Bayesian Inference

STA 427/527, Fall 2019, Xin Wang

Probability

- An **experiment** is the process by which an observation/outcome is made.
- An **outcome** of an experiment is any possible observation of that experiment. Outcomes are often called sample points.
- The **sample space** of an experiment is the set consisting of all possible sample points. (S)
- An **event** is a set of outcomes of an experiment, or a subset of S. That is a set of sample points.

ex) flip a coin → outcome : heads, tails

$S = \{\text{Heads, tails}\}$

toss a die → outcome : 1, 2, 3, 4, 5, 6

$S = \{1, 2, 3, 4, 5, 6\}$

→ Event → rolling even # $A = \{2, 4, 6\}$

- **Probability:** Suppose S is a sample space, A is a subset of S. A probability measure $P(\cdot)$ is a function that maps events in S to real numbers.

→ Properties of a probability function

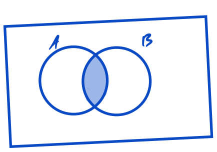
→ 1) For any event A, $P(A) \geq 0$

2) $P(S) = 1$

3) For any countable collection $A_1, A_2, \ldots$ (finite sequence of events) of pairwise mutually exclusive events ($A_i \cap A_j = \emptyset$ $i \neq j$)

$P(A_1 \cup A_2 \cup \cdots) = P(A_1) + P(A_2) + \cdots$

- Consider $P(A \cap B)$



→ Different ways to write this

$$\boxed{\begin{aligned} P(A \cap B) &= P(A|B)\,P(B) \\ &= P(B|A)\,P(A) \end{aligned}}$$

→ Conditional probability

→ $P(A|B) = \dfrac{P(A \cap B)}{P(A)}$

$P(B|A) = \dfrac{P(A \cap B)}{P(B)}$

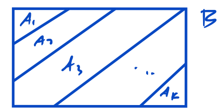
- **Bayes' theorem**

$\rightarrow P(A|B) = \dfrac{P(A\cap B)}{P(B)}$ $\qquad \rightarrow P(B|A) = (\ldots)$

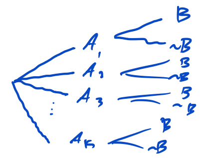
$\boxed{= \dfrac{P(B|A)\,P(A)}{P(B)}}$

Conditional probability is very important in Bayesian statistics $\Longrightarrow$ need to know the relationships between joint, marginal, + conditional dists

$\rightarrow$ Extension $\rightarrow$ $B = A_1 \cup A_2 \cup \ldots \cup A_k$ $\qquad A_i \cap A_j = \emptyset \quad i \neq j$



$$\rightarrow P(A_i|B) = \frac{P(B|A_i)\,P(A_i)}{P(B)} = \frac{P(B|A_i)\,P(A_i)}{\sum P(B\cap A_i)} = \frac{P(B|A_i)\,P(A_i)}{\sum P(B|A_i)\,P(A_i)}$$



- **Example:** The prevalence of heart disease in a certain population is 10%. A screening test for heart disease has 99% sensitivity and 90% specificity, where sensitivity measures the probability of positive that is correctly identified as such and specificity measures the probability of negative that is correctly identified as such. Suppose everyone in the population is given the screening test. What is the probability that one individual with positive test result actually has heart disease?

$\rightarrow$ Define events $\rightarrow$ A = has heart disease
B = positive test result

$\rightarrow P(A|B) = ??$

$\rightarrow$ start with what we know

$\rightarrow P(A) = 0.1 \Longrightarrow P(\sim A) = 0.9$

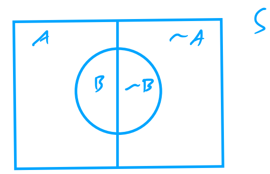
$P(B|A) = 0.99$

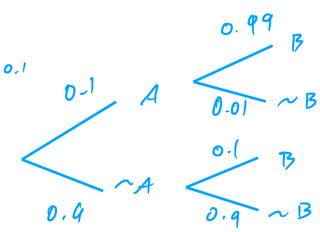
$P(\sim B|\sim A) = 0.9$

$$P(A|B) = \frac{P(A\cap B)}{P(B)} = \frac{P(B|A)\cdot P(A)}{P(B|A)\,P(A) + P(B|\sim A)\cdot P(\sim A)}$$

$\hookrightarrow = 1 - P(\sim B|\sim A) = 0.1$ $\quad = 0.9$

$$= \frac{0.99 * 0.1}{0.99 * 0.1 + 0.1\,(0.9)} = 0.524$$





Frequentist vs Bayesian

Scenario → • Suppose θ is the unknown parameter, and $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)$ are observations.

