Lecture 4

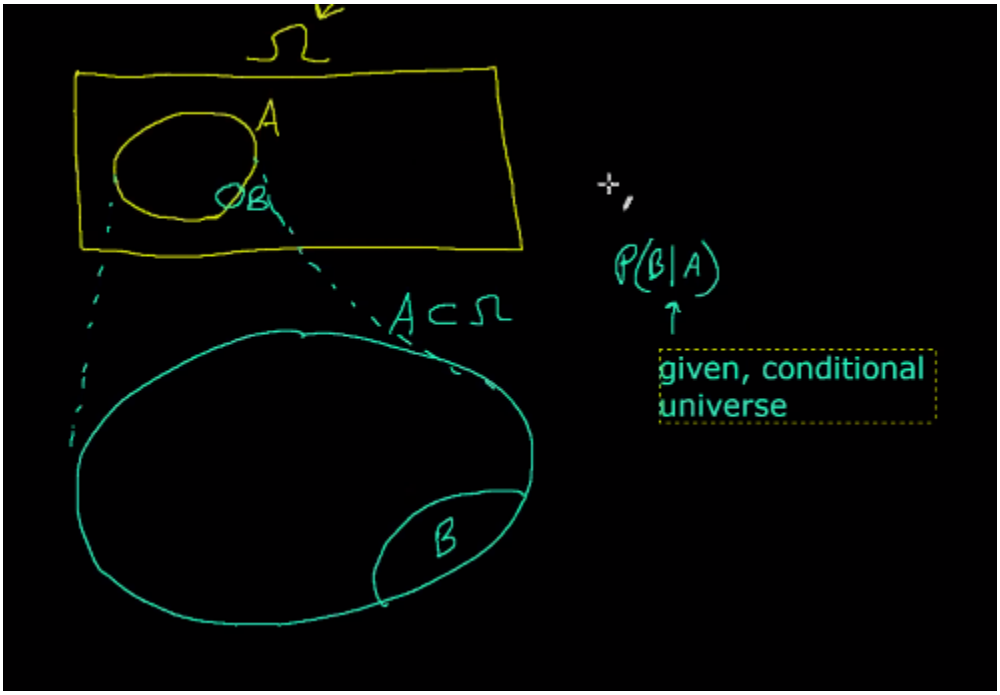
Bayes From the Ground Up

Let the event A be smoking, and let B be the event of lung cancer.

$$P(A) = 0.200$$

$$P(B) = 0.060$$

$$P(AB) = P(A \cap B) = 0.36$$



Note the diagram, Ω is how we denote the universe. The event A $\subset\Omega$, and the event $B \subset A$. $P(B|A)$ is how we denote a condition, read, the probability of B given A.

$$P(B|A) \propto P(AB) \rightarrow P(B|A) = cP(AB) = \frac{P(\Omega)}{P(A)}P(AB)$$

Because $P(\Omega) = 1$, we get,

$$\frac{P(\Omega)}{P(A)}P(AB) = \frac{P(AB)}{P(A)}$$

Similarly,

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

This leads us to Bayes' Rule,

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

Using the definition of conditional probability we get,

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

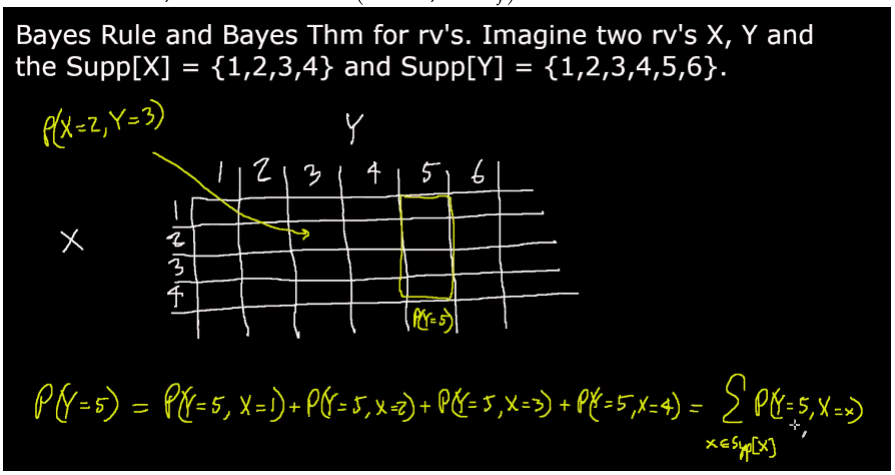
If the events$B_1, B_2, ..., B_k$ are mutually exclusive and collectively exhaustive, where mutual exclusivity means that the intersection of these events is the empty set, and where collectively exhaustive means the union of the sets gives you the universe, as in there must me an event that occurs, we get this rule,

$$P(A) = \Sigma_{k=1}^{K} P(A, B_i) = \Sigma_{k=1}^{K} P(A|B_k)P(B_k)$$

In the equation above, we "margin out" the B_ks or "integrate out" the B_ks. So lets go back to Bayes' rule, substituting in what we have found so far, we have Bayes' theorem,

$$P(B_i|A) = \frac{P(A, B_i)}{P(A)} = \frac{P(A|B_i)P(B_i)}{\Sigma_{k=1}^{K} P(A|B_k)P(B_k)}$$

So far we have worked with the probabilities of events. Now we will deal with the probability of events of discrete random variables, or realizations. So imagine two random variables, X, Y, and $Supp[X] = \{1,2,3,4\}$ and $Supp[Y] = \{1,2,3,4,5,6\}$. Now imagine a grid, with X on the diagonal axis, and Y on the horizontal axis. At the intersections, we will have $P(X = i, Y = j)$.



