IFI 9000 Analytics Methods
Bayesian Statistics

by **Houping Xiao**

Spring 2021

Georgia State University | J. MACK ROBINSON COLLEGE OF BUSINESS

Bayesian Statistics

Bayesian statistics [url, b] is a mathematical procedure that:

- applies probabilities to statistical problems
- provides people the tools to **update their beliefs**
- in the evidence of **new data**

And, Bayesian statistics is built based on:

- Bayes theorem
- conditional probability

Example 1:

- Out of all the 4 championship races (F1) between Niki Lauda and James Hunt
    - Niki won 3 times
    - James won only 1 time
- If you were to bet on the winner of next race, who would he be?

New information:

- It rained once when James won and once when Nike Won
- It is definite that it will rain on the next date
- Who would you bet your money on now?

Example 2: Cancer diagnosis [url, a]

- Assume that 1% of a population have cancer
- A screening test has 80% sensitivity and 95% specificity
- Given a person have a positive result

What is the change that this person actually has the cancer?

- $\mathbb{P}(cancer|positiveresult) \approx 14\%$
- Most positive results are actually false alarms
- Sensitivity:
  - True positive rates
  - Given a person has cancer, the chance that the test will say positive
- Specificity:
  - True negative rates
  - Given a person does not has cancer, the chance that the test will say negative

Bayesian Statistics via an Example

Example 2: Cancer diagnosis [url, a]

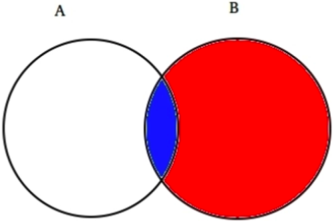
- Assume that 1% of a population have cancer Prior Knowledge
- A screening test has 80% sensitivity and 95% specificity Data
- Given a person have a positive result Data

What is the change that this person actually has the cancer?

- $\mathbb{P}(cancer|positiveresult) \approx 14\%$ Updated belief
- Most positive results are actually false alarms
- Sensitivity:
  - True positive rates
  - Given a person has cancer, the chance that the test will say positive
- Specificity:
  - True negative rates
  - Given a person does not has cancer, the chance that the test will say negative

Conditional Probability and Bayes Theorem

- **Conditional Probability**: Probability of an event A given B equals the probability of B and A happening together divided by the probability of B



$$\mathbb{P}(A|B) = \frac{\text{Blue Area}}{\text{Blue Area + Red Area}}$$

Conditional Probability and Bayes Theorem

- $\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$
- $\mathbb{P}(B|A) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)} \Rightarrow \mathbb{P}(A \cap B) = \mathbb{P}(B|A)\mathbb{P}(A)$

Bayes' Theorem tell us that

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

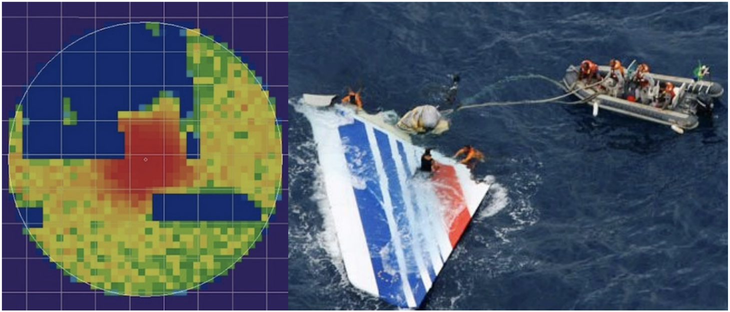
- **Example 1**
    - A: Niki wins; $\mathbb{P}(A) = \frac{1}{4}$
    - B: Event of raining; $\mathbb{P}(B) = \frac{2}{4}$
    - $\mathbb{P}(B|A) = 1$

Bootstrap General Framework

- Suppose a black-box calculates $\hat{\alpha}$ from a sample set, e.g., a coefficient in linear or logistic regression
- We are interested in estimating the variability of $\hat{\alpha}$ without examining many new sample sets
- Denoting the first bootstrap data set by Z^{*1}, we use Z^{*1} to produce a new bootstrap estimate for α, which we call $\hat{\alpha}^{*1}$
- This procedure is repeated B (say 100 or 1,000) times, in order to produce B different bootstrap data sets, $Z^{*1}, Z^{*2}, \cdots, Z^{*B}$, and the corresponding α estimates $\hat{\alpha}^{*1}, \cdots, \hat{\alpha}^{*B}$

Bayesian Inference

Bayes' Theorem used in practice



- During the search for Air France 447, from 2009-2011, knowledge about the black box location was described via probability – i.e., using Bayesian inference
- Eventually, the black box was found in the read area

Bayesian Inference

Bayesian inference can help us to:

- update knowledge, as data is obtained

It can also be used for

- parameter estimation
- density estimation
- regression function estimating
- ...