- Frequentist

  1. parameter θ is an unknown fixed value
  2. use $\underset{\sim}{y}$ to estimate θ, denoted $\hat{\theta}$
  3. $\hat{\theta}$ is a random variable, $\hat{\theta}(y)$

     → Different $\hat{\theta}$s for different sets of $\underset{\sim}{y}$
  4. The inference is based on the assumption that the data are "repeatable"

  → Example → $Y_i \sim N(\mu, \sigma^2)$, $i = 1, \ldots, n$, σ^2 is unknown

  observations $Y_1, Y_2, \ldots, Y_n$

  Construct a confidence interval for μ $100(1-\alpha)\%$ $(\alpha = 0.05)$

  $$\bar{Y} \pm t_{n-1, \alpha/2} \frac{s}{\sqrt{n}} \implies \left[\bar{Y} - t_{n-1,\alpha/2}\frac{s}{\sqrt{n}}, \; \bar{Y} + t_{n-1,\alpha/2}\frac{s}{\sqrt{n}}\right]$$

  95% CI deals w/ foundation of inference

  (i.e. how we collected data + estimate $\hat{\theta}$)

- Bayesian

1. θ is a random variable (prior)

   → This is the biggest difference in assumptions to the frequentist approach

2. How can we make inference about θ?

   → In order to make inference on θ, we need to have a posterior distribution

   → First we have a general idea of the shape of the distribution of θ (before collecting data)

   → This is called a prior

   → Then we get observations $\underset{\sim}{y} = (y_1, \ldots, y_n)^T$

   → Then we update the prior w/ the new data by finding the distribution of θ conditional on the given dataset $\underset{\sim}{y}$ (i.e. the posterior)

  - Steps in Bayesian

    1. Data distribution (likelihood)

       → specify a model $p(\underset{\sim}{y}|\theta)$

       → Likelihood example → $Y_i \overset{iid}{\sim} N(\mu, \sigma^2)$, σ^2 is known → $(Y_1, \ldots, Y_n)$

       → $L(\mu|\underset{\sim}{y}) = \prod_{i=1}^{n} p(Y_i|\mu)$

       $= p(\underset{\sim}{y}|\mu)$ → data distribution (i.e. our observations) are based on an unknown parameter

    2. Prior: → prior knowledge of θ → $\pi(\theta)$

    3. Posterior ☆ → $p(\theta|\underset{\sim}{y}) = \dfrac{p(\underset{\sim}{y}|\theta)\,\pi(\theta)}{p(\underset{\sim}{y})}$

       → marginal of data → $p(\underset{\sim}{y}) = \int_\Theta p(\underset{\sim}{y}, \theta)\, d\theta = \int_\Theta p(\underset{\sim}{y}|\theta)\,\pi(\theta)\, d\theta$

       "omega" = parameter space

       $P(A \cap B) = P(B|A)P(A)$

       → in practice → $\boxed{p(\theta|\underset{\sim}{y}) \propto p(\underset{\sim}{y}|\theta)\,\pi(\theta)}$

       proportional to

Bayes theorem

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

$A \to \theta$, $B \to \underset{\sim}{y}$

$$\boxed{p(\theta|\underset{\sim}{y}) = \frac{p(\underset{\sim}{y}|\theta)\,\pi(\theta)}{p(\underset{\sim}{y})}}$$

$\pi(\theta)$ → prior

$p(\underset{\sim}{y})$ → normalizing constant

- **Example:** A novelty coin company produces coins with varying levels of bias. For an individual coin, the probability of spinning "heads" is θ, where θ is drawn from a Uniform(0, 1) distribution. Consider the following thought experiment. Suppose it were possible to spin each coin in the population 100 times. Let Y represent the number of "heads" resulting from a coin's 100 spins.

