STA261: Probability and Statistics II

Shahriar Shams

Week 10 (Bayesian Inference)

UNIVERSITY OF TORONTO

Winter 2020

Recap of Week 9

- Likelihood ratio test (LRT)
  - LRT for single population
  - LRT for two populations
  - Confidence Interval using LRT
- Goodness of Fit (GOF) test

Learning goal for this week

- Part-1 of lecture:
  - Idea of Bayesian Inference
  - Prior, Likelihood and Posterior
  - Some examples of calculating posterior dist
- Part-2 of lecture:
  - Inference using posterior dist
    - Summary of posterior dist
    - Credible Region
  - Different types of prior

These are selected topics from E&R chapter 7.1, 7.2, 7.4

Section 1

Idea of Bayesian Inference

Frequentist approach: the concept of FIXED parameter

- Previously, all the inferences that we made had one common assumption:
  - Parameter, θ is a fixed number (though unknown)
- Hence, we can not write statements like $P[3 \leq \theta \leq 5] = 0.95$
- But we defined likelihood function, $L(\theta)$ and differentiated with respect to θ (!)
- This is known as the *frequentist* approach.
- The interpretation of the confidence intervals calculated using the frequentist approach are often criticized for not having real intuitive meaning.

Bayesian: parameter is a random variable

- "Whether a parameter, θ is a fixed unknown number or a random variable", this debate is rather philosophical.
- We will not get into that debate.
- I will simply introduce the idea of a parameter being a random variable by giving few real life examples.
- One huge positive side of using Bayesian inference is that we can recover all the inferences made in frequentist approach as a special case.

One example

- We are interested in calculating the average height of all UofT students. We say height of a single student follows a Normal distribution with mean μ and variance σ^2.
  - When estimating μ, we use maximum likelihood estimation, we define $L(\mu)$
  - We treat μ as a completely unknown quantity.

  - Is μ completely unknown? Do we know nothing about μ?
  - I believe we all will agree that μ is a number between 140cm and 200cm

  - Question: how do we incorporate this prior belief into our calculation.

One more example (a sad but current one)

- Suppose we want to estimate the true proportion of COVID-19 deaths in Canada. We say, whether a Canadian with COVID-19 will die or not follows $Bernoulli(\theta)$.
  - We can estimate θ, by taking a representative sample of size n from the confirmed cases and by counting the number of deaths(say X)
  - X/n is an estimator θ.

  - Is θ completely unknown?
  - Can we use the estimated value from China or Italy and incorporate that into our calculation?

  - Question: how do we incorporate this prior belief into our calculation.

Incorporating Prior belief

- In our first example , intuitively we can say, $\mu \sim Unif(140, 200)$
  - here, all we are saying μ could be any number between 140 and 200. We are not supporting any part of our belief.

- In our second example, intuitively we can say, $\theta \sim Unif(0, 0.2)$

Section 2

Prior, Likelihood and Posterior

Prior and Posterior distribution

- In **Bayesian** setting, parameters are believed to be random variables following some distributions.
- Distribution of the parameter is called *prior*, $\pi(\theta)$.
- Generally speaking, $\pi(\theta)$ is a valid *pdf* of θ
- Our interest is in updating this prior belief using the observed data $(X_1, X_2, ..., X_n)$.
- If $(X_1, X_2, ..., X_n)$ is the observed data, then our goal is to calculate $\pi(\theta | X_1, X_2, ..., X_n)$
    - This is the conditional distribution of θ given the observations $(X_1, X_2, ..., X_n)$
    - In E&R, $(X_1, X_2, ..., X_n)$ are represented as s.
    - So in short we are interested in $\pi(\theta | s)$
- $\pi(\theta | \boldsymbol{s})$ is called the *posterior* distribution of θ

Calculating the Posterior

- Under the Bayesian setting, likelihood is the conditional distribution for the data $\boldsymbol{s}$ given θ.
- Recall: $L(\boldsymbol{s}|\theta) = f(x_1, x_2, ..., x_n|\theta)$
- If we multiply this by the prior, $\pi(\theta)$, we will get the joint distribution of s and θ
- Marginal distribution of the data $\boldsymbol{s}$ is given by

$$m(\boldsymbol{s}) = \int_{\Omega} L(\boldsymbol{s}|\theta) * \pi(\theta)\, d\theta$$

- Posterior distribution is given by,

$$\pi(\theta|\boldsymbol{s}) = \frac{L(\boldsymbol{s}|\theta) * \pi(\theta)}{m(\boldsymbol{s})}$$

Posterior distribution (cont...)