Bayesian Inference

Formally:

- **Prior distribution** $\pi(\beta)$: what you know about parameter β, excluding the information in the data
- **Likelihood** $f(y|\beta)$: based on modeling assumptions, how likely to observe y if the truth is β
- **Posterior distribution** stating what we know about β, combining the prior with the data:

$$\mathbb{P}(\beta|y) \propto f(y|\beta) \times \pi(\beta)$$

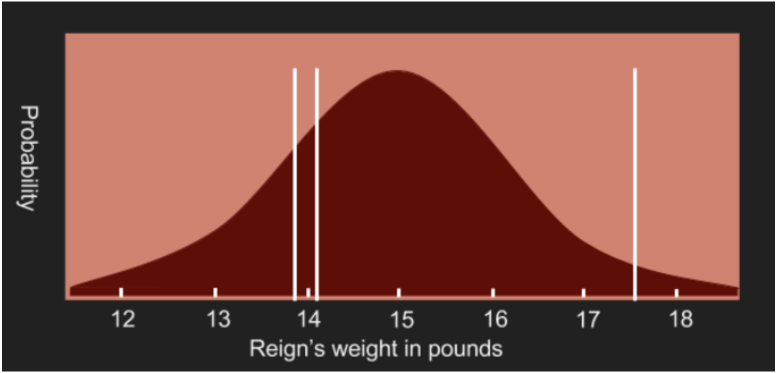
$$\textit{posterior} \propto \textit{likelihood} \times \textit{prior}$$

Bayesian Inference - How much does she weigh?

How much does she weigh?

- Three measures for a dog: 13.9 lb, 17.5 lb, and 14.1 lb

Likelihood: $f(x_1, x_2, x_3|\mu, \sigma^2) = \phi(\frac{x_1-\mu}{\sigma})\phi(\frac{x_2-\mu}{\sigma})\phi(\frac{x_3-\mu}{\sigma})$



Bayesian Inference - How much does she weigh?

$$\mathbb{P}(\mu|m) = \frac{\mathbb{P}(m|\mu)\mathbb{P}(\mu)}{\mathbb{P}(m)}$$

- m - measurement; μ - weight
- $\mathbb{P}(\mu)$ - prior
- $\mathbb{P}(m\mu)$ - likelihood
- $\mathbb{P}(\mu|m)$ - posterior

Start by making assumptions:

- assume dog's weight is equally likely to be 13 pounds or 15 pounds or 1 pounds or 1,000,000 pounds
- assume a uniform prior: $\mathbb{P}(\mu)$ is constant for all values

So by Bayes' Theorem: $\mathbb{P}(\mu|m) = \mathbb{P}(m|\mu)$

$$\prod_{i=1}^{3} \phi(\frac{x_i - \mu}{\sigma}) = \frac{1}{\sigma^3(\pi)^{3/2}} \exp\left(-\frac{\sum_{i=1}(x_i - \mu)^2}{2\sigma^2}\right)$$

Bayesian Inference - How much does she weigh?

Last time she weighted 14.2 pounds

- assume an prior $\mu \sim \mathcal{N}(14.2, 0.5^2)$

So by Bayes' Theorem: $\mathbb{P}(\mu|m) \propto \mathbb{P}(m|\mu)\mathbb{P}(\mu)$

- assume dog's weight is equally likely to be 13 pounds or 15 pounds or 1 pounds or 1,000,000 pounds
- assume a uniform prior: $\mathbb{P}(\mu)$ is constant for all values

