Lec 7 2/21/18 Math 341

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

Yesterday we learned that for $\mathcal{F}$ = iid Bernoulli, $\theta | X \sim \text{Beta}\left(\sum x_i + 1,\ n - \sum x_i + 1\right)$

~~We saw before for $X_1, \ldots, X_n$ from $\mathcal{F}$ = Bernoulli, only the sum~~

i.e. only the sum of the x's mattered. Let's consider just $\mathcal{F}$ = Binomial

Recall from Math 241:

$$X \sim \text{Binom}(n, \theta) := \binom{n}{x} \theta^x (1-\theta)^{n-x}$$

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

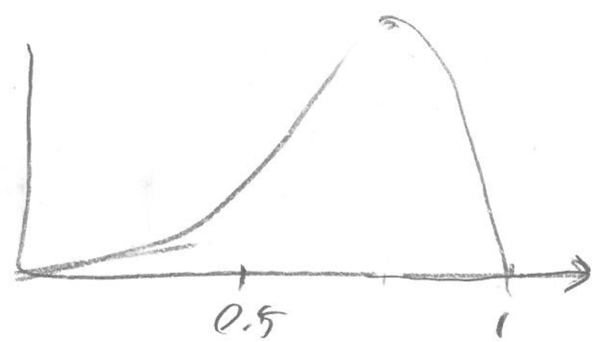
Same! From now on $\mathcal{F}$ = Binomial

$$P(\theta | x) = \frac{P(x|\theta)\, P(\theta)}{P(x)} = \frac{\cancel{\binom{n}{x}}\, \theta^x (1-\theta)^{n-x}}{\int_0^1 \cancel{\binom{n}{x}}\, \theta^x (1-\theta)^{n-x}} = \text{Beta}(\underbrace{x+1}_{\alpha'},\ \underbrace{n-x+1}_{\beta'})$$

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

if $n = 10$, $x = 7$

$$\theta | x \sim \text{Beta}(8, 4)$$



$$\hat{\theta}_{MAP} = \frac{\alpha' - 1}{\alpha' + \beta' - 2} = \frac{8-1}{8+4-2} = \frac{7}{10} = 0.7$$

What about the average $\theta | x$?

$$E[\theta | x] = \frac{\alpha'}{\alpha' + \beta'} = \frac{8}{8+4} = 0.66$$

Avg

$$E[\theta] = \frac{\alpha}{\alpha + \beta} = 0.5$$

what's this? Prior mean!!

What about the median $\theta | x$?

$\text{Med}[\theta | x]$ ... no closed form... in R... qbeta(0.5, 8, 4) $\approx$ 0.676...

Three different kinds of pt. estimation.

$$\hat{\theta}_{MAP} = \text{argmax}\{P(\theta|x)\} = \text{mode}[\theta|x] \quad \text{posterior mode}$$

$$\hat{\theta}_{MMSE} = E[\theta|x] \quad \text{posterior expectation / mean}$$

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

turns out... $\hat{\theta}_{MMSE}$ minimizes squared error loss

$$= \text{argmin}\, E\left[(\theta - \hat{\theta}_{MMSE})^2\right]$$

$\hat{\theta}_{MAE}$ minimizes abs err loss

$$= \text{argmin}\, E\left[\,|\theta - \hat{\theta}_{MAE}|\,\right]$$

We will be using all 3. Default is $\hat{\theta}_{MMSE} = E[\theta|x]$

but $\hat{\theta}_{MAP}$ is easiest to get! No need to compute $P(x)$!

New idea... priors...

$\theta \sim U(0,1) = \text{Beta}(1,1)$

Why not let

$\theta \sim \text{Beta}(\alpha, \beta)$

where I choose α, β to reflect my prior information!

Forget how I choose α, β for now...

