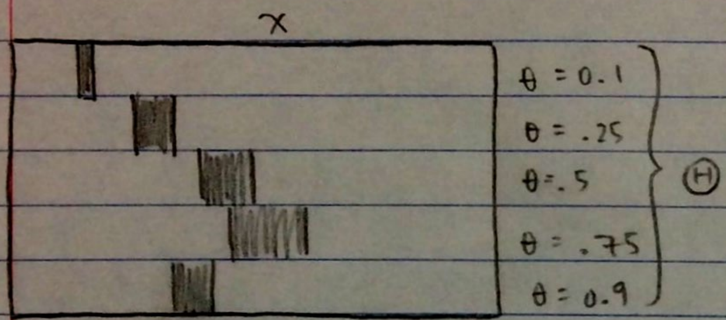
2/14/17

F = Bernoulli

$X = \langle 0, 1, 1 \rangle \Rightarrow n = 3$

$\Theta_0 = \{0.1, 0.25, 0.5, 0.75, 0.9\}$

$\theta \sim U(\Theta_0)$. we want $P(\theta | X)$.



~~$P(x|\theta)$~~

Prob/likelihood

$P(X | \theta = 0.1) = .009$

$P(X | \theta = .25) = .047$

$P(X | \theta = .5) = .125$

$P(X | \theta = .75) = .141$

$P(X | \theta = .9) = .061$

Idea to find "best" θ.

* $\hat{\theta}_{MAP} := \underset{\theta \in \Theta_0}{\text{argmax}} \{ P(\theta | X) \} = \text{argmax} \left\{ \frac{P(X|\theta)P(\theta)}{P(X)} \right\}$ bayes rule

↳ maximum of posterior (posterior mode)

$= \text{argmax} \{ P(X|\theta) P(\theta) \}$ because $P(X)$ is constant $\neq f(\theta)$

$= \text{argmax} \{ P(X|\theta) \}$ due to principle of indifference $P(\theta) \neq f(\theta)$

$= \hat{\theta}_{MLE}$

$$P(\theta | X) = P(X|\theta) \cdot \underbrace{P(\theta)}_{\text{scale by prior belief}} \cdot \underbrace{\frac{1}{P(X)}}_{\text{normalize constant so all strips sum to 1}}$$

$$P(\theta|X) = \frac{P(X|\theta)P(\theta)}{P(X)} = \frac{P(X|\theta)P(\theta)}{\sum_{\theta_0 \in \Theta_0} P(X, \theta_0)} = \frac{P(X|\theta)P(\theta)}{\sum_{\theta_0 \in \Theta_0} P(X|\theta_0)P(\theta_0)} \overset{\text{under principle of indifference}}{=} \frac{P(X|\theta)}{P(X|\theta_1) + \dots + P(X|\theta_n)}$$

$$P(\theta = .75 | X) = \frac{0.141}{.009 + .047 + .125 + .141 + .061} = \frac{.141}{.383} \approx 37\%$$

$\hat{\theta}_{MAP} = \hat{\theta}_{MLE}$

$.75 \neq .67$ why? → prior did not cover entire parameter space $\Theta_0 = \Theta = (0, 1)$

Main reason to be skeptical of Bayesianism:

- prior could be wrong.

$\Theta_0 = \{.25, .75\}$

$x = \langle 0,1,1 \rangle$ | $\theta \sim U(\Theta_0) \Rightarrow$ principle of indifference

$F = $ Bernoulli | x^* the 4^{th} future observation

"see" data sequentially

$x_1 = 0$

$$P(\theta = .25 \mid x_1 = 0) = \frac{P(x_1 = 0 \mid \theta = 0.25)}{P(x_1 = 0 \mid \theta = .25) + P(x_1 = 0 \mid \theta = .75)}$$

