Lec 4 2/17/17 Math 341

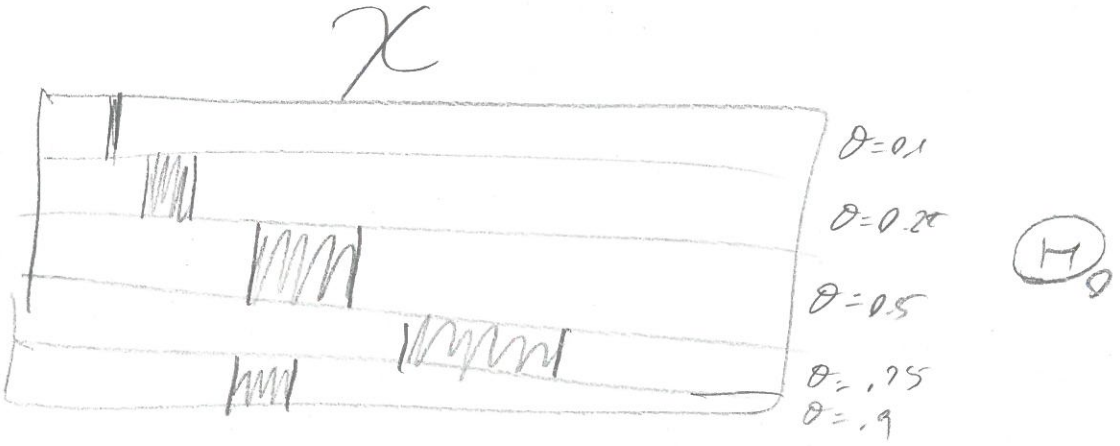
F = Bernoulli

$X = (0, 1, 1)$

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

$\theta \sim U(\Theta_0)$ i.e. discrete uniform on the elements of Θ_0 i.e. $P(\theta) = 0.2$ $\forall \theta \in \Omega$

We want to find $P(\theta|x)$. Before we do that... draw picture



$P(x|\theta)$ represents the prop. of area in any stripe.

$P(x|\theta=0.1) = 0.099$

$P(x|\theta=0.25) = 0.047$

$P(x|\theta=0.5) = 0.125$

$P(x|\theta=.75) = 0.141$

$P(x|\theta=0.9) = 0.061$

$P(\theta|x)$ is each of the slices div by the total area of all slices

What is best model already? Biggest slice of the slices $\theta = .75$

Already you see its $P(\theta|x)$ is going to be the largest

Bigger slice of the shows is a form of pt. estimation (best guess of Θ)!

$$\hat{\theta}_{MAP} = \underset{\theta \in H_0}{\arg\max} \{ P(\theta|x) \} = \underset{\theta \in H_0}{\arg\max} \left\{ \frac{P(x|\theta) P(\theta)}{P(x)} \right\}$$

↑ Max. a posteriori Bayesian estimate (AKA "posterior mode")

$$= \underset{\theta \in H_0}{\arg\max} \{ P(x|\theta) P(\theta) \}$$

since $P(x)$ is a normalizing constant $\neq f(\theta)$

$$= \underset{\theta \in H_0}{\arg\max} \{ P(x|\theta) \}$$

since $1(\theta)$ is the same for all H_0 $\neq f(\theta)$

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

$$P(\theta|x) = P(x|\theta) \cdot P(\theta) \cdot \frac{1}{P(x)}$$

- $P(x|\theta)$: slide based on prior belief (height)
- $P(\theta)$: slide based on prior belief (height)
- $\frac{1}{P(x)}$: normalize so all $P(\theta|x)$'s add up to 1 (relative to other strips)

Under principle of indifference ...

$$P(\theta|x) = \frac{P(x|\theta) P(\theta)}{\sum_{\theta \in H_0} P(x|\theta) P(\theta)} = \frac{P(x|\theta)}{P(x|\theta_1) + \dots P(x|\theta_k)}$$

See this here... $P(\theta = .75 | x = (0,1))$

$$= \frac{.191}{.009 + .042 + .125 + .141 + .061} = \frac{.141}{.383}$$

$\approx$ 37%

$\hat{\theta}_{MAP} = \hat{\theta}_{MLE}$ but $0.75 \neq 0.66$ Why? Our choice of prior did not cover all of the parameter space! $H_0 \neq H = (0,1)$ (for T = bernoulli)

Main skeptics of Bayesian stats: Prior could be wrong! reason to be

Let's look at down one set of the $\{.25, .75\}$ (indifference) $X_1 = 0$

What we know how ... prior and X_1

let $P(\theta | X_1)$ be our new prior.