So by Bayes' Theorem: $\mathbb{P}(\mu|m) = \mathbb{P}(m|\mu)$

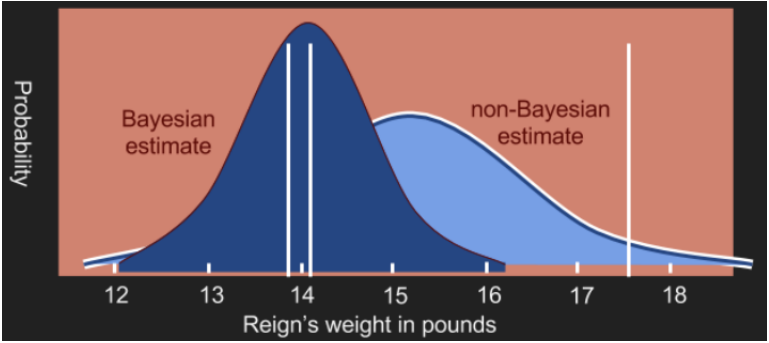
$$\prod_{i=1}^{3} \phi(\frac{x_i - \mu}{\sigma})\phi(\frac{\mu - 14.2}{0.5})$$

$$= \frac{1}{\sigma^3(\pi)^{3/2}} \exp\left(-\frac{\sum_{i=1}(x_i - \mu)^2}{2\sigma^2} - \frac{\sum_{i=1}(\mu - 14.2)^2}{2 \times 0.5^2}\right)$$

Bayesian Inference - How much does she weigh?

Bayesian vs. not

- The Bayesian estimate ignores 17.5 lb like an outlier
- The distribution is narrower. Confidence is greater
- The answer is probably much closer to correct



Ordinary Least Squares (OLS)

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

- $y = \boldsymbol{X}\beta + \epsilon$
- $RSS(\beta) = \sum_{i=1}^{n}(y_i - \beta^\top x_i)^2$
- $\hat{\beta} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top y$

Bayesian Linear Regression

Bayesian linear regression:

$$y_i = x_i^\top \beta + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

The likelihood:

$$p(y|\boldsymbol{X}, \beta) \propto (\sigma^2)^{-n/1)} \exp\left(-\frac{1}{2\sigma^2}(y - \boldsymbol{X}\beta)^\top (y - \boldsymbol{X}\beta)\right)$$

The prior $\beta|\sigma^2 \sim \mathcal{N}(\mu_0, \sigma^2 \Lambda_0^{-1})$:

$$p(\beta|\sigma^2) \propto (\sigma^2)^{-k/2} \exp\left(-\frac{1}{2\sigma^2}(\beta - \mu_0)^\top \Lambda_0 (\beta - \mu_0)\right)$$

The posterior:

$$p(\beta|y, \boldsymbol{X}) \propto p(\beta|\sigma^2) p(y|\boldsymbol{X}, \beta)$$

Bayesian Linear Regression

The prior $\beta|\sigma^2 \sim \mathcal{N}(\mu_0, \sigma^2\Lambda_0^{-1})$:

$$p(\beta|\sigma^2) \propto (\sigma^2)^{-k/2} \exp\left(-\frac{1}{2\sigma^2}(\beta-\mu_0)^\top\Lambda_0(\beta-\mu_0)\right)$$

The posterior:

$$(\sigma^2)^{-(k+n)/2} \exp\left(-\frac{1}{2\sigma^2}(\beta-\mu_0)^\top\Lambda_0(\beta-\mu_0) - \frac{1}{2\sigma^2}(y-\boldsymbol{X}\beta)^\top(y-\boldsymbol{X}\beta)\right)$$

- $p(\beta|y,\boldsymbol{X}) \sim \mathcal{N}(\mu_n, \Lambda_n)$
- $\Lambda_n = (\boldsymbol{X}^\top\boldsymbol{X} + \Lambda_n)$, $\mu_n = \Lambda_n^{-1}(\boldsymbol{X}^\top\boldsymbol{X}\hat{\beta} + \Lambda_0\mu_0)$
- $(\beta-\mu_0)^\top\Lambda_0(\beta-\mu_0)+(y-\boldsymbol{X}\beta)^\top(y-\boldsymbol{X}\beta) = (\beta-\mu_n)^\top\Lambda_n(\beta-\mu_n)+C$

In the example above, the posterior of β follows a known distribution (multivariate normal)

- Posterior inference is straightforward
- posterior mean, variance
- hypothesis testing

Summarizing the posterior involves integrals.

- For simple problems, this can be done with pencil and paper
- For hard problems, we usually use MCMC

MCMC Sampling

Markov Chain Monte Carlo (MCMC) sampling is

- the predominant method for Bayesian inference
- approximate the posterior by drawning samples from the posterior distribution

E.g.

- posterior mean can be approximated by the sample mean of MCMC samples
- posterior sd can be approximated by the sd of the MCMC samples
- percentile, confidence interval, etc.

