Probability and Inference

February 19, 2023

Overview

QQ Group for resources and discussions

- I will upload course slides, assignments, and other related resources or reading materials to the group.
- You can discuss any Bayesian statistics related questions with other students or me here.

The three steps of Bayesian data analysis

- Setting up a full probability model—a joint probability distribution for all observable and unobservable quantities in a problem.
- Conditioning on observed data: calculating and interpreting the appropriate posterior distribution—the conditional probability distribution of the unobserved quantities of ultimate interest, given the observed data.
- Evaluating the fit of the model and the implications of the resulting posterior distribution.

Bayesian vs. Frequentist

- A primary motivation for Bayesian thinking is that it facilitates a common-sense interpretation of statistical conclusions.
- A Bayesian (probability) interval for an unknown quantity of interest can be directly regarded as having a high probability of containing the unknown quantity.
- A frequentist (confidence) interval, which may strictly be interpreted only in relation to a sequence of similar inferences that might be made in repeated practice.

Notation

- Inference is from sample to the population. To draw conclusion with unknown quantities based on the observed (known) quantities.
- Estimands:
  - Potentially observable quantities, such as future observations of a process, or the outcome under the treatment not received in the clinical trial example ($\tilde{y}$)
  - Quantities that are not directly observable, that is, parameters that govern the hypothetical process leading to the observed data (θ)
- Observed data: y

Notation

In general, we use

- Greek letters for parameters,
- lower case Roman letters for observed or observable scalars and vectors (and sometimes matrices),
- and upper case Roman letters for observed or observable matrices.

When using matrix notation, we consider vectors as column vectors throughout; for example, if u is a vector with n components, then $u^T u$ is a scalar and uu^T an $n \times n$ matrix.

General Notation for Statistical Inference

- Observational units and variables: $y = (y_1 \ldots, y_n)$
- Exchangeability: The usual starting point of a statistical analysis is the assumption that the n values y_i may be regarded as exchangeable, meaning that we express uncertainty as a joint probability density $p(y_1, \ldots, y_n)$ that is invariant to permutations of the indexes.
- We commonly model data from an exchangeable distribution as independently and identically distributed (i.i.d.) given some unknown parameter vector θ with distribution $p(\theta)$.

General Notation for Statistical Inference

- Explanatory variables: X with $n \times k$ dimension.
- Hierarchical models: models with different levels indicated by some specific variables.

Probability Notation

- Conditional probability: $p(\cdot|\cdot)$
- Marginal probability: $p(\cdot)$
- Probability of an event: $Pr(\theta > 2) = \int_{\theta>2} p(\theta)d\theta$
- When using a standard distribution, we use a notation based on the name of the distribution. For example, if θ has a normal distribution with mean μ variance σ^2, we write $\theta \sim N(\mu, \sigma^2)$ or $p(\theta) = N(\theta|\mu, \sigma^2)$

Bayes' Rule

- In order to make probability statements about θ and y. The joint probability mass or density function can be written as a product of two densities that are often referred to as the prior distribution $p(\theta)$ and the sampling distribution (or data distribution) $p(y|\theta)$, respectively:

$$p(\theta, y) = p(\theta)p(y|\theta)$$

- According to Bayes' rule, the conditional probability of θ given y, known as posterior distribution density:

$$p(\theta \mid y) = \frac{p(\theta, y)}{p(y)} = \frac{p(\theta)p(y \mid \theta)}{p(y)}$$

Bayes' Rule

- For discrete parameters: $p(y) = \sum_\theta p(\theta)p(y|\theta)$
- For continuous parameters: $p(y) = \int_\theta p(\theta)p(y|\theta)d\theta$
- Sometimes $p(y)$ is very hard to be directly analytically estimated, so we may use unnormalized posterior density:

$$p(\theta|y) \propto p(\theta)p(y|\theta)$$

The term $p(y|\theta)$ is taken as a function of θ with fixed y named as likelihood function.

Bayes' Rule

- The technical core of Bayesian inference: the primary task of any specific application is to develop the model $p(\theta, y)$ and perform the computations to summarize $p(\theta|y)$ in appropriate ways.

Prediction

- Unknown observable y has the marginal distribution (prior predictive distribution)

$$p(y) = \int p(y, \theta) d\theta = \int p(\theta) p(y|\theta) d\theta.$$

- After observe y, we can predict unobserved $\tilde{y}$ with the posterior predictive distribution:

$$\begin{aligned} p(\tilde{y}|y) &= \int p(\tilde{y}, \theta|y) d\theta \\ &= \int p(\tilde{y}|\theta, y) p(\theta|y) d\theta \\ &= \int p(\tilde{y}|\theta) p(\theta|y) d\theta. \end{aligned}$$

Evaluation

- The ratio of the posterior density $p(\theta|y)$ at the points θ_1 and θ_2 under a given model is called the posterior odds for θ_1 and θ_2.
- the posterior odds are equal to the prior odds multiplied by the likelihood ratio:

$$\frac{p(\theta_1|y)}{p(\theta_2|y)} = \frac{p(\theta_1)p(y|\theta_1)/p(y)}{p(\theta_2)p(y|\theta_2)/p(y)} = \frac{p(\theta_1)}{p(\theta_2)}\frac{p(y|\theta_1)}{p(y|\theta_2)}.$$

- The most familiar application of this concept is with discrete parameters, with θ_2 taken to be the complement of θ_1.