$\Rightarrow$ No longer indifferent!

$$P(\theta = .25 | X_1 = 0) = \frac{P(X_1 = 0 | \theta = .25)}{P(X_1 = 0 | \theta = .25) + P(X_1 = 0 | \theta = .75)} = \frac{.25}{.75 + .25} = .75$$

$\Rightarrow P(\theta = .75 | X_1 = 0) = .25 = 1 - P(\theta = .25 | X_1$

$X_2 = 1$

$$P(\theta = 0.25 | X_2 = 1) = \frac{P(X_2 = 1 | \theta = 0.25)\, P(\theta = 0.25 | X_1)}{P(X_2 = 1 | \theta = 0.25)\, P(\theta = 0.25 | X_1) + P(X_2 = 1 | \theta = 0.75)\, P(\theta = 0.75 | X_1)} = \frac{0.25 \cdot 0.75}{0.25 \cdot 0.75 + 0.75 \cdot 0.25} = 0.5$$

were back to square 1. For this prior, no information learned. Make sense?

Now we know prior, X_1, X_2. Use this as prior

let $P(\theta | X_2, X_1)$ be our prior where $X_3 = 1$.

$$P(\theta = 0.25 | X_3 = 1) = \frac{P(X_3 = 1 | \theta = 0.25)\, P(\theta = 0.25 | X_2, X_1)}{P(X_3 = 1 | \theta = 0.75)\, P(\theta = 0.25 | X_2, X_1) + P(X_3 = 1 | \theta = 0.75)\, P(\theta = 0.75 | X_2, X_1)} = \frac{0.25 \cdot 0.5}{0.25 \cdot 0.5 + 0.75 \cdot 0.5} = 0.25$$

Same as $P(\theta = 0.25 | X = \langle 0, 1, 1 \rangle)$ from previously.

Is this true in general?

$$P(\theta | X_1, \ldots, X_n) = \frac{P(X_1, \ldots, X_n | \theta)\, P(\theta)}{P(X_1, \ldots, X_n)}$$

Why? iid

$$= \frac{P(X_n | \theta) \cdot \ldots \cdot P(X_2 | \theta)\, P(X_1 | \theta)\, P(\theta)}{P(X_n, \ldots, X_2 | X_1)\, P(X_1)} = P(\theta | X_1)$$

$$= \frac{P(X_n | \theta) \cdot \ldots \cdot P(X_3 | \theta)\, P(X_1, X_2 | \theta)\, P(\theta)}{P(X_n, \ldots, X_3 | X_1, X_2)\, P(X_1, X_2)} = P(\theta | X_1, X_2)$$

etc...

New question: we haven't seen X_4 yet. What is it doing?

Of course $P(X_4|\theta) = \theta^{X_4}(1-\theta)^{1-X_4}$ but... you don't know θ.)

