Introduction to Bayesian Statistics (I)

Xiaoquan (William) Wen

July 3, 2019
AN OVERVIEW

- Named after Thomas Bayes (1701 - 1761)

- What is Bayesian statistics
  - a mathematical procedure that applies probabilities to statistical problems
  - provides the tools to update people’s beliefs in the evidence of new data.

- Bayesian approach is trending in big data era
A brief history of Bayesian statistics

- 1700s, Bayes’ Theorem
- 1800s, Pierre-Simon Laplace formalized and popularized Bayesian inference
- 1940s, Alan Turing’s Bayesian system decoded German Enigma Machine, but in general Bayesianism is considered in decline
- 1960s, revival of the Bayes’ theorem: theory and computation work
- Current day practice:
  - Election prediction (FiveThirtyEight.com)
  - Aviation incidents investigations
  - Broadly used in medicine, economy and all branches of sciences
**Conditional Probability**

\[ \Pr(A \mid B) \]

- Probability of event A given event B has occurred

\[ \Pr(A \mid B) = \frac{\Pr(A \cap B)}{\Pr(B)} \]

- Fundamental in probability theory and statistics
Bayes’ Theorem

\[
\Pr(\theta \mid \text{data}) = \frac{\Pr(X) \Pr(\text{data} \mid X)}{\Pr(\text{data})} \\
= \frac{\Pr(X) \Pr(\text{data} \mid X)}{\Pr(\text{data})} \\
= \frac{\Pr(X) \Pr(\text{data} \mid X)}{\Pr(X) \Pr(\text{data} \mid X) + \Pr(X^c) \Pr(\text{data} \mid X^c)}
\]

- \(\Pr(X):\) prior probability/distribution
- \(\Pr(\text{data} \mid X):\) likelihood
- \(\Pr(X \mid \text{data}):\) posterior probability/distribution
Application of Bayes Theorem

If 1% of a population have a specific form of cancer, for a screening test with 80% sensitivity and 95% specificity, **What is the chance that a patient has the cancer if he tests positive?**

Sensitivity: $\text{Pr}(\text{test}^+ | \text{cancer}) = 80\%$

Specificity: $\text{Pr}(\text{test}^- | \text{no cancer}) = 95\%$

$\text{Pr}(\text{cancer} | \text{test}^+) = \frac{\text{Pr}(\text{cancer}) \times \text{Pr}(\text{test}^+ | \text{cancer})}{\text{Pr}(\text{test}^+)}$

$= \frac{0.01 \times 0.80}{0.01 \times 0.80 + 0.99 \times 0.05} 
\approx 13.9\%$
APPLICATION OF BAYES THEOREM

If 1% of a population have a specific form of cancer, for a screening test with 80% sensitivity and 95% specificity, What is the chance that a patient has the cancer if he tests positive?

- Sensitivity: $Pr(\text{test }+ \mid \text{ cancer}) = 80\%$
- Specificity: $Pr(\text{test }- \mid \text{ no cancer}) = 95\%$
APPLICATION OF BAYES THEOREM

If 1% of a population have a specific form of cancer, for a screening test with 80% sensitivity and 95% specificity, What is the chance that a patient has the cancer if he tests positive?

- Sensitivity: \( \Pr(\text{test + } | \text{ cancer}) = 80\% \)

- Specificity: \( \Pr(\text{test – } | \text{ no cancer}) = 95\% \)

\[
\Pr(\text{cancer } | \text{ test + } ) = \frac{\Pr(\text{cancer}) \Pr(\text{test + } | \text{ cancer})}{\Pr(\text{test + })} \\
= \frac{0.01 \times 0.80}{0.01 \times 0.80 + 0.99 \times 0.05} \\
\approx 13.9\%
\]
Most positive tests (≈ 86%) are actually false alarms

But is the prior Pr(cancer) = 0.01 reasonable to use here?
The Process of Bayesian Inference

The Bayesian Machinery

1. Define a parametric model (prior, likelihood)

2. Apply Bayes Theorem and compute the posterior for the parameters of interest

3. Posterior distributions contain full information of inference result
The Bayesian Philosophy

- The Bayesian inference process is a byproduct of multiple statistical principles

