StatsM254 Statistical Methods in Computational Biology Lecture 3 - April 6, 2015

Subsampling vs. Bootstrap & Priors

Lecturer: Jingyi Jessica Li

Scribe: Kimberly Insigne

1 Subsampling vs. bootstrap

Data: $X_1, ..., X_n$

Size of subsample: b

 $X_1^*, ..., X_b^* \in (X_1, ..., X_n)$ without replacement At most $\binom{n}{b}$ subsamples and for each calculate a test statistic, giving $\binom{n}{b}$ subsample statistics.

Two main differences between subsample and bootstrap

- 1. Size of sample: b vs. n
- 2. without replacement vs. with replacement

Bootstrap: Treat X_1, \ldots, X_n as a new population, draw n data points from it with replacement. True population: P - probability/distribution "New" population (sample): \hat{P}_n - empirical probability/distribution

Subsampling: draw smaller samples of size b from PBootstrap: draw samples of size n from \hat{P}_n (\hat{P}_n needs to be close to P)

With subsampling, it is a difficult issue to choose b

To get desirable theory:

subsampling: $\frac{b}{n} \xrightarrow{n \to \infty} 0$

Don't want b too small, but if too large the condition won't hold

Bootstrap - sometimes it fails when

 $\hat{P}_n \to P \not\Rightarrow T(\hat{P}_n) \to T(P)$

Example:

$$X_1, ..., X_n \sim U(0, \theta)$$
$$T(X_1, ..., X_n) = X_{(n)}$$

X follows a uniform distribution with minimum value 0 and maximum value θ . The test statistic uses the maximum value $X_{(n)}$ to estimate the data. When you do bootstrap, the distribution of the bootstrap test statistic will not converge to distribution of $X_{(n)}$.

Which one to choose?

- 1. Choose subsampling if you have an appropriate way to choose b.
- 2. Otherwise, choose bootstrap for its simplicity, less worrisome.

2 Choice of prior

Uninformative/non-informative/objective priors aim to provide no subjective information about the distribution of parameters. However, the term uninformative is somewhat of a misnomer.

2.1 Flat prior

A flat prior gives uniform density.

For the distribution $X \sim N(\mu, 1)$ the prior is $p(\mu) \propto 1$ for every possible μ , giving each the same density e.g