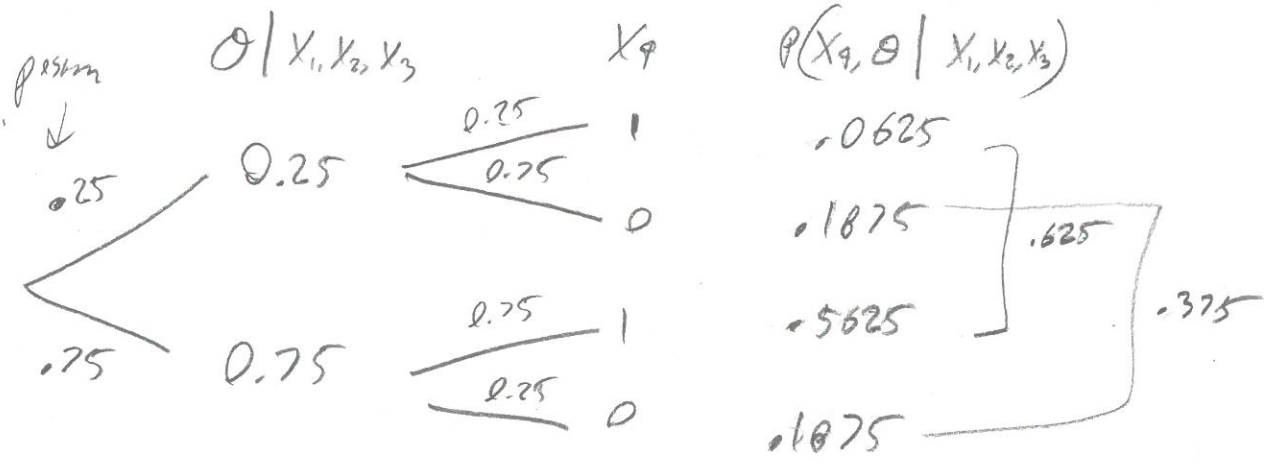
Recently... what did you do?

$$P(X_4|\theta) \approx P(X_4|\theta = \hat{\theta}_{MLE} = 0.66) = \text{Bern}(0.66)$$

What's the problem? Uncertainty in $\hat{\theta}_{MLE}$...

Bayesian Solution:

Seek: $P(X_4|X_1, X_2, X_3)$



$$\Rightarrow P(X_4|X_1 X_2 X_3) = \text{Bern}(.625)$$

this incorporates all uncertainty of θ using the prior & data.

What did we do?

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

Same as

$$P(Y) = \sum_{X \in S_{p(X)}} P(Y, X)$$ here is the world of X_1, X_2, X_3, θ

Bayes Rule

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

why?

$$= \sum_{\theta \in \{H\}_\theta} P(X_4 \mid \theta)\, P(\theta \mid X_1, X_2, X_3) = \sum_{\theta \in \{H\}_\theta} P(X_4 \mid \theta)\, \frac{P(X_1, X_2, X_3 \mid \theta)\, P(\theta)}{P(X_1, X_2, X_3)}$$

procedure: draw a θ from prior, Bernie draw of X_4 under θ, repeat for all θ's

Another take on the procedure: draw θ from prior. Examine how likely the data is under θ relative to all possible θ's, $P(X_1, X_2, X_3) = \sum_{\theta \in \{H\}} P(X_1, X_2, X_3 \mid \theta)\, P(\theta)$

then use this θ to see what draw of X_4 is

"Posterior Predictive Distribution"

$$P(X_4 \mid \theta) \overset{?}{=} P(X_4 \mid \theta, X_1, X_2, X_3)$$

$$= \frac{P(X_4, X_1, X_2, X_3, \theta)}{P(\theta, X_1, X_2, X_3)}$$

$$= \frac{P(X_1, X_2, X_3, X_4 \mid \theta)\, P(\theta)}{P(X_1, X_2, X_3 \mid \theta)\, P(\theta)} \quad \text{ici}$$

$$= \frac{P(X_1 \mid \theta)\, P(X_2 \mid \theta)\, P(X_3 \mid \theta)\, P(X_4 \mid \theta)}{P(X_1 \mid \theta)\, P(X_2 \mid \theta)\, P(X_3 \mid \theta)}$$

Intuition: Once θ is known ... data is useless when considering distr. of a future obs.