  1. Find the distribution of Y. That is, what is the distribution of the numbers of "heads" observed in 100 spins?

→ setup → flip a coin 100 times

$Y = \#$ heads

$p(y) = ??$ → This is often difficult to find + computationally expensive if using numerical methods

→ $p(y) = \int_0^1 \underbrace{p(y, \theta)}_{= p(y \mid \theta)\pi(\theta)} \, d\theta$

→ [data dist] → $Y \mid \theta \sim \text{Binomial}(100, \theta)$ → choose this based on situation or shape of data after collection

$p(y \mid \theta) = \binom{100}{y} \theta^y (1-\theta)^{100-y}, \quad y = 0, 1, \ldots, 100$

→ [prior dist] → $\pi(\theta) = 1, \quad 0 < \theta < 1$

→ uniformative prior

→ marginal of data → $p(y) = \int_0^1 \binom{100}{y} \theta^y (1-\theta)^{100-y} \, d\theta$

$$= \binom{100}{y} \int_0^1 \theta^y (1-\theta)^{100-y} \, d\theta$$

$$= \binom{100}{y} \int_0^1 \theta^{(y+1)-1} (1-\theta)^{(101-y)-1}$$

$$= \binom{100}{y} \left[\frac{\Gamma(y+1)\Gamma(101-y)}{\Gamma(y+1+101-y)} \right]$$

$$= \binom{100}{y} \left[\frac{\Gamma(y+1)\Gamma(101-y)}{\Gamma(102)} \right]$$

$$= \frac{100!}{\cancel{(100-y)!}\cancel{y!}} \left[\frac{\cancel{y!}\cancel{(100-y)!}}{101!} \right]$$

$$= \frac{1}{101}, \quad y = 0, 1, 2, \ldots, 100$$

$\sim$ Discrete uniform $(0, 100)$

if $X \sim \text{Beta}(\alpha, \beta) \quad 0 < x < 1$

→ $f(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1} (1-x)^{\beta-1}$

→ $1 = \int_0^1 f(x) \, dx$

$$= \int_0^1 \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1} (1-x)^{\beta-1} \, dx$$

$$= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \int_0^1 x^{\alpha-1} (1-x)^{\beta-1} \, dx$$

$$\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} = \int_0^1 x^{\alpha-1} (1-x)^{\beta-1} \, dx$$

$\Gamma(x) = (x-1)!$ if $x \in \mathbb{Z}^+$ positive integer

$+ \; \Gamma(0) = 1$

2. Now think about the case that exactly 50 "heads" are observed in the coins we observed. Among these coins, determine the distribution of θ, the actual probabilities of "heads".

→ [posterior dist] → $p(\theta|y) = \dfrac{p(y|\theta)\,\pi(\theta)}{p(y)}$

$$= \frac{\left[\binom{100}{y}\theta^{y}(1-\theta)^{100-y}\right]\cdot 1}{1/101}$$

$$= 101\left(\frac{100!}{(100-y)!\,y!}\right)\theta^{y}(1-\theta)^{100-y}$$

$$= \frac{101!}{(100-y)!\,y!}$$

$$= \frac{\Gamma(102)}{\Gamma(101-y)\,\Gamma(y+1)}\,\theta^{(y+1)-1}\,(1-\theta)^{(101-y)-1}$$

$$\sim Beta(y+1, 101-y)$$

Showing exact calculation of the posterior (i.e. not using proportionality)
→ much easier in practice to use proportionality
+ just looks for kernel

→ If $y=50$ → $\theta|y \sim Beta(51, 51)$

→ [posterior dist] → $p(\theta|y) \propto p(y|\theta)\,\pi(\theta)$
★ (in practice)

$$= \left[\binom{100}{y}\theta^{y}(1-\theta)^{100-y}\right]\cdot 1$$

$$\propto \theta^{y}(1-\theta)^{100-y}$$

$$= \theta^{(y+1)-1}(1-\theta)^{(101-y)-1}$$

$$\sim Beta(y+1, 101-y)$$

Just a constant (w/ respect to θ) ⇒ drop it

→ Kernel

$X \sim Beta(\alpha, \beta)$ $\quad 0 < x < 1$, $\alpha, \beta > 0$

→ $f(x) = \dfrac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\, x^{\alpha-1}(1-x)^{\beta-1}$

Kernel of Beta w/ shape parameters α & β

→ $E(x) = \dfrac{\alpha}{\alpha+\beta}$

→ [posterior mean] → $E(\theta|y) = \dfrac{y+1}{(y+1)+(101-y)}$
(expected value)

$$= \frac{y+1}{102}$$

↳ Bayesian point estimate

Frequentist approach

→ [Data dist] → $p(y|\theta) = \binom{100}{y}\theta^{y}(1-\theta)^{100-y}$

→ [log-likelihood] → $\ell(\theta|y) = \log\left[\binom{100}{y}\theta^{y}(1-\theta)^{100-y}\right]$

$$= \log\binom{100}{y} + y\log(\theta) + (100-y)\log(1-\theta)$$

→ < ... > → < ... derivative & set to zero ... >

→ point estimate → $\hat{\theta}_{MLE} = \dfrac{y}{100}$