$$= \frac{.75}{.75 + .25} = .75.$$

$$P(\theta = .75 \mid x_1 = 0) = .25$$

Now $x_2 = 1$

Let's let our prior here be the posterior from previous data

$$P(\theta = .25 \mid x_2 = 1) = \frac{P(x_2 = 1 \mid \theta = .25)\, P(\theta = .25 \mid x_1 = 0)}{P(x_2 = 1 \mid \theta = .25) P(\theta = .25 \mid x_1 = 0) + \underbrace{P(x_2 = 1 \mid \theta = .25) P(\theta = .25 \mid x_1 = 0)}_{P(x_2)}}$$

$$= \frac{.25 \cdot .75}{.25 \cdot .75 + .75 \cdot .25} = .5$$

$$P(\theta = .75 \mid x_2 = 1) = .5$$

Now $x_3 = 1$

$$P(\theta = .25 \mid x_3 = 1) = \frac{P(x_3 = 1 \mid \theta = 0.25) P(\theta = .25 \mid x_1 = 0, x_2 = 1)}{P(x_3 = 1 \mid \theta = 0.25) P(\theta = .25 \mid x_1 = 0, x_2 = 1) + P(x_3 = 1 \mid \theta = .25) P(\theta = .25 \mid x_1 = 0, x_2 = 1)}$$

$$= \frac{.25 \cdot .5}{.25 \cdot .5 + .75 \cdot .5} = .25$$

$$P(\theta = .25 \mid x = \langle 0,1,1 \rangle) = .25$$

$$P(\theta \mid x_1, \ldots, x_n) = \frac{P(x_1, \ldots, x_n \mid \theta) P(\theta)}{P(x_1, \ldots, x_n)} \quad \text{bayes Rule}$$

$$= \frac{P(x_n \mid \theta) \cdots P(x_2 \mid \theta) P(x_1 \mid \theta) P(\theta)}{P(x_n, \ldots, x_2 \mid x_1)\, P(x_1)}$$

$$= \frac{P(x_2, \ldots, x_n \mid \theta) P(\theta \mid x_1)}{P(x_n, \ldots, x_2 \mid x_1)} = \frac{P(x_n \mid \theta) \cdots P(x_3 \mid \theta) P(x_1, x_2 \mid \theta) P(\theta)}{P(x_n, \ldots, x_3 \mid x_1, x_2)\, P(x_1, x_2)}$$

$$\frac{P(x_1, x_2 \mid \theta) P(\theta)}{P(x_1, x_2)} = P(\theta \mid x_1, x_2)$$

What is the distritribution of 4^{th}?

$\underbrace{X^*}_{\text{the next unseen observation}} \sim \text{Bern}(?)$

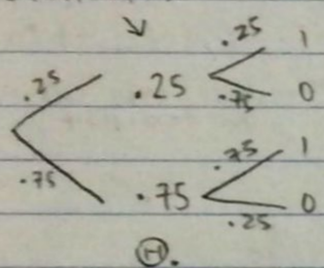
frequentist Approach:

Previously, $\underbrace{P(X^* | X_1, X_2, X_3)}_{\text{posterior predictive distribution}} \approx P(X^* | \theta = \hat{\theta}_{MLE}) = \text{Bern}(.67)$

↳ inaccurate

does not account for uncertainty in $\hat{\theta}_{MLE}$

$\theta \mid X_1, X_2, X_3$	X^*	$P(X^*, \theta \mid X_1, X_2, X_3)$
.25	1	.0625
.25	0	.1875
.75	1	.5625
.75	0	.1875



$P(X^* = 1 | X_1, X_2, X_3) = .625$

$\Rightarrow X^* | X_1, X_2, X_3 \sim \text{Bern}(.625)$

$$P(X^* | X_1, X_2, X_3) = \sum_{\theta \in \Theta_0} P(X^*, \theta | X_1, X_2, X_3) = \sum_{\theta \in \Theta_0} P(X^* | \theta, X_1, X_2, X_3) P(\theta | X_1, X_2, X_3)$$

if you know θ then it doesn't matter what you got previously

$$= \sum_{\theta \in \Theta_0} P(X^* | \theta)\, P(\theta | X_1, X_2, X_3)$$

$$= \sum_{\theta \in \Theta_0} P(X^* | \theta) \frac{P(X_1, X_2, X_3 | \theta) P(\theta)}{P(X_1, X_2, X_3)}$$