Shrinkage Prior and Bayesian Lasso

The lasso estimates:

$$\hat{\beta}^{lasso} = \arg\min_{\beta} ||y - \boldsymbol{X}\beta||_2^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

- $||y - \boldsymbol{X}\beta||_2^2$ - goodness of fit
- $\lambda \sum_{j=1}^{p} |\beta_j|$ - penalty

Tibshirani (1996):

- lasso estimates can be viewed as the posterior mode
- when β's follow iid Laplace (or double-exponential) priors

Shrinkage Prior and Bayesian Lasso

The likelihood:

- $p(y|\beta, \sigma^2) = \mathcal{N}(y|\boldsymbol{X}\beta, \sigma^2 \boldsymbol{I}_n)$

The prior:

- $p(\beta|\gamma) = (\tau/2)^p \exp(-\tau||\beta||)_1$

The lasso estimates equal the mode of the posterior distribution of β

$$\hat{\beta}_L = \arg\max_{\beta} \quad p(\beta|y, \sigma^2, \tau)$$

- n is the sample size, p is the number of covariates

The regression example

How Bayesian statistics convinced me to hit the gym.

- weight as a function of height
- weight percentile

The complete example can be found here:
https://towardsdatascience.com/how-bayesian-statistics-convinced-me-to-hit-the-gym-fa737b0a7ac
More examples for using JAGS:
https://www4.stat.ncsu.edu/reich/ABA/notes/JAGS.pdf

Bayesian Lasso Example

Election prediction using census data.

References:

- https://www4.stat.ncsu.edu/reich/ABA/code/BLASSO
- https://github.com/ncsu-statistics/bayesian-learning-with-R

Naive Bayesian Classifier

	Age	Income	Student	Credit	Buys_computer
P1	31…40	high	no	fair	no
P2	<=30	high	no	excellent	no
P3	31…40	high	no	fair	yes
P4	>40	medium	no	fair	yes
P5	>40	low	yes	fair	yes
P6	>40	low	yes	excellent	no
P7	31…40	low	yes	excellent	yes
P8	<=30	medium	no	fair	no
P9	<=30	low	yes	fair	yes
P10	>40	medium	yes	fair	yes

Naive Bayesian Classifier

	Age	Income	Student	Credit	Buys_computer
P1	31…40	high	no	fair	no
P2	<=30	high	no	excellent	no
P3	31…40	high	no	fair	yes
P4	>40	medium	no	fair	yes
P5	>40	low	yes	fair	yes
P6	>40	low	yes	excellent	no
P7	31…40	low	yes	excellent	yes
P8	<=30	medium	no	fair	no
P9	<=30	low	yes	fair	yes
P10	>40	medium	yes	fair	yes

- To classify means to determine the highest $P(H_i|X)$ among all classes $C_1, \cdots, C_m$
  - If $P(H_1|X) > P(H_0|X)$, then X buys computer
  - If $P(H_0|X) > P(H_1|X)$, then X does not buy computer
  - Calculate $P(H_i|X)$ using the Bayes Theorem

$$P(H_i|X) = \frac{P(H_i)P(X|H_i)}{P(X)}$$

Naive Bayesian Classifier: Class Prior Probability

- $P(H_i)$ is class prior probability that X belongs to a particular class C_i
  - Can be estimated by n_i/n from training data samples
  - n is the total number of training data samples
  - n_i is the number of training data samples of class C_i

	Age	Income	Student	Credit	Buys_computer
P1	31…40	high	no	fair	no
P2	<=30	high	no	excellent	no
P3	31…40	high	no	fair	yes
P4	>40	medium	no	fair	yes
P5	>40	low	yes	fair	yes
P6	>40	low	yes	excellent	no
P7	31…40	low	yes	excellent	yes
P8	<=30	medium	no	fair	no
P9	<=30	low	yes	fair	yes
P10	>40	medium	yes	fair	yes

- H_1: $Buys\_computer = yes$, $P(H_1) = 6/10 = 0.6$
- H_0: $Buy\_computer = no$, $P(H_0) = 4/10 = 0.4$

$$P(H_i|X) = \frac{P(H_i)P(X|H_i)}{P(X)}$$

Naive Bayesian Classifier: Descriptor Prior Probability

- $P(X)$ is prior probability that X
  - Probability that observe the attribute values of X
  - Suppose $X = (x_1, \cdots, x_p)$ and they are independent, then $P(X) = P(x_1) \cdots P(x_p)$
  - $p(x_i) = \frac{n_i}{n}$, where n_i is the number of training samples having value x_i for feature A_i; n is the total number of training samples
  - constant for all classes

