STA 365: Applied Bayesian Statistics

Boris Babic
Assistant Professor, University of Toronto

Week 1B: Bayes' Theorem

Bayes' Theorem

$$\Pr(E|I) = \frac{\Pr(I|E)\Pr(E)}{\Pr(I)}$$

- Simple example: Pr of die landing on 4, conditional on it landing on an even number.
- $\Pr(4|even) = \frac{\Pr(even|4)\Pr(4)}{\Pr(even)}$
- $\Pr(4|even) = \frac{1 \times 1/6}{1/2} = 1/6 \times 1/2 = 1/3$

Total probability

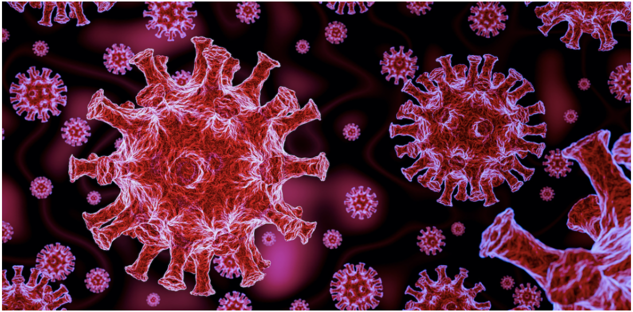
Law of total probability:

$$\mathrm{P}(E) = \sum_{j=1}^{n} \mathrm{P}(E|I_j)\mathrm{P}(I_j)$$

Medical Screening

Bayes' Theorem

The Bayesian Approach



- Suppose that the base rate (prevalence) of COVID-19 is 1% (ha!).
- Suppose PCR tests have 99% sensitivity. That is, $p(+|D) = 0.99$
- Suppose they also have specificity of 95%. That is, $p(-|\overline{D}) = 0.95$
- Find $p(D|+)$.

Medical Screening

- This beecomes:

$$p(D|+) = \frac{.99 \times .01}{.99 \times .01 + .05 \times .99} = 0.167$$

- Notice the difference between $p(D|+)$ (0.167) and $p(+|D)$ (0.99)!
- High specificity and sensitivity can still lead to extremely low posterior probability
- A lesson to remember for AI and machine learning!

Medical screening

$P(D) = 0.01$

$P(\bar{D}) = 0.99$

sensitivity: $P(+|D) = 0.99$

specificity: $P(-|\bar{D}) = 0.95$

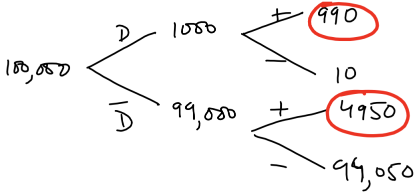
$\underbrace{P(-|D) = 0.01}_{\text{False -ve}}$

$\underbrace{P(+|\bar{D}) = 0.05}_{\text{False +ve}}$

$$P(D|+) = \frac{P(D)\,P(+|D)}{P(+|D)P(D) + P(+|\bar{D})P(\bar{D})}$$

$= 0.167$

Medical Screening



$$P(D|+) = \frac{990}{990+4950} = 0.167$$

- Find $p(D|-)$

$$p(D|-) = \frac{p(D) \times p(-|D)}{p(D) \times p(-|D) + p(-|\overline{D} \times p(\overline{D}))}$$

- $p(D|-) = (0.01 \times 0.01)/(0.01 \times 0.01 + 0.95 \times 0.99) = 0.0001$

Marginal and Conditional Independence

- E_2 is marginally independent of E_1 if learning about E_1's occurrence doesn't affect the probability of E_2:
$$p(E_2|E_1) = p(E_2)$$
- In this case, it would be completely redundant to laern about E_1.
- Medical tests are not like this. Once you test positive a second time, it is still informative that you have already tested positive once.
- E_2 is conditionally independent of E_1 if learning about E_1's occurrence doesn't affect the probability of E_1 provided that one has learned about the occurrence of a third event, E_3:
$$p(E_2|E_1, D) = p(E_2|D)$$
- This is true of medical tests: they are indepnedent conditional on the true disease state.

Medical Screening

Conditional Independence : $P(+_2 | D, +_1) = P(+_2 | D)$

Total Redundancy

Marginal independence : $P(+_2 | +_1) = P(+_2)$

---

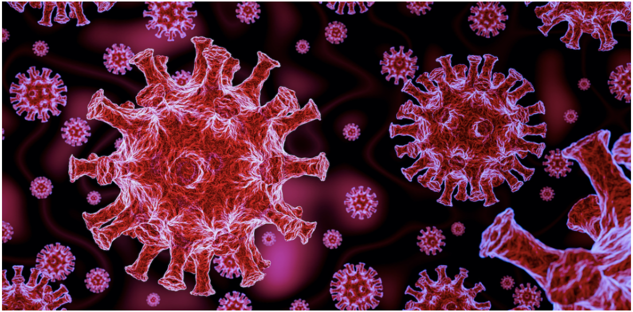
with conditional independence :

$$P(D) \xrightarrow{1^{st} \text{ test}} P(D | +_1)$$

$$P(D | +_1) \xrightarrow{2^{nd} \text{ test}} P(D | +_1, +_2)$$

$$P(D | +_1, +_2) = \frac{P(D) P(+_1, +_2 | D)}{P(+_1, +_2 | D) P(D) + P(+_1, +_2 | \bar{D}) P(\bar{D})}$$