- They start from different perspectives and all conclude that statistical inference results should be summarized in form of posterior distributions

- This also leads to different interpretations of probabilities

  - Bayesian: probability is simply a quantification of uncertainty

  - Frequentist: probability reflects a long-run frequency
**Argument 1: Coherence of Decision Making**

- Need principled approach to make decision accounting for uncertainty

- Consider make a prediction, $\delta(x)$, with respect to an unknown parameter $\theta$ based on observed data $x$

- Coherent decision should be informed by the posterior distribution $p(\theta | x)$

- Inevitably, it requires a prior distribution and apply Bayes theorem
ARGUMENT 2: THE LIKELIHOOD PRINCIPLE

- Sufficiency Principle (S): irrelevance of observations independent of a sufficient statistic

- Conditionality Principle (C): irrelevance of (component) experiments not actually performed

- The voltmeter story: https://en.wikipedia.org/wiki/Likelihood_principle

- Likelihood Principle (L): irrelevance of outcomes not actually observed
The Likelihood Principle (cont’d)

- (almost) All statisticians accept S and C
- It has been shown (Birnbaum, 1962) that

\[ S + C \rightarrow L \]

i.e., all data scientists should accept L

- Bayesian inference process follows the likelihood principle

- Some commonly used frequentist procedures, e.g., p-values, confidence intervals, violate the likelihood principle
**Argument 3: Exchangeability**

Bayesian inference provides more **flexible** and **realistic** modeling options.

Consider tossing a coin with a sequence of outcomes: $X_1, X_2, \ldots$.  

- The random sequence is often modeled as *independent identically distributed (i.i.d)*.  
- Are the sequence of outcomes really independent? Note that, independence implies

\[
\Pr(X_1, X_2, \ldots, X_p) = \prod_{i=1}^{p} \Pr(X_i)
\]

\[
\Pr(X_p \mid X_1, X_2, \ldots, X_{p-1}) = \Pr(X_p)
\]
ARGUMENT 3: EXCHANGEABILITY

Bayesian inference provides more flexible and realistic modeling options.

Consider tossing a coin with a sequence of outcomes: $X_1, X_2, ...$

- The random sequence is often modeled as independent identically distributed (i.i.d).
- Are the sequence of outcomes really independent? Note that, independence implies

$$\Pr(X_1, X_2, ..., X_p) = \prod_{i=1}^{p} \Pr(X_i)$$

$$\Pr(X_p \mid X_1, X_2, ..., X_{p-1}) = \Pr(X_p)$$

- But the sequence of outcomes share the information on the biasness of the coin!
EXCHANGEABILITY (CONT’D)

- A more realistic modeling assumption is to treat the sequence *exchangeable*

- Mathematically, it means $\Pr(X_1, ..., X_p)$ is invariant to the permutations of indexes $(1, ..., p)$.

- An independent sequence is obviously exchangeable, but an exchangeable sequence does not need to be independent!
**De Finetti Theorem**

The de Finetti theorem indicates

\[
\Pr(X_1, \ldots, X_p) = \int \left[ \prod_{i=1}^{p} \Pr(X_i \mid \theta) \right] p(\theta) d\theta,
\]

for any exchangeable sequence.

- \(\theta\) represents the biasness of the coin
- Conditional on \(\theta\), the sequence is i.i.d

\[
\Pr(X_1, \ldots, X_p \mid \theta) = \prod_{i=1}^{p} \Pr(X_i \mid \theta)
\]

- Because \(\theta\) is unknown, the probability of the sequence has to be averaged over the uncertainty of \(\theta\) (a prior distribution!)
MODEL EXCHANGEABILITY

- Requires a prior distribution
- Give rise to a hierarchical model
- Hierarchical model as a probabilistic generative model
SUMMARY

▶ What is Bayesian statistics

▶ The machinery of Bayesian inference

▶ The foundations of Bayesianism

Next time

▶ Apply Bayesian principle to build statistical models for data analysis problems