P(\theta \in CI_{\theta,1-\alpha}) \overset{=\{0,1\}}{\neq} 1-\alpha$ frequentist conundrum

But look



Can we build a "Bayesian" "credible region" where θ has a $1-\alpha$ prob?

$P(\theta \in CR) = 1-\alpha$? Yes! What we always wanted to say!!!

Why not just take the "center" 95% of $P(\theta|x)$

$$CR_{\theta,1-\alpha} = \Big[\underbrace{\text{Quantile}\left[\theta|x, \tfrac{\alpha}{2}\right], \text{Quantile}\left[\theta|x, 1-\tfrac{\alpha}{2}\right]}_{\text{range of values for } \theta}\Big]$$

technically, a "2-sided credible region" or "2-sided Bayesian CI"

In our example, ...

$$\left[\text{Quantile}\left[\text{Beta}(\alpha+x, \beta+n-x), \tfrac{\alpha}{2}\right], \text{Quantile}\left[\text{Beta}(\alpha+x, \beta+n-x), 1-\tfrac{\alpha}{2}\right]\right]$$

Note:

$$F(x) = \frac{\Gamma(\alpha+\beta+n)}{\Gamma(\alpha+x)\Gamma(\beta+n-x)} \int_0^x t^{\alpha+x-1}(1-t)^{\beta+n-x-1}\,dt \overset{SET}{=} 2.5\%$$

for Beta

solve for x

i.e. only a computer can do this!

One of the reasons Bayesian stats did not become popular until now...

But is this a sensible idea? Consider a weird-looking posterior:

CR

But what about A U B U C U D U E U F U G

What's the advantage? Shorter! More parsimonious!

Define $\mu(A)$ to be the Lebesgue measure of a set A.

For our purposes $\mu([a,b]) = b-a$ i.e. defined as its length

$$\mu([a,b] \cup [c,d]) = (b-a)+(d-c)$$

s.t. $a<b<c<d$

Consider:

$$CR_{\theta, 1-\alpha} := \underset{A \subset \Theta}{\operatorname{argmin}} \{\mu(A) : P(\theta \in A \mid X) = 1-\alpha\}$$

AKA "highest density region" (HDR)

i.e. the shortest length interval that contains $1-\alpha$ of the prob.

Dis's

1. Tremendous computational problem

2. Not palatable to have a non-contiguous set for θ.

$CR = [0.01, 0.3] \cup [0.8, 0.84]$ i.e. low or high.

vs

$CR = [0.2, 0.82]$

For our class, first definition is the one we'll use