Medical Screening



- Suppose again that the base rate (prevalence) of COVID-19 is 1%.
- As before PCR tests have 99% sensitivity. That is, $p(+|D) = 0.99$
- And they also have specificity of 95%. That is, $p(-|\overline{D}) = 0.95$
- YOu have tested positive once and found that $p(D|+)0.167$
- Find $p(D|+_1, +_2)$.

Medical Screening

$P(D|+_1,+_2) = ??$

$$P(D|+_1,+_2) = \frac{P(D|+_1)P(+_2|D,+_1)}{P(+_2|D,+_1)P(D|+_1) + P(+_2|\bar{D},+_1)P(\bar{D}|+_1)}$$

$$\left.\begin{aligned} P(+_2|D,+_1) &= P(+_2|D) \\ P(+_2|\bar{D},+_1) &= P(+_2|\bar{D}) \end{aligned}\right\} \rightsquigarrow \text{conditional independence !}$$

with conditional independence,

$$P(D|+_1,+_2) = \frac{0.167(0.99)}{0.167(0.99) + 0.833(0.05)} = 0.80$$

Generalizing Further

3 or more events

$E_1, E_2, \ldots, E_k$ form a partition on S

$$P(E_i \mid I) = \frac{P(E_i)P(I \mid E_i)}{\sum_{i=1}^{k} P(I \mid E_i)P(E_i)} \leftarrow P(I)$$

k prior prob'ys: $P(E_1), \ldots, P(E_k)$

k likelihoods: $P(I \mid E_1), \ldots, P(I \mid E_k)$

k posterior prob'ys: $P(E_1 \mid I), \ldots, P(E_k \mid I)$

If you test positive for a certain disease...

You may want to ask

- Do I have the disease or not?
- What is the chance that I have the disease?

Possible answers

- Frequentist: I do not know. You're asking the wrong question. Whether you have the disease or not is not a random variable. It is a fixed value. Therefore, the question does not make sense.
- Bayesian: The chance that you have the disease is ... % (How?)

Bayesian Statistics

Differences from frequentist statistics

- On the Bayesian approach, the parameter θ is considered as a random quantity.
- We describe our uncertainty about θ by a probability distribution, referred to as the prior distribution.
- A sample is taken from a population indexed by θ, and the prior is then updated, using Bayes' Rule, to get a posterior distribution for θ given the sample.
- Inferences are then made from the posterior distribution.

The Three Steps of the Bayesian Approach

- Set up a full probability model
  - A joint probability distribution for all observable and unobservable quantities in a problem.
  - The model should be consistent with knowledge about the underlying scientific problem and the data collection process
- Condition on the observed data
  - Calculate and interpret the appropriate posterior distribution
  - We are interested in the conditional probability distribution of the unobserved quantities of interest given the observed data
  - We are often interested in the marginal distribution of a subset of unobserved quantities
- Evaluate the model
  - Does the model fit the data?
  - Are substantive conclusions reasonable?
  - How sensitive are results to model assumptions?

Random Variables and the probability density function

- In Bayesian inference a random variable is defined as an unknown numerical quantity about which we make probability statements.
- Quantitative outcomes of experiments are random variables.
- And fixed but unknown parameters are also random variables.
- Recall that $F(y) = \Pr(Y \leq y)$ is called the cumulative distribution function..
- If F is continuous we say that Y is a continuous random variable.
- A theorem from mathematics says that for every continuous cdf F there exists a positive function $\pi(y)$ such that
$$F(a) = \int_{\infty}^{a} \pi(y)d(y)$$
- $\pi(y)$ is the probability density function of Y
- Its essential properties are that it is positive, and it integrates to 1.
- Due to the linearity of integration it is also additive.
- Integration for continuous distributions behaves similarly to summation for discrete distributions.
- In fact, integration can be thought of as a generalization of summation for situations in which the sample space is not countable.
- $p(y)$ is not the probability that $Y = y$.