Generally ...

$$P(X^* \mid X_1, \ldots, X_n) = \sum_{\theta \in \{H\}_0} P(X^* \mid \theta)\, P(\theta \mid X_1, \ldots, X_n) \quad \text{(discrete)}$$

$$= \int_{\{H\}_0} P(X^* \mid \theta)\, P(\theta \mid X_1, \ldots, X_n)\, d\theta \quad \text{(continuous)}$$

$$\neq P(X^* \mid \hat{\theta}) \leftarrow \text{Bad idea since one pt. estimate cannot express the same distr!}$$

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

$0.75 \neq 0.667$ why? $H_0 \neq H$

This is likely a bad idea... you shouldn't put 0 prob on elements in the parameter space without good reason!

So what prior can we use?

$\theta \sim U(0,1)$ this has $\bullet$ $Supp(\theta) = [0,1] = H$

(wide range!)

- principle of indifference... no θ's are given special priority..

$$P(\theta) = \begin{cases} 1 & \text{if } \theta \in H \\ 0 & \text{o/e} \end{cases}$$

density function of std. uniform.

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

$$P(\theta \mid X) = \frac{P(X_1, X_2, X_3 \mid \theta)\, P(\theta)}{P(X_1, X_2, X_3)} \propto P(X_1, X_2, X_3 \mid \theta) = \theta^2(1-\theta)$$

$$\hat{\theta}_{MAP} = \arg\max \{\theta^2(1-\theta)\} \quad \frac{d}{d\theta}\left[\theta^2(1-\theta)\right] = \frac{d}{d\theta}\left(\theta^2 - \theta^3\right) = 2\theta - 3\theta^2$$

$$\Rightarrow 0 = 2 - 3\theta \Rightarrow \hat{\theta}_{MAP} = \frac{2}{3} \checkmark$$

"the less model"

New question... what if I'm interested in $P(\theta \in [0.6, 0.7] \mid X)$

i.e. what we conditio got before!

We need to solve for $P(\theta|x)$ explicitly ... not just find its mode ...

$$P(\theta|x=(0,1,1)) = \frac{P(x_1,x_2,x_3|\theta)\,P(\theta)}{P(x_1,x_2,x_3)} = \frac{\theta^2(1-\theta)}{\int P(x_1,x_2,x_3|\theta)\,P(\theta)\,d\theta} = 12\theta^2(1-\theta)$$

$P(x_1,x_2,x_3|\theta)$ under H_0

$\sum P(x_1,x_2,x_3|\theta)$ under every θ!

$$\int_0^1 \theta^2(1-\theta)\,(1)\,d\theta = \int_0^1 (\theta^2-\theta^3)\,d\theta = \frac{\theta^3}{3} - \frac{\theta^4}{4}\Big]_0^1 = \frac{1}{3}-\frac{1}{4} = \frac{1}{12}$$

$P(\theta)$

$$P(\theta \in [0.6, 0.7]\,|\,x) = \int_{0.6}^{0.7} P(\theta|x)\,d\theta = \int_{0.6}^{0.7} 12\,\theta^2(1-\theta)\,d\theta = 12\left[\frac{\theta^3}{3} - \frac{\theta^4}{4}\right]_{0.6}^{0.7} =$$

$$= 12\,(0.0543 - 0.0396) = \boxed{0.1765}$$

this is the probability that "θ" is between 0.6 and 0.7 assuming the prior.

Let's solve for general data $x_1, \dots, x_n$ and same prior.

$$P(\theta|x) = \frac{P(x|\theta)\,P(\theta)}{P(x)} = \frac{\prod_{i=1}^n P(x_i|\theta)}{\int P(x|\theta)\,P(\theta)\,d\theta} = \frac{\prod_{i=1}^n P(x_i|\theta)}{\int_0^1 \prod_{i=1}^n P(x_i|\theta)\,d\theta}$$

H_0

$$\prod_{i=1}^n \theta^{x_i}(1-\theta)^{1-x_i} = \theta^{\sum x_i}(1-\theta)^{n-\sum x_i} \Rightarrow \text{posterior only is dependent on } \sum x_i \Rightarrow \text{let } x := \sum x_i \text{ from now on...}$$

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

Form of $B(x,y) := \int_0^1 t^{x-1}(1-t)^{y-1}\,dt$

"capital β"