Keep $\mathcal{F}$ = Binomial and let's see what happens!

recall...

$$\theta \sim \text{Beta}(\alpha,\beta) := \frac{1}{B(\alpha,\beta)}\,\theta^{\alpha-1}(1-\theta)^{\beta-1}\ ,\ E[\theta]=\frac{\alpha}{\alpha+\beta}\ ,\ \text{Var}(\theta)=\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

$$p(\theta|x)=\frac{p(x|\theta)\,p(\theta)}{\underbrace{p(x)}_{\int_{\Theta} p(x|\theta)p(\theta)\,d\theta}}=\frac{\binom{n}{x}\theta^x(1-\theta)^{n-x}\frac{1}{B(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}}{\int_0^1\binom{n}{x}\theta^x(1-\theta)^{n-x}\frac{1}{\cancel{B(\alpha,\beta)}}\theta^{\alpha-1}(1-\theta)^{\beta-1}\,d\theta}=\frac{\theta^{x+\alpha-1}(1-\theta)^{n-x+\beta-1}}{\int_0^1\theta^{x+\alpha-1}(1-\theta)^{n-x+\beta-1}\,d\theta}$$

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

$$\underset{\text{Beta}(\alpha,\beta)}{\theta}\ \xrightarrow{\ x\ }\ \underset{\text{Beta}(x+\alpha,\,n-x+\beta)}{\theta|x}$$

"Beta is the conjugate prior for the Binomial."

Posterior pt estimation

$$\hat\theta_{MMSE}=E[\theta|x]=\frac{x+\alpha}{n+\alpha+\beta}$$

$$\hat\theta_{MAP}=\text{Mode}[\theta|x]=\frac{x+\alpha-1}{n+\alpha+\beta-2}\quad\text{if } x+\alpha>1\ \&\ n-x+\beta>1$$

$$\hat\theta_{MAE}=\text{Med}[\theta|x]=\text{qbeta}(0.5,\ x+\alpha,\ n-x+\beta)$$

no closed form sol!

Now let's ask the question... the next x^*... what is its distr? $n^*=1$ (one trial)

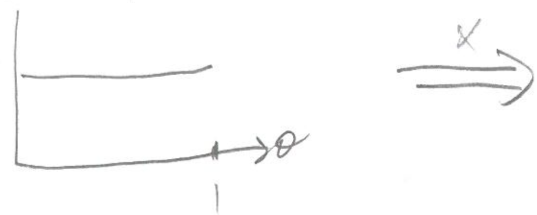
We know it must be Bernoulli since $Supp[X^*] = \{0,1\}$

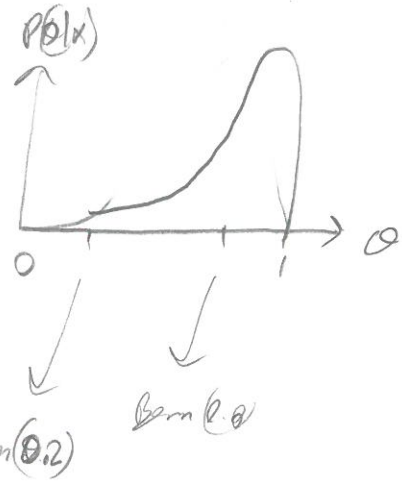
Recall:

$$P(X^*|X) = \int_{\Theta} P(X^*|\theta)\, P(\theta|X)\, d\theta$$ same as $$f(x|z) = \int_{Supp[Y]} f(x,y|z)\, dy$$

$$f(x,y|z) = f(x|y,z)\, f(y|z)$$

$\theta \sim U(0,1)$ $\alpha = \beta = 1$





$W \sim Bern(\theta)$

$P(W=1) = \theta$

Draw out many, many θ's (prop. to how likely these θ's are after seeing the data) and avg them

$$P(X^*|X) = \int_0^1 \left(\theta^{X^*}(1-\theta)^{1-X^*}\right) \frac{1}{B(x+\alpha, n-x+\beta)}\, \theta^{x+\alpha-1}(1-\theta)^{n-x+\beta-1}\, d\theta$$

