STA 250 Lecture 4 Notes

Taken by Teresa Filshtein

Oct 09, 2013

1 Intro to Bayesian Inference

- In Bayesian models, parameters are not fixed but treated as random variables that follow a probability distribution
- ***Definition:*** **Prior Distribution** ($\pi(\theta)$) - the distribution of the parameter (θ) assumed, *before* observing any data.
- A goal of Bayesian inference is to make inference about θ (the parameter).
- The 100(1-α)% confidence interval from a classical stand point is one defined for a fixed parameter under repeated sampling. In other words, we could make statements like "the 100(1-α)% confidence interval is expected to contain our true parameter θ, 100(1-α)% of the time under repeated sampling from the model". For any given interval, however, the true parameter is either in the interval or it isn't, since the parameter is fixed. In the classical case, therefore, under repeated sampling, the parameter is fixed, and the CI is random.
- Bayesian Inference is based on the idea that the parameter itself (or our state of knowledge about it) is random (or described by a probability distribution) and allows us to make probability statements about θ. For example, we can make statements like: 'there is a (1-α)% chance that $theta$ is between 0.12 and 0.85'.
- The idea is that with the likelihood function $p(\mathbf{y}|\theta)$, the probability of our data $\mathbf{y}$ given the parameter θ, we would like to find $\pi(\theta|\mathbf{y})$, the probability of the parameter θ given our data $\mathbf{y}$
- ***Definition:*** **Posterior Distribution** ($\pi(\theta|\mathbf{y})$): the conditional distribution of θ given the observed data $\mathbf{y}$

$$\pi(\theta|\mathbf{y}) = \frac{p(\mathbf{y}|\theta)\pi(\theta)}{\int p(\mathbf{y}|\theta)\pi(\theta)d\theta} \tag{1}$$

2 Prior Distribution

- A main topic in Bayesian inference is the determination of the prior distribution. How do you know what the prior distribution of θ should be?

- There are three main methods for determining which distributions are appropriate for θ.

  1. Reference Priors (Jose Bernado, Jim Berger ̃1970s)
     - Idea: Maximize the "'distance"' (e.g. the K-L divergence) between the prior distribution and the posterior distribution
     - This method puts the most "'impact"' to the data. Because the prior is 'far away' it is not informative, and therefore most of the information is contained in the data.
     - Excellent rule but tricky to derive for complex models
  2. Probability Matching Prior (Welch/Peers 1956)
     - Idea: select a prior distribution such that the posterior distribution allows the construction of intervals with frequentist coverage (i.e. confidence intervals).
     - Nice in theory but is not practical and can be hard to derive.
  3. Invariance
     - Idea: construct a rule such that the prior distributions constructed in different parametizations are consistent.
     - The most famous invariant prior is Jeffreys Prior,

       $$\pi(\theta) \propto |I(\theta)|^{\frac{1}{2}}, \tag{2}$$

     - where $I(\theta)$ is the Fisher Information and this is the square root of the determinant

3 Jeffreys Prior

- The idea behind the invariance property of Jeffreys Prior is that (for example) take $\theta = e^{\mu}$. You could derive the JP for μ and then transform it accordingly and you would obtain the results as if you were to find the JP on θ directly.

- **Example**

  $$y_i|\theta \sim^{ind} Bin(n_i, \theta), (i = 1, ...m)$$

  - Therefore the likelihood function for all m observations is:

  $$p(\mathbf{y}|\theta) = \prod_{i=1}^{m} \left(\begin{array}{c} n_i \\ y_i \end{array} \right) \theta^{y_i}(1-\theta)^{n_i - y_i}$$

  $$\propto \theta^{\sum y_i}(1-\theta)^{\sum n_i - y_i}$$

- And the Posterior Distribution is

  $$\pi(\theta|\mathbf{y}) \propto \pi(\theta)p(\mathbf{y}|\theta) = \pi(\theta)\theta^{\sum y_i}(1-\theta)^{\sum n_i - y_i} \tag{3}$$

  If you do the calculations, you find that

  $$I(\theta) = \frac{\sum n_i}{\theta(1-\theta)}$$

  and Jefferey's Prior is

  $$\pi(\theta) \propto \theta^{-\frac{1}{2}}(1-\theta)^{-\frac{1}{2}}$$

  Which changes 3 to

  $$\pi(\theta|\mathbf{y}) \propto \theta^{\sum y_i - \frac{1}{2}}(1-\theta)^{\sum (n_i - y_i) - \frac{1}{2}} \tag{4}$$

- We recognize this as a Beta(α, β) distribution with Beta parameters $\alpha = \sum y_i + \frac{1}{2}$ and $\beta = \sum (n_i - y_i) + \frac{1}{2}$ and we use the property that for any probability density $f(x)$, $\int f(x) = 1$, to find the proportionality constant. We need

  $$\int \pi(\theta|\mathbf{y}) = 1,$$

  Recall that if $X \sim$ Beta(α, β), for $x \in (0, 1)$

  $$p(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$$

  Therefore 4 can be expressed as

  $$\pi(\theta|\mathbf{y}) \propto \theta^{(\sum y_i + \frac{1}{2}) - 1}(1-\theta)^{(\sum (n_i - y_i) + \frac{1}{2}) - 1} \tag{5}$$

  And we can obtain our proportionality constant

  $$\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}$$

- Note that Jefferey's Prior, $\pi(\theta) \sim$ Beta$(\frac{1}{2}, \frac{1}{2})$
- When the prior and posterior distribution are in the same family, we call the prior a ***Conjugate Prior***
- Some examples of Conjugate/Likelihoods are Normal/Normal, Binomial/Beta.
- Lets continue. We now have our posterior distribution

  $$\theta|\mathbf{y} \sim Beta\left(\frac{1}{2} + \sum y_i, \frac{1}{2} + \sum (n_i - y_i)\right)$$

  Suppose we get data with # of successes and # of failures, respectively, $\sum y_i = 10, \sum (n_i - y_i) = 20$, our posterior distribution (now that we have observed this data) is

  $$\theta|\mathbf{y} \sim Beta(10.5, 20.5)$$

- Now we need a point estimate
  - posterior mean
  - posterior median
  - posterior mode (usually this is great to use but can be hard to compute in practice)

  We also need an uncertainty quantification/interval or "'Credible Interval"' (in Bayesian Context a Posterior Interval is called a Credible Interval).

  $S^{1-\alpha}(y)$ is defined to be the $100(1-\alpha)\%$ Credible Interval. If

$$\int_{S^{1-\alpha}(y)} \pi(\theta|\mathbf{y}) = (1-\alpha) \tag{6}$$

4 Credible Intervals

- There are two main types of Intervals
  1. Central Interval
     - this is defined by the $\frac{\alpha}{2}, 1-\frac{\alpha}{2}$ percentiles of the posterior
  2. HPD, Highest Posterior Density Interval
     - This is an interval S such that:

$$S = \{\theta : \pi(\theta) > \pi(\theta'), \int_S \pi(\theta|\mathbf{y}) = (1-\alpha)\} \tag{7}$$

$\forall \;\; \theta \in S$ and $\;\; \theta' \notin S$