Lecture 4 - 2/10/21

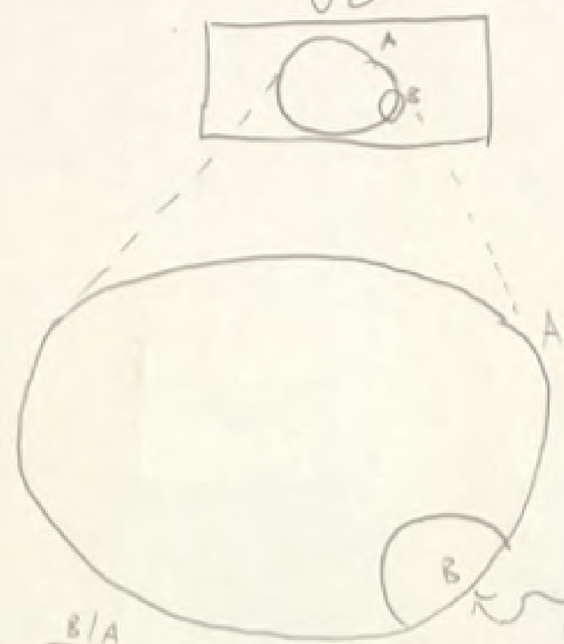
Let A : smoking and B : lung cancer

$P(A) = 0.200$

$P(B) = 0.060$

$P(AB) = 0.036$

$\Omega \leftarrow$ universe



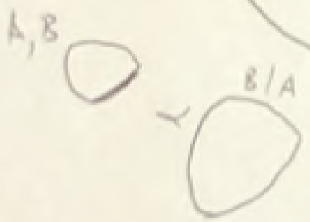
$A \subset \Omega$

P (lung cancer | Smoking)

$P(B|A)$

$\uparrow$

given, conditional universe
(B, relative to A)



"proportional"

$P(B|A) \propto P(AB)$

c = scale factor

$$P(B|A) = c\,P(AB) = \frac{P(\Omega)}{P(A)} P(AB) = \frac{P(AB)}{P(A)}$$

Definition of conditional probability

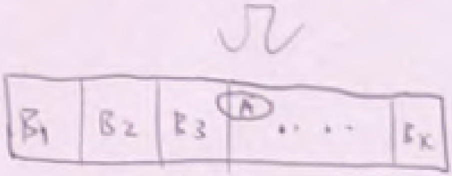
$$P(A|B) = \frac{P(AB)}{P(B)} \Rightarrow P(AB) = P(A|B)\,P(B) \Rightarrow P(B|A) = \frac{P(A|B)\,P(B)}{P(A)}$$

Bayes rule

$$P(A) = P(AB) + P(AB^c) = P(A|B)\,P(B) + P(A|B^c)\,P(B^c)$$

If $B_1, B_2, \ldots, B_K$ are mutually exclusive and collectively exhaustive

$\forall i \neq j \quad B_i \cap B_j = \phi$ $\qquad\qquad \Omega = \bigcup_{i=1}^{K} B_i$

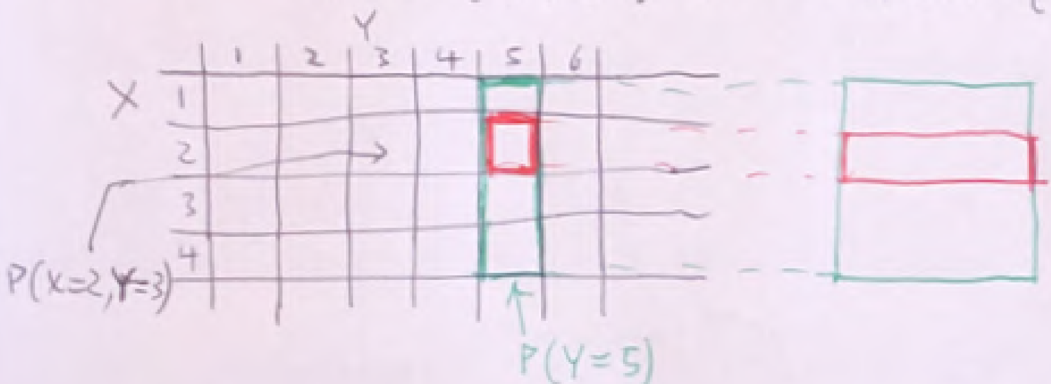


$$P(A) = \underbrace{\sum_{k=1}^{K} P(A, B_K)}_{} = \sum_{k=1}^{K} P(A \mid B_K) P(B_K)$$

"margining out the B_K's or integrating out the B_K's"

$$P(B_i \mid A) = \frac{P(A, B_i)}{P(A)} = \frac{P(A \mid B_i) P(B_i)}{\sum_{k=1}^{K} P(A \mid B_K) P(B_K)}$$ Bayes Theorem

Bayes Rule and Bayes Thm for r.v.'s. Imagine two r.v.'s X, Y and the $Supp[X] = \{1,2,3,4\}$ and $Supp[Y] = \{1,2,3,4,5,6\}$



$$P(Y=5) = P(Y=5, X=1) + P(Y=5, X=2) + P(Y=5, X=3) + P(Y=5, X=4)$$
$$= \sum_{x \in Supp[X]} P(Y=5, X=x)$$

$$P(X=2 \mid Y=5) = \frac{P(X=2, Y=5)}{P(Y=5)}$$

$$P(y) := P(Y=y) = \sum_{x \in Supp[X]} P(Y=y, X=x) = \sum_{x} p(y,x)$$

$$P(X=x \mid Y=y) = P(X=x, Y=y) = \frac{P(x,y)}{P(x)}$$ $\leftarrow$ JMF

Conditional PMF

$$f(y) = \int_{Supp[X]} f(x,y)\, dx$$

$$f_{X|Y}(x,y) = \frac{f(x,y)}{f(y)}$$ $\leftarrow$ JDF

Back to the story... can we use Bayes Rule to tell us anything about inference for parameter θ given data x ($x = \langle x_1, \dots, x_n \rangle$).

