Bayesian Statistics

Basics

Nan Lin
Department of Mathematics
Washington University in St. Louis

Bayes' theorem

- $P(A|B) = \frac{P(B|A)P(A)}{P(B)} = \frac{P(B|A)P(A)}{P(B|A)P(A)+P(B|\bar{A})P(\bar{A})}$
- Suppose we have a model $y|\theta \sim p(y|\theta)$
  - y: data
  - θ: parameter
- Prior distribution: $p(\theta)$
  - Very often, the notation $\pi(\theta)$ is used
- Posterior distribution:
  - $p(\theta, y) = p(\theta)p(y|\theta)$
  - $p(y) = \sum_\theta p(\theta)p(y|\theta)$
  - $p(\theta|y) = \frac{p(\theta,y)}{p(y)} = \frac{p(\theta)p(y|\theta)}{p(y)} \propto p(\theta)p(y|\theta)$

Likelihood principle

- <u>Likelihood Principle</u>. In the inference about θ, after y is observed, all relevant experimental information is contained in the likelihood function for the observed y. Furthermore, two likelihood functions contain the same information about θ if they are proportional to each other.
- Consider testing the fairness of a coin.

$$H_0: \theta = \frac{1}{2} \, vs \, H_1: \theta > \frac{1}{2}$$

- Data: An experiment is conducted and 9 heads and 3 tails are observed.

Two possible experiments

- Binomial: 12 toss in total
- Negative binomial: keep tossing until getting three tails

- Likelihoods are proportional
- Conclusions based on pvalues are contradictive → violation of likelihood principle
- Bayesian method has no difficult → the same conclusion under both scenarios

What did Bayes solve initially? → Binomial model

Thomas Bayes

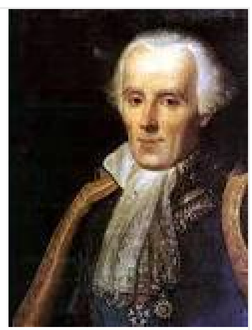
Thomas Bayes was an English mathematician and Presbyterian minister, known for having formulated a specific case of the theorem that bears his name: Bayes' theorem. Wikipedia

Born: 1701, London

Died: April 7, 1761, Royal Tunbridge Wells

Education: University of Edinburgh

Pierre-Simon Laplace



Pierre-Simon, marquis de Laplace was a French mathematician and astronomer whose work was pivotal to the development of mathematical astronomy and statistics. Wikipedia

Born: March 23, 1749, Beaumont-en-Auge

Died: March 5, 1827, Paris

Education: Caen University

Spouse: Marie-Charlotte de Courty de Romanges

Books: A philosophical essay on probabilities

- A ball W is randomly thrown (according to a uniform distribution) on a rectangular table. The horizontal position of the ball on the table is θ, expressed as a fraction of the table width.
- A ball O is randomly thrown n times. The value of y is the number of times O lands to the right of W.
- Question: What is the "inverse probability" $P(\theta_1 < \theta < \theta_2|y)$?

In the Bayesian language

- Prior distribution of θ: $U[0,1]$
- Likelihood: $p(y|\theta)$, i.e. $y|\theta \sim Binomial(n,\theta)$
- Posterior probability

$$P(\theta_1 < \theta < \theta_2|y) = \frac{P(\theta_1 < \theta < \theta_2, y)}{p(y)}$$

$$= \frac{\int_{\theta_1}^{\theta_2} p(y|\theta)p(\theta)d\theta}{p(y)} = \frac{\int_{\theta_1}^{\theta_2} \binom{n}{y} \theta^y (1-\theta)^{n-y} d\theta}{p(y)}$$

- Marginal distribution: Bayes succeeded in evaluating the denominator, for $y = 0, \ldots, n$,

$$p(y) = \int_0^1 \binom{n}{y} \theta^y (1-\theta)^{n-y} d\theta = \frac{1}{n+1}$$

  - All possible values of y are equally likely *a priori*

Laplace's application

- Estimate the proportion of female births in a population.
- A total of 241,945 girls and 251,527 boys were born in Paris from 1745 to 1770.
- What is the probability the female birth rate is above 50%?
- Let θ be the probability that any birth is female, Laplace showed that

$$P(\theta \geq 0.5 | y = 241945, n = 251945 + 251527) \approx 1.15 \times 10^{-42}$$

Posterior distribution

$$p(\theta|y) \propto p(y|\theta)p(\theta) = \theta^y(1-\theta)^{n-y}$$

- What is it?
  - In general, we will need to find the normalizing constant $c^{-1} = \int \theta^y(1-\theta)^{n-y} d\theta$. But generally, this can be difficult to solve.
  - Alternative solution: Look up among commonly known probability distributions
    - Here, we can see this is a beta distribution, $beta(n+1, n-y+1)$
  - What if it does not belong to any commonly known distribution?
    - Use simulation
    - But how do we simulate from a distribution when we do not know the normalizing constant?