	Age	Income	Student	Credit	Buys_computer
P1	31…40	high	no	fair	no
P2	<=30	high	no	excellent	no
P3	31…40	high	no	fair	yes
P4	>40	medium	no	fair	yes
P5	>40	low	yes	fair	yes
P6	>40	low	yes	excellent	no
P7	31…40	low	yes	excellent	yes
P8	<=30	medium	no	fair	no
P9	<=30	low	yes	fair	yes
P10	>40	medium	yes	fair	yes

- $X = (age : 31 \cdot 40, income : medium, student : yes, credit : fair)$
- $P(age = 31 \cdot 40) = 3/10$, $P(income : medium) = 3/10$
- $P(student = yes) = 5/10$, $P(credit = fair) = 7/10$
- $P(X) = P(age = 31 \cdot 40) \cdot P(income : medium) \cdot P(student = yes) \cdot P(credit = fair) = 0.3 \cdot 0.3 \cdot 0.5 \cdot 0.7 = 0.0315$

$$P(H_i|X) = \frac{P(H_i)P(X|H_i)}{P(X)}$$

Naive Bayesian Classifier: Descriptor Posterior Probability

- $P(X|H_i)$ is posterior probability of X given H_i
  - Probability that observe X in class C_i
  - Suppose $X = (x_1, \cdots, x_p)$ and they are independent, then $P(X|H_i) = P(x_1|H_i) \cdots P(x_p|H_i)$
  - $p(x_i) = \frac{n_{i,j}}{n_i}$, where $n_{i,j}$ is the number of training samples in class C_i having value x_i for feature A_i; n_i is the total number of training samples in class $_{C\,i}$

	Age	Income	Student	Credit	Buys_computer
P1	31…40	high	no	fair	no
P2	<=30	high	no	excellent	no
P3	31…40	high	no	fair	yes
P4	>40	medium	no	fair	yes
P5	>40	low	yes	fair	yes
P6	>40	low	yes	excellent	no
P7	31…40	low	yes	excellent	yes
P8	<=30	medium	no	fair	no
P9	<=30	low	yes	fair	yes
P10	>40	medium	yes	fair	yes

- $X = (age : 31 \cdot 40, income : medium, student : yes, credit : fair)$
- $H_1 = X$ buys a computer
- $n_1 = 6, n_{11} = 2, n_{21} = 2, n_{31} = 4, n_{41} = 5$
- $P(X|H_1) = \frac{2}{6} \times \frac{2}{6} \times \frac{4}{6} \times \frac{5}{6} = \frac{5}{81} = 0.062$

$$P(H_i|X) = \frac{P(H_i)\, P(X|H_i)}{P(X)}$$

	Age	Income	Student	Credit	Buys_computer
P1	31…40	high	no	fair	no
P2	<=30	high	no	excellent	no
P3	31…40	high	no	fair	yes
P4	>40	medium	no	fair	yes
P5	>40	low	yes	fair	yes
P6	>40	low	yes	excellent	no
P7	31…40	low	yes	excellent	yes
P8	<=30	medium	no	fair	no
P9	<=30	low	yes	fair	yes
P10	>40	medium	yes	fair	yes

- $X = (age : 31 \cdot 40, income : medium, student : yes, credit : fair)$
- $H_0 = X$ does not buy a computer
- $n_0 = 4, n_{10} = 1, n_{20} = 1, n_{30} = 1, n_{40} = 2$
- $P(X|H_1) = \frac{1}{4} \times \frac{1}{4} \times \frac{1}{4} \times \frac{2}{4} = \frac{1}{128} = 0.0078$

$$P(H_i|X) = \frac{P(H_i)\, P(X|H_i)}{P(X)}$$

Bayesian vs Frequentist

Pros:

- Scientific knowledge incorporation via the prior
- More information for decision making
- Computationally easier for complex model, e.g., hierarchical models
- A framework to incorporate data/info from multiple sources

Cons:

- Picking a prior is subjective
- Computing can be slow or unstable for some problems

Reference: https://www4.stat.ncsu.edu/reich/ABA/notes/Intro.pdf

Reference

http://faculty.washington.edu/kenrice/BayesIntroClassEpi2018.pdf.

https://www.analyticsvidhya.com/blog/2016/06/bayesian-statistics-beginners-simple-english/.

The End