$P(\theta) = \sum_{x \in \text{supp}(X)} P(x, \theta)$

↳ marginalization

Procedure for posterior predictive distribution:

1. draw θ from posterior
2. examine $X^* | \theta$
3. repeat for all θ's and average

$$P(X^* | \theta) \overset{?}{=} P(X^* | \theta, X_1, X_2, X_3)$$

$$= \frac{P(X^*, X_1, X_2, X_3, \theta)}{P(X_1, X_2, X_3, \theta)}$$

$$= \frac{P(X^*, X_1, X_2, X_3 | \theta) \cancel{P(\theta)}}{P(X_1, X_2, X_3 | \theta) \cancel{P(\theta)}}$$

$$= \frac{P(X^* | \theta) \cancel{P(X_1 | \theta) P(X_2 | \theta) P(X_3 | \theta)}}{\cancel{P(X_1 | \theta) P(X_2 | \theta) P(X_3 | \theta)}}$$

$$P(x^* \mid x_1, \ldots, x_n) = \sum_{\theta \in \Theta_0} P(x^* \mid \theta)\, P(\theta \mid x_1, \ldots, x_n)$$

$$= \int_{\Theta_0} P(x^* \mid \theta)\, P(\theta \mid x_1, \ldots, x_n)\, d\theta \qquad \neq P(x^* \mid \hat{\theta})$$

$\hat{\theta}_{MAP} = \hat{\theta}_{MLE}$

if principle of indifference.

$0.75 \neq 0.67$

$\Theta_0 \neq \Theta = (0,1)$

What prior should we use?

$supp[\theta]$ = parameter space of $F = (0,1)$

An idea: $\theta \sim U(0,1) := 1$ if $\theta \in (0,1)$

continuous random variable

↳ any of these number have equal chance (non-zero probability)

$x = \langle 0, 1, 1 \rangle$

$$P(\theta \mid x) = \frac{P(x \mid \theta) \overset{1}{P(\theta)}}{\underbrace{P(x)}_{\to \text{constant}}} \propto P(x \mid \theta) = (1-\theta)(\theta)(\theta) = \theta^2 - \theta^3$$

If I care about the $\hat{\theta}_{MAP}$

$$\hat{\theta}_{MAP} = \arg\max \{ P(\theta \mid x) \}$$

$$= \arg\max \{ P(x \mid \theta) \} = \arg\max \{ \theta^2 - \theta^3 \}$$ How? Take derivative set = 0

if principle of indifference

$$= \frac{d}{d\theta} [\theta^2 - \theta^3] \overset{set}{=} 0$$

$$2\theta - 3\theta^2 = 0$$

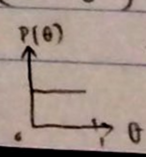
$$2 - 3\theta = 0 \Rightarrow \hat{\theta}_{MAP} = .67 = \hat{\theta}_{MLE}$$

frequentists can't do this

$$P(\theta \in [0.6, 0.7] \mid x) = \int_{0.6}^{0.7} P(\theta \mid x)\, d\theta = \int_{0.6}^{0.7} 12(\theta^2 - \theta^3) = 12 \left[\frac{\theta^3}{3} - \frac{\theta^4}{4} \right]_{0.6}^{0.7}$$

$$= .1765$$

$$P(\theta \mid x) = \frac{P(x \mid \theta) P(\theta)}{P(x)} = \frac{\theta^2 - \theta^3}{\underbrace{\int_0^1 P(x \mid \theta) P(\theta)\, d\theta}_{\overset{\shortparallel}{1}}} = \frac{\theta^2 - \theta^3}{\underbrace{\int_0^1 (\theta^2 - \theta^3)\, d\theta}_{\frac{\theta^3}{3} - \frac{\theta^4}{4} \Big]_0^1 = \frac{1}{12}}} = 12(\theta^2 - \theta^3)$$



$\langle 0, 1, 1 \rangle \Rightarrow$