- $m(\boldsymbol{s})$ is free of θ (Since we have integrated θ out)
- $m(\boldsymbol{s})$ plays the role of the *inverse normalizing constant* for the posterior distribution.
- In other words, $L(\boldsymbol{s}|\theta) * \pi(\theta)$ is not always a valid or [sum/integration $\neq 1$]
- $m(\boldsymbol{s})$ makes sure that

$$\int_{\Omega} \pi(\theta|\boldsymbol{s}) = 1$$

- Since $m(\boldsymbol{s})$ is constant, we can write

$$\pi(\theta|\boldsymbol{s}) \propto L(\boldsymbol{s}|\theta) * \pi(\theta)$$

- From the expression of $L(\boldsymbol{s}|\theta) * \pi(\theta)$ we try to deduce the probability distribution.

Section 3

Some examples of Calculating Posterior Distribution

$(X_1, X_2, ..., X_n) \sim Bern(\theta)$ and $\theta \sim Unif[0, 1]$

$$L(\theta) = \theta^{\sum x_i} (1-\theta)^{n-\sum x_i}$$

Posterior $= 1 I(0 \leq \theta \leq 1)$

$$L(\theta | S) = k\, \theta^{\sum x_i + 1 - 1} (1-\theta)^{n - \sum x_i + 1 - 1}$$

$$\sim Beta\left(\sum x_i + 1,\ n - \sum x_i + 1\right)$$

$$L(\theta) = \theta^{\sum x_i} (1-\theta)^{n-\sum x_i}$$

Posterior: $\pi(\theta)\ \theta^{\alpha-1} (1-\theta)^{\beta-1}$

$L(\theta) \times$ Posterior: $k\ \theta^{\sum x_i + \alpha - 1} (1-\theta)^{n - \sum x_i + \beta - 1}$

$$\sim Beta\left(\sum x_i + \alpha,\ n - \sum x_i + \beta\right)$$

$(X_1, X_2, ..., X_n) \sim N(\mu, \sigma_0^2)$ where σ_0^2 is known and $\mu \sim N(\mu_0, \tau_0^2)$

Assignment (Non-credit) for part-1

Evans and Rosenthal

Exercise: 7.1.1, 7.1.4, 7.1.5, 7.1.7, 7.1.9

Section 4

Inference using posterior distribution

Estimation using Posterior Distribution

- Though *posterior* distribution sounds fancy, in reality this is just a pdf (or pmf) of a random variable θ.
- We can calculate mean, variance, mode, quantiles of this distribution etc. just the way we calculated these for any distribution in STA257.
- The corresponding summaries will then be called by the same name just with the word posterior added to it.
- For example, the mean of the posterior distribution is called "posterior mean". Similarly the other summaries.

Examples of calculating posterior mean

- The two posterior distributions calculated on slide 15 and 16 were both Beta distributions.
- We know for any $Beta(\alpha, \beta)$ distribution, mean=$\frac{\alpha}{\alpha+\beta}$

- For the example, $(X_1, X_2, ..., X_n) \sim Bern(\theta)$ and $\theta \sim Unif[0, 1]$
  - Posterior dist, $\theta | s \sim Beta(\sum x_i + 1, n - \sum x_i + 1)$
  - Posterior mean $= E[\theta|s] = \frac{\sum x_i + 1}{n+2}$

- For the example, $(X_1, X_2, ..., X_n) \sim Bern(\theta)$ and $\theta \sim Unif[0, 1]$
  - Posterior dist, $\theta | s \sim Beta(\sum x_i + \alpha, n - \sum x_i + \beta)$
  - Posterior mean $= E[\theta|s] = \frac{\sum x_i + \alpha}{n+\alpha+\beta}$

- Similarly for the Normal example.

Note: $Unif[0, 1] \equiv Beta(1, 1)$

Other summaries

- We can calculate the posterior variance in the same way we calculated posterior mean in the previous slide.
- We can calculate the median just by calculating the 50th percentile of the posterior distribution which is by solving this equation

$$\int_{-\infty}^{m} \pi(\theta|s)\, d\theta = \frac{1}{2}$$

- We can calculate mode by finding the value of θ at which the posterior density is the maximum.

Credible Interval/Region

- Credible Interval (or sometime called region) is the Bayesian equivalent idea of confidence interval.
- Recall: In confidence interval, we wanted to construct a statement like
$$P[l() < \theta < u()] = \gamma$$
- From posterior distribution, since its a distribution of θ, we can simply find two quantiles of the distribution that covers 100*γ% of the distribution.

HPD Interval

- If we remember from week 6, we can construct infinitely many γ level confidence intervals.
- In the same way for credible intervals we also have a lot of different ways of calculating it.
- Intuitively it makes much more sense to include those θ values that corresponds to the higher posterior densities.
- Hence the name, highest posterior density interval or HPD interval.
- HPD interval is the narrowest of all possible intervals just because of the way it is constructed.