Now we will look at events of continuous random variables and their realizations.

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

$$P(y) = P(Y = y) = \Sigma_{x \in Supp[X]} P(Y = y, X = x) = \Sigma_x P(y, x)$$

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

Conditional PMF = $P(X = x|Y = y)$, JMF = $\frac{P(X=x,Y=y)}{P(x)}$.

Back to the story, can we use Bayes' rule to tell us anything about inference for parameter θ given data x, $x =< x_1, x_2, ..., x_n >$. The JMF is also defined below, $P(x|\theta)$, which is also called the likelihood.

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

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

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

$$P(\theta \neq \theta|x) = \frac{P(x|\theta)0}{\Sigma_{\theta \in \Theta} P(x|\theta)P(\theta)} = \frac{0}{P(x|\theta)} = 0$$

This was a mean exam problem but its not super interesting since you do not know θ and even if you did, this doesn't help with the three goals of inference. So where is the big leap? How do we make this make sense?

Big Leap:
Let θ be a random variable. The $P(\theta)$ has a distribution (either discrete or continuous). Why is this a big leap? Well, θ is a constant. This is the big philosophical problem with Bayesian Statistics/Bayesian Inference. θ is a constant but we are assigning it to be a random variable. Some authors say its still a constant, but $P(\theta)$ represents uncertainty in its value. Frequentists say that's nonsense, subjective, and not real. Now lets do an anatomy of the Bayesian Rule.

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

Posterior: $P(\theta|x)$
Likelihood: $P(x|\theta)$

Prior: $P(\theta)$
Prior predictive distribution: $P(x)$

The prior represents thoughts summed up in a distribution over Θ the parameter space *prior* to seeing any data. There is no x within. The posterior represents thoughts summed up in a distribution over Θ, the parameter space *after* seeing the data, x, which is why its conditional on x.

Notation for the rest of class:
p now denotes discrete PMF/ conditional madd dunction *or* continuous PDF/ conditional density function. We will not be using f anymore.

If θ is discrete:

$$P(x) = \Sigma_{\theta \in \Theta} P(x|\theta)P(\theta)$$

If θ is continuous:

$$P(x) = \int_{\Theta} P(x|\theta)P(\theta)$$

Lets say the event is the toss of a coin, and we wish to see if the coin is weighted. Where, $F = iidBernoulli$, x = <0, 1, 1>, where $P(x|\theta) = \theta^2(1-\theta)$. Let $\Theta_0 = \{0.5, 0.75\} \neq (0, 1)$. Now can we say the following,

$$P(\theta = 0.75|x) >? P(\theta = 0.50|x)$$

Well, lets calculate both sides with Bayes' Theorem.

$$P(\theta = 0.75|x) = \frac{P(x|\theta = 0.75)P(\theta = 0.75)}{P(x|\theta = 0.50)P(\theta = 0.50) + P(x|\theta = 0.75)P(\theta = 0.75)}$$

We already have $P(x|\theta)$, so we'll plug that in.

$$P(x|\theta = 0.75) = (0.75)^2(0.25) = 0.141, P(x|\theta = 0.5) = (0.5)^3 = 0.125$$

Now we only need $P(\theta = 0.75)$ and $P(\theta = 0.50)$ to complete the calculation. Thats the prior, $P(\theta)$. It's subjective. What do we thing it should be? Well any two numbers that equal 100% works. Say, $P(\theta = 0.75) = 0.2$ and $P(\theta = 0.50) = 0.8$ because I say it should be so. An automatic rule is called the *"principle of indifference"*, which is sometimes called the "Laplace prior". This principle says that ll values of $\theta \in \Theta$ are equally likely. In our case,

$$P(\theta) = \{^{0.5 \; if \; \theta = 0.75}_{0.5 \; if \; \theta = 0.50}$$

In general, $P(\theta) = \frac{1}{|\Theta|}$. This is also called the principle of equal priors. We'll use this philosophy to answer the problem.

$$P(\theta = 0.75|x) - \frac{0.141 \times 0.5}{0.125 \times 0.5 + 0.141 \times 0.5} = 0.53$$

$$P(\theta = 0.75|x) - \frac{0.125 \times 0.5}{0.125 \times 0.5 + 0.141 \times 0.5} = 0.47$$

This makes sense only if you consider that you saw more heads than tails. You can see that the principle of indifference impacts the result of the data. This process is called Bayesian Conditionalism.