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

$$= \frac{B(X^*+x+\alpha, -X^*+n-x+\beta+1)}{B(x+\alpha, n-x+\beta)}$$

this is the answer... but we must re-write it as a Bernoulli

Recall...
$\Gamma(x+1) = x\Gamma(x)$

$$P(X^*=1|X) = \frac{B(1+x+\alpha, n-x+\beta)}{B(x+\alpha, n-x+\beta)} = \frac{\frac{\Gamma(1+x+\alpha)\,\cancel{\Gamma(n-x+\beta)}}{\Gamma(1+n+\alpha+\beta)}}{\frac{\Gamma(x+\alpha)\,\cancel{\Gamma(n-x+\beta)}}{\Gamma(n+\alpha+\beta)}} = \frac{\frac{(x+\alpha)\,\cancel{\Gamma(x+\alpha)}}{(n+\alpha+\beta)\,\cancel{\Gamma(n+\alpha+\beta)}}}{\frac{\cancel{\Gamma(x+\alpha)}}{\cancel{\Gamma(n+\alpha+\beta)}}} = \frac{x+\alpha}{n+\alpha+\beta}$$

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

$$\uparrow$$

$$E[\theta|X]$$

Why does this make sense? The avg of many draws will be the average $\theta|X$!

Let's take a look at the posterior again

$$\text{Beta}(\alpha,\beta) \xrightarrow{x} \text{Beta}(\alpha + x,\ \beta + n - x)$$

x: # of data successes

$n - x$: # of data failures

α: # of prior successes (AKA "pseudosuccesses")

β: # of prior failures (AKA "pseudofailures")

α, β are "pseudocounts"

Conjugate prior parameters have interpretations as "pseudodata." They are as if you've seen data before!

$\theta \sim U(0,1) = \text{Beta}(1,1) \Rightarrow$ as if you've seen $\alpha = 1$ success previously
and " " " " " $\beta = 1$ failure previously

$\theta \sim U(0,1)$ is <u>not</u> devoid of information. The principle of indifference <u>is a statement</u> about a belief!

$E(\theta) = \frac{\alpha}{\alpha+\beta} = \frac{1}{1+1} = 0.5 \Rightarrow$ You believe that your prior uncertainty is centered around 0.5 (the prior expectation)

$\hat{\theta}_{MMSE} = E(\theta \mid x) = \frac{\alpha + x}{n + \alpha + \beta}$. Are $E(\theta)$ and $E(\theta \mid x)$ related??

$E[\theta] \qquad \bar{X} = \hat{\theta}_{MLE}$

$$\hat{\theta}_{MMSE} = E[\theta|x] = \frac{\alpha+x}{n+\alpha+\beta} = \frac{\alpha+\beta}{\alpha+\beta}\cdot\frac{\alpha}{n+\alpha+\beta} + \frac{x}{n+\alpha+\beta}\cdot\frac{n}{n}$$

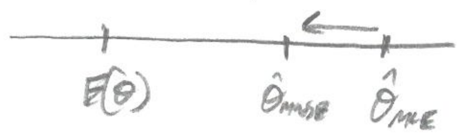
$$= \frac{\alpha+\beta}{n+\alpha+\beta} E[\theta] + \frac{n}{n+\alpha+\beta}\hat{\theta}_{MLE}$$

Note $\underbrace{\frac{\alpha+\beta}{n+\alpha+\beta}}_{\rho} + \underbrace{\frac{n}{n+\alpha+\beta}}_{1-\rho} = 1$

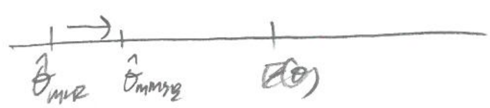
$$\hat{\theta}_{MMSE} = \rho\, E[\theta] + (1-\rho)\,\hat{\theta}_{MLE}$$

this is known as a "shrinkage estimator" because it shrinks the purely data-driven estimate toward the prior mean.

ρ is the shrinkage proportion



shrink down



shrink up

Larger vals of $\rho = \frac{\alpha+\beta}{n+\alpha+\beta}$ shrink "harder"

If α,β large compared to n ... $\Rightarrow$ "strong" prior or small sample size

" small " " " " $n \Rightarrow$ "weak" prior or large sample size

$n\uparrow \Rightarrow \rho\downarrow$

$\lim_{n\to\infty} \rho = 0$. In the limit, large sample sizes "drown out" the prior.

Goals of Inference

① Pt est for θ ✓

(1b) Prediction for future ✓ (for 1 obs)

② Conf. interval for θ ←

③ Testing values of θ