Example using Location Normal model

- The example 7.1.2 on page 377 of E& R involves calculating posterior distribution of μ given σ^2 known.
- $\mu|s \sim N\left((\frac{1}{\tau_0^2}+\frac{n}{\sigma_0^2})^{-1})(\frac{1}{\tau_0^2}\mu_0+\frac{n}{\sigma_0^2}\bar{x})\,,\,(\frac{1}{\tau_0^2}+\frac{n}{\sigma_0^2})^{-1})\right)$
- Posterior mean $= (\frac{1}{\tau_0^2}+\frac{n}{\sigma_0^2})^{-1})(\frac{1}{\tau_0^2}\mu_0+\frac{n}{\sigma_0^2}\bar{x})$
- Posterior variance $= (\frac{1}{\tau_0^2}+\frac{n}{\sigma_0^2})^{-1}$
- γ-level credible interval

$$(\frac{1}{\tau_0^2}+\frac{n}{\sigma_0^2})^{-1})(\frac{1}{\tau_0^2}\mu_0+\frac{n}{\sigma_0^2}\bar{x}) \pm z_{\frac{1-\gamma}{2}}\sqrt{(\frac{1}{\tau_0^2}+\frac{n}{\sigma_0^2})^{-1}}$$

Section 5

Different types of prior

Conjugate Prior

- Conjugate prior corresponds to a prior that will result in a posterior distribution that belongs to the same family of distribution as the prior.
- For example, In the data from $Bernoulli(\theta)$ with $Beta(\alpha, \beta)$ prior example,
  - We got a posterior distribution that is also $Beta$ just the parameters are different.
  - Since Prior and Posterior both follows $Beta$, this prior for this particular example is called a conjugate prior.
- Some other known examples of conjugate prior:
  - data follows $N(\mu, \sigma^2)$ + Prior, $\mu \sim N(\mu_0, \tau_0^2)$
  - data follows $Poisson(\lambda)$ + Prior, $\lambda \sim Gamma(\alpha, \beta)$

Improper priors

- As we have mentioned previously, a prior distribution is generally a valid pdf of θ.
- Sometimes a function is used as a prior that is not a valid pdf. ie. $\int_\Omega \pi(\theta) \neq 1$
- These types of priors are called improper priors.
- For example $Beta(0,0)$ is sometime used as prior which is not a valid pdf.

Non-informative prior

- Often we don't know anything about θ.
- We then use priors that are non-informative or vague.
  - In the $Bernoulli(\theta)$ example, saying $\theta \sim Unif[0,1]$ is non-informative.
  - In the location normal example, saying $\tau_0^2 \to \infty$ adds no information to the analysis.
- The idea of non-informative prior is that we don't want to add any prior information rather want the posterior to be completely based on the data (same as saying likelihood).

Retrieving MLE from Bayesian analysis

Some hand waving:

- We know,

$$posterior \propto likelihood * prior$$

- under non-informative prior,

$$posterior \propto likelihood$$

- Then

$$posterior\ mode \equiv maximum\ likelihood\ estimate$$

Re-visit Location Normal example

- $\mu|s \sim N\left((\frac{1}{\tau_0^2} + \frac{n}{\sigma_0^2})^{-1}(\frac{1}{\tau_0^2}\mu_0 + \frac{n}{\sigma_0^2}\bar{x}) \, , \, (\frac{1}{\tau_0^2} + \frac{n}{\sigma_0^2})^{-1})\right)$
- Under non-informative prior
- Posterior mean $= (\frac{1}{\tau_0^2} + \frac{n}{\sigma_0^2})^{-1}(\frac{1}{\tau_0^2}\mu_0 + \frac{n}{\sigma_0^2}\bar{x}) \;\; \rightarrow \bar{x}$
- Posterior variance $= (\frac{1}{\tau_0^2} + \frac{n}{\sigma_0^2})^{-1} \;\; \rightarrow \frac{\sigma_0^2}{n}$
- γ-level credible interval

$$(\frac{1}{\tau_0^2} + \frac{n}{\sigma_0^2})^{-1}(\frac{1}{\tau_0^2}\mu_0 + \frac{n}{\sigma_0^2}\bar{x}) \pm z_{\frac{1-\gamma}{2}} \sqrt{(\frac{1}{\tau_0^2} + \frac{n}{\sigma_0^2})^{-1}}$$

$$\rightarrow \quad \bar{x} \pm z_{\frac{1-\gamma}{2}} \frac{\sigma_0}{\sqrt{n}}$$

Assignment (Non-credit) for part-2

Evans and Rosenthal

Exercise: 7.1.2, 7.1.3, 7.2.1, 7.2.10, 7.2.12(a), 7.2.20, 7.4.1