Notation

- $X = x, Y = z, Z = z$ for observed data/ observable random variables.
- θ unobservable (vector) of quantities/population parameters. Eg: $\theta' = (\mu, \sigma^2)'$.
- $\widetilde{y}$ unknown, potentially observable data (e.g., the outcome for the next patient enrolled in a trial)
- $\pi(\cdot)$ is a probability density/mass function. Sometimes we use $\Pr(\cdot)$.
- $\pi(\cdot|\cdot)$ denotes a conditional density/mass function.
- $f(x|\theta)$ is the sampling distribution of X. Equivalently: $\pi(x|\theta)$
- $\theta \sim f(\varphi)$ means that the distribution of θ is given by $f(\varphi)$ (and hence φ is a (vector) of hyperparameter(s)).
- Parameter space: Θ, sample space: $\mathcal{Y}$.

Bayesian Framework

- Prior distribution for θ:
$$\theta \sim \pi(\theta)$$
- Sample distribution (or likelihood) of $\boldsymbol{X}$ given θ:
$$\boldsymbol{X}|\theta \sim f(\boldsymbol{x}|\theta) = \pi(\boldsymbol{x}|\theta)$$
- Joint distribution of $\boldsymbol{X}$ and θ (this is our full model):
$$f(x, \theta) = f(\boldsymbol{x}|\theta)\pi(\theta)$$
- Recall, chain rule of probability: $p(E_1 \cap E_2) = p(E_2|E_1)p(E_1)$
- Marginal distribution of $\boldsymbol{X}$:
$$m(\boldsymbol{x}) = \int_{\theta\in\Omega} f(\boldsymbol{x}, \theta)d\theta = \int_{\theta\in\Omega} f(\boldsymbol{x}|\theta)\pi(\theta)d\theta$$
- Posterior distribution of θ (conditional distribution of θ given $\boldsymbol{X}$):
$$\pi(\theta|\boldsymbol{x}) = \frac{f(\boldsymbol{x}, \theta)}{m(\boldsymbol{x})} = \frac{f(\boldsymbol{x}|\theta)\pi(\theta)}{m(\boldsymbol{x})} \propto f(\boldsymbol{x}|\theta)\pi(\theta) \qquad \text{(Bayes' Rule)}$$

Exchangeability

- At the beginning of class 1A, we assumed the data generating process is iid. This is a typcial assumption in classical inference.
- In Bayesian inference, we will assume something weaker, *exchangeability*.
- For example, a sequence of coin tosses is exchangeable if $\pi(x_1, \ldots, x_n) = \pi(x_{\sigma(1)}, \ldots, x_{\sigma(n)})$ for every permutation σ of the order.
- Position and order is irrelevant, for any length of the sequence.
- Exchangeability is an assumption about the underlying symmetry in the inference problem.
- For example: compare the probability of observing (H, H, T) with the probability of observing (H, T, H).
- Independent (Bernoulli) trials with x successes and $n - x$ failures are exchangeable: for any length, the probability is proportional to $\theta^x(1-\theta)^{n-x}$
- Can you think of an exchangeable sequence that is not iid?

DeFinetti's Theorem

- Let $Y_i \in \mathcal{Y}$ for all $i \in \{1, 2, ...\}$. Suppose that, for any n, our belief model for $Y_1, ..., Y_n$ is exchangeable:

$$\pi(y_1, ...y_n) = \pi(y_{p1}...y_{pn})$$

  for all permutations p of $\{1, ...n\}$.

- Then our model can be written as:

$$\pi(y_1, ...y_n) = \int \left[\prod_{i=1}^{n} \pi(y_i|\theta) \right] \pi(\theta) d\theta$$

- This θ, whose existence is guaranteed for exchangeable sequences, can be interpreted as the Bayesian prior. But notice that it is not imposed into the problem! Its existence follows from a symmetry assumption weaker than iid.

Sketch proof of exchangeability for the Bernoulli process

- Let $p_{k,n}$ denote $P(X, n-x)$ where $X_1, \ldots, X_m$ is an exchangeable sequence of Bernoulli random variables.
- Let $q_r = P(\sum_{i=1}^m X_i = r)$
- Then,
$$p_{k,n} = \sum_{r=0}^{m} \frac{(r)_k (m-r)_{n-k}}{(m)_n}$$
where $(x)_k = \prod_{j=0}^{k-1}(x-j)$
- From exchangeability, it follows that given r ones, the distribution of $X_1, \ldots, X_m$ is the same as that obtained by drawing from an urn containing r ones and $m-r$ zeros.
- Thus, the rth term of the series is
$$P\left[X_1 = 1, \ldots, X_r = 1, X_{r+1} = 0, \ldots, X_m = 0 | \sum_{j=1}^{m} X_j = r\right] \times P\left[\sum_{j=1}^{m} X_j = r\right]$$
- So we can rewrite the first eq. as
$$p_{k,n} = \int_0^1 \frac{(\theta m)_k ((1-\theta)m)_{n-k}}{(m)_n} F_m(d\theta)$$
- where F_m is the distribution function concentrated on $\{r/m : 0 \leq r \leq m\}$ whose jump at r/m is q_r.