Consider:
$$P(\theta|X) = \frac{\overbrace{P(X|\theta)}^{P(x_1,\dots,x_n;\theta) = \mathcal{L}(\theta;x_1,\dots,x_n) \to \text{JMF, but called "likelihood"}} P(\theta)}{P(x)}$$

What is wrong with this equation? Previously, θ, the unknown parameter was assumed to be a fixed real value. Thus, $\theta \sim \text{Deg}(\theta)$. Then, this equation is trivial. If you plug in the actual value of $\theta = \theta_0$ on the right hand side then you get:

$$P(\theta = \theta_0 | x) = \frac{P(x|\theta)(1)}{\sum_{\theta \in \Theta} P(x|\theta)P(\theta)} = \frac{P(x|\theta_0)}{P(x|\theta_0)} = 1$$

$$P(\theta \neq \theta_0 | x) = \frac{P(x|\theta_0)(0)}{\sum P(x|\theta)P(\theta)} = \frac{0}{P(x|\theta_0)} = 0$$

This was a mean exam problem but not super interesting since you don't know θ_0 and even if you did, this doesn't help with the three goals of inference.

The big leap: let θ be a r.v.! Then $P(\theta)$ has a distribution (either discrete or continuous). But it's a constant! This is the big philosophical problem in Bayesian Statistics / Bayesian Inference. Some authors say it's still a constant but $P(\theta)$ represents uncertainty in its value. Purists say that's nonsense.

$$\underbrace{P(\theta|X)}_{\text{posterior}} = \frac{\overbrace{P(X|\theta)}^{\text{likelihood}}\ \overbrace{P(\theta)}^{\text{prior}}}{\underbrace{P(x)}_{\leftarrow \text{ prior predictive distribution}}}$$

$$P(X) \begin{cases} = \sum_{\theta \in \Theta} P(x|\theta)P(\theta) & \text{if } \theta \text{ discrete} \\ = \int_{\Theta} P(x|\theta)\,p(\theta)\,d\theta & \text{if } \theta \text{ contin.} \end{cases}$$

<u>prior</u>: thoughts summed up in a distribution over Θ the parameter space **before** seeing any data. There is no x within it.

Frequentists say this is "subjective" and not real!

<u>posterior</u>: thoughts summed up in a distribution over Θ the parameter space **after** seeing the data x which is why it's conditional on x!

Notation for the rest of class: "p" now discrete PMF/conditional mass function **or** continuous PDF/conditional density function. I won't use "f" anymore.

$\mathcal{F}$ = iid Bernoulli, $x = \langle 0,1,1 \rangle$ $\quad P(x|\theta) = \theta^2(1-\theta)$

Let $\Theta_0 = \{0.5, 0.75\} \neq (0,1)$

$$P(\theta = 0.75|x) \overset{?}{>} P(\theta = 0.5|x)$$

$$P(\theta = 0.75|x) \overset{\text{Bayes Thm}}{=} \frac{P(x|\theta = 0.75)\,P(\theta = 0.75)}{P(x|\theta = 0.5)\,P(\theta = 0.5) + P(x|\theta = 0.75)\,P(\theta = 0.75)}$$

$P(x|\theta = 0.75) = (0.75)^2(.25) = 0.141$, $P(x|\theta = 0.5) = (0.5)^3 = 0.125$

We need $P(\theta = 0.75)$ and $P(\theta = 0.5)$ to complete the calculations. That's the prior, $P(\theta)$. It's subjective. What do you think it should be? Amir says $P(\theta = 0.75) = 0.2$ and $P(\theta = 0.5) = 0.8$ because he feels that way.

An automatic rule is called the "principle of indifference" (Laplace's idea so it's sometimes called the "Laplace prior"). This principle says that all values of θ in the parameter space are equally likely. In our case,

$$P(\theta) = \begin{cases} 0.5 & \text{if } \theta = 0.7 \\ 0.5 & \text{if } \theta = 0.5 \end{cases} \quad \text{In general, } P(\theta) = \frac{1}{|\Theta|},$$

this formula only works for finite parameter spaces.

$$P(\theta = 0.75|x) = \frac{(0.141)(0.5)}{(0.125)(.5) + (0.141)(.5)} \approx 0.53^*$$

$\theta = 0.75$ is your point estimate.

$$P(\theta = 0.5|x) = \frac{(0.125)(.5)}{(0.125)(.5) + (.141)(.5)} = 0.47$$

$$P(\theta = 0.75) = 0.5 \overset{x}{\Longrightarrow} P(\theta = 0.75|x) = 0.53$$

This is called Bayesian Conditionalism.