Example: Spelling Correction

- Suppose someone types 'radom'. How should that be read? It could be a misspelling or mistyping of 'random' or 'radon' or some other alternative, or it could be the intentional typing of 'radom'. What is the probability that 'radom' actually means random?

Example: Spelling Correction

- If we label y as the data and θ as the word that the person was intending to type, then

$$\Pr(\theta \mid y = \text{`radom'}) \propto p(\theta) \Pr(y = \text{`radom'} \mid \theta).$$

- For simplicity, consider only 3 possibilities for the intended word, θ (random, radon, or radom), we can compute the posterior probability of interest by computing

$$p(\text{random}|\text{`radom'}) = \frac{p(\theta_1)p(\text{`radom'}|\theta_1)}{\sum_{j=1}^{3} p(\theta_j)p(\text{`radom'}|\theta_j)},$$

where $\theta_1 = \textit{random}$, $\theta_2 = \textit{radon}$, and $\theta_3 = \textit{radom}$.

Example: Spelling Correction

- The prior probabilities $p(\theta_j)$ can most simply come from frequencies of these words in some large database, for example, Google:

θ	$p(\theta)$
random	7.60×10^{-5}
radon	6.05×10^{-6}
radom	3.12×10^{-7}

  The renormalization is not necessary.

- The likelihoods $p(y|\theta_j)$ can come from some modeling of spelling and typing errors, for example again, the Google's model of spelling:

θ	$p(\text{`radom'}|\theta)$
random	0.00193
radon	0.000143
radom	0.975

Example: Spelling Correction

- The word 'radom' in Wikipedia is a medium-sized city: home to 'the largest and best-attended air show in Poland, also the popular unofficial name for a semiautomatic 9 mm Para pistol of Polish design'.
- The likelihood function is not a probability distribution, but a set of conditional probabilities of a particular outcome ('radom') from three different probability distributions.
- The first conditional probability could be explained as 0.2% chance that this particular word is a mistyping from 'random'.

Example: Spelling Correction

- we multiply the prior probability and the likelihood to get joint probabilities and then renormalize to get posterior probabilities:

θ	$p(\theta)p(\text{`radom'}\|\theta)$	$p(\theta\|\text{`radom'})$
random	1.47×10^{-7}	0.325
radon	8.65×10^{-10}	0.002
radom	3.04×10^{-7}	0.673

- Question1: According to the Bayes' rule, what is the naive method of the renormalization?
- Question2: What if you find another model of spelling and typing errors who provides a different likelihood? Can you integrate the likelihoods together?

Example: Spelling Correction

Decision making, model checking, and model improvement.

- We can accept the two-thirds probability that the word was typed correctly.
- We can also question the probability by saying that 'radom' looks like a typo and that the estimated probability of it being correct seems much too high.
- When we dispute the claims of a posterior distribution, we are saying that the model does not fit the data or that we have additional prior information not included in the model so far.
- In statistics context, 'random' should be highly frequent, 'radon' occurs occasionally, and 'radom' is entirely new to us.

Example: Spelling Correction

- If we label x as the contextual information used by the model, the Bayesian calculation then becomes,

$$p(\theta|x, y) \propto p(\theta|x)p(y|\theta, x).$$

- This is not a perfect assumption but could reduce the burden of modeling and computation.
- The practical challenges in Bayesian inference involve setting up models to estimate all these probabilities from data. At that point, as shown above, Bayes' rule can be easily applied to determine the implications of the model for the problem at hand.

Probability as a measure of uncertainty

- In Bayesian statistics, probability is used as the fundamental measure of uncertainty.
- In Bayesian statistics, it would become as natural to consider the probability that an unknown estimand lies in a particular range of values as it is to consider the probability that the mean of a random sample of 10 items from a known fixed population of size 100 will lie in a certain range.
- The guiding principle is that the state of knowledge about anything unknown is described by a probability distribution.
- Probabilities may be a reasonable approach to summarizing uncertainty in applied statistics, but the ultimate proof is in the success of the applications.

Some useful results from probability theory

- The relationship among joint, marginal, and conditional distribution function.
- To factor a joint density.
- Expectation and variance
- covariance matrix of a vector
- $E(u) = E(E(u|v))$
- $var(u) = E(var(u|v)) + var(E(u|v))$
- Suppose $p_u(u)$ is the continuous density of the vector u and we transform to $v = f(u)$, then $p_v(v) = |J|p_u(f^{-1}(v))$
- Logistic transformation and its inverse transformation.