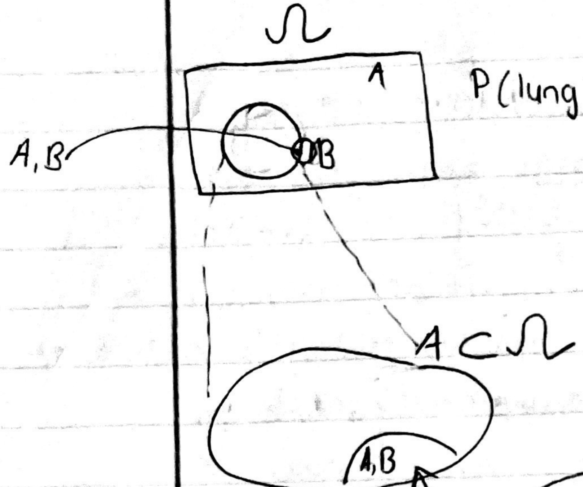
Feb 10
Lec 4.

Let A: smoking and B: lung cancer

$P(A) = 0.200$
$P(B) = 0.060$
$P(AB) = 0.036$

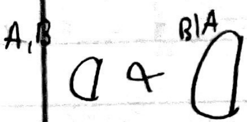


P(lung cancer | smoking)

given, conditional universe

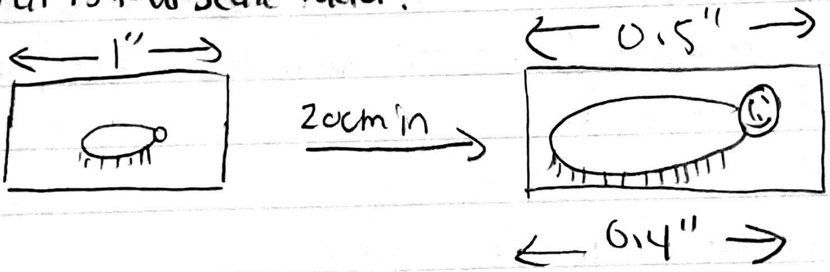
$P(B|A)$

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



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

what is that scale factor?



zoom function $= \frac{1}{0.5} = 2$

c is the zoom factor

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

Probability of conditional probability

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

Bayes Rule

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

we use conditional probability

If B_1, B_2 . . . B_k are mutually exclusive and collectively exhaustive.

$\forall i \neq j \quad B_i \cap B_j = \emptyset$

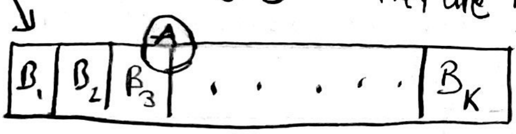
if we get back the universe we get

$$\Omega = \bigcup_{i=1}^{k} B_i$$

If you land in one of these squares you can't land into others

Ω

they are mutally exclusive



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

"marging out the K_K's or integrating out the B_K's just to get A back

Bayes Theorem

$$P(B_i|A) = \frac{P(A, B_i)}{P(A)} = \frac{P(A|B_i)P(B_i)}{\sum_{K=1}^{K} P(A|B_K)\,P(B_K)}$$

Bayes Rule and Bayes Theroem for Random Variable.

Imagine two random variable X, Y and the $supp[X] = \{1,2,3,4\}$ and $Supp[Y] = \{1,2,3,4,5,6\}$

$P(X=2, Y=3)$

X \ Y	1	2	3	4	5	6
1						
2						
3						
4					$P(Y=5)$	

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

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

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

Joint Mass Function

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

conditional PMF

Continous

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

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

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

$P(x_1, \ldots x_n; \theta) = \mathcal{L}(\theta; x_1, \ldots x_n)$ — JMF but called likelihood

Consider: $$P(\theta|X) = \frac{P(X|\theta)\,P(\theta)}{P(X)}$$

what is wrong w/ this equation?

Previously, θ, the parameter was assumed to be a fixed real value. Thus, $\theta \sim Deg(\theta)$. Then, this equation is trivial. If you plugin the actual value of θ on the r.h.s. then you get:

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

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

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 w/ the three goals of inference

What can I change, how can I make sense, here is the big leap.

* Let θ be a random variable! 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 $P(\theta)$ represent uncertainty in it's value. Purists say that's nonsense.

$$\underbrace{P(\theta|X)}_{\text{Posterior}} = \frac{\overbrace{P(X|\theta)}^{\text{likelihood}}\ \overbrace{P(\theta)}^{\text{prior}}}{P(X)}$$

$P(X) \leftarrow$ prior predictor distribution

Why is it called prior? prior is thoughts summed up in a distribution over θ the parameter space **prior** to seeing any data. There is no X within in. Frequentist say this is "subjective and not real

posterior: 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 denote discrete PMF / conditional mass function **or** continous PDF / conditional density Function. I won't use "F" anymore.

$\mathcal{F}$ = iid Bernoulli, $X < 0,1,1 >$ $P(X|\theta) = \theta^2(1-\theta)$

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

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

Bayes thm.

$$P(\theta = 0.75|X) \overset{\vee}{=} \frac{P(X|\theta = 0.75)P(\theta = 0.75)}{P(X|\theta = 0.5)P(\theta = 0.5) + P(X|\theta = 0.75)P(\theta = 0.75)}$$

$P(X|\theta = 0.75) = 0.75^2 \cdot 0.25 = .141$, $P(X|\theta = 0.5) = .5^3 = .125$

We need $P(\theta = 0.75)$ and $P(\theta = 0.5)$ to complete the calculation. that's the prior $P(\Theta)$. It's subjective. What do you think it should be?

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

$$P(\theta) = \begin{cases} 0.5 & \text{if } \theta = 0.75 \\ 0.5 & \text{if } \theta\ 0.5 \end{cases}$$

In general $P(\theta) = \dfrac{1}{|\Theta|}$ — this formula only works for finite parameter space

use $\dfrac{P(X|\theta)\,P(\theta)}{P(X)}$ formula

$$P(\theta = 0.75 \mid X) = \frac{.141 \cdot (.5)}{.125 \cdot (.5) + .141 \cdot (.5)} = 0.53$$

$\theta = 0.75$ is your point estimate

$$P(\theta = 0.5 \mid X) = \frac{.125 \cdot (.5)}{.125 \cdot (.5) + .141 \cdot (.5)} = .47$$

$$P(\theta = 0.75) = 0.5 \implies P(\theta = 0.75 \mid X) = .53$$

This is called Bayesian Conditionalism

Scanned with CamScanner