Prediction

- Interested in the outcome of one new trial
- Let $\tilde{y}$ denote the outcome of a new trial, and it is exchangeable with the previous n trials
  - Exchangeability: n values of y_i are regarded as exchangeable if the join probability density $p(y_1, \dots, y_n)$ is invariant to permutations of the indexes.
    - Independently and identically distributed (i.i.d.) random variables are exchangeable
- Predictive distribution:

$$P(\tilde{y} = 1|y) = \int_0^1 p(\tilde{y} = 1|\theta, y)p(\theta|y)d\theta = \int_0^1 \theta p(\theta|y)d\theta$$

$$= E(\theta|y) = \frac{y+1}{n+2}$$

Some general facts about Bayesian inference

- On average, posterior distribution is less variable than the prior distribution
  - $var(\theta) = E\big(var(\theta|y)\big) + var\big(E(\theta|y)\big) \geq E\big(var(\theta|y)\big)$
  - Prior variance: $var(\theta)$
  - Posterior variance: $var(\theta|y)$
- Sequential updates in Bayesian inference
  - Prior: $p(\theta)$
  - After the first batch data of $y_1 \rightarrow p(\theta|y_1) \propto p(y_1|\theta)p(\theta)$
  - After the second batch data of y_2 (assume it is conditionally independent with y_1)$\rightarrow p(\theta|y_1, y_2) \propto p(y_2|\theta)p(\theta|y_1) \propto p(y_2|\theta)p(y_1|\theta)p(\theta) = p(y_1, y_2|\theta)p(\theta)$
  - This is the same as if we have y_1, y_2 together

Some general facts about Bayesian inference (cont)

- Inference can be performed based on the sufficient statistics
  - Sufficient statistics
    - Heuristic definition: We say T is a sufficient statistic if the statistician who knows the value of T can do just as good a job of estimating the unknown parameter θ as the statistician who knows the entire random sample.
    - Mathematical definition: A statistic T is a *sufficient statistic* if for each t, the conditional distribution of data given $T = t$ and θ does not depend on θ.
- For example, $y|\theta \sim Binomial(n, \theta)$, this model can be viewed as summarized from i.i.d. $Bernoulli(\theta)$ random variables $x_1, \dots, x_n$, where $y = x_1 + x_2 + \cdots + x_n$. One can show that $\theta|y$ and $\theta|x_1, \dots, x_n$ have the same distribution.

How to identify sufficient statistics?

- Factorization theorem
  - Let $X_1, \dots, X_n$ be a random sample (i.i.d. random variables) from a distribution with density $p_\theta(x)$. Then $T(X_1, \dots, X_n)$ is a sufficient statistic of θ <u>if and only if</u>

$$\prod_{i=1}^{n} p_\theta(x_i) = g(T, \theta) h(X_1, \dots, X_n),$$

  where
    - $g(T, \theta)$ depends on the data only through the statistic T,
    - $h(X_1, \dots, X_n)$ depends on the data but is the same for every θ.
- Example: $\bar{X}$ is a sufficient statistic of μ for data from $N(\mu, \sigma^2)$ if σ^2 is known.

- Consider $y|\theta \sim p(y|\theta)$ and T is a sufficient statistic of θ
- By factorization theorem, we can write $p(y|\theta) = p(T|\theta)h(y)$
- Then

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)} = \frac{p(T|\theta)h(y)p(\theta)}{\int p(T|\theta)h(y)p(\theta)d\theta}$$

$$= \frac{p(T|\theta)p(\theta)}{\int p(T|\theta)p(\theta)d\theta} = \frac{p(T|\theta)p(\theta)}{p(T)} = p(\theta|T)$$

Some important results from probability theory

- Conditional expectation: $E(X) = E[E(X|Y)]$
- Conditional variance: $var(X) = E[var(X|Y)] + var[E(X|Y)]$
- Change-of-variable formula
  - $Y = g(X)$: one-to-one transformation

$$p_y(\mathbf{y}) = p_x(\mathbf{x})|\det\left(\frac{\partial \mathbf{x}}{\partial \mathbf{y}}\right)| = p_x(\mathbf{x})|\det \mathbf{J}_{\mathbf{y}\to\mathbf{x}}| = p_x(\mathbf{x})|J_{\mathbf{y}\to\mathbf{x}}|$$

  - Jacobian matrix

$$\mathbf{J}_{\mathbf{x}\to\mathbf{y}} \stackrel{\text{def}}{=} \frac{\partial(y_1,\ldots,y_m)}{\partial(x_1,\ldots,x_n)} \stackrel{\text{def}}{=} \begin{pmatrix} \frac{\partial y_1}{\partial x_1} & \cdots & \frac{\partial y_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial y_m}{\partial x_1} & \cdots & \frac{\partial y_m}{\partial x_n} \end{pmatrix}$$

Simulate normal random variables: Box-Muller transformation

For the Box-Muller transform, we require two random variables U, V, uniformly distributed on $[0, 1]$. Set

$$R = \sqrt{-2 \log V} \qquad \text{and} \qquad \Theta = 2\pi U.$$

and

$$Z_1 = R \cos \Theta = \sqrt{-2 \log V} \cos(2\pi U), \qquad \text{and} \qquad Z_2 = \sqrt{-2 \log V} \sin(2\pi U).$$

Then X and Y are independent standard normal random variables. To obtain two standard normal random variables with correlation ρ, take

$$X = Z_1 \qquad \text{and} \qquad Y = \rho Z_1 + \sqrt{1 - \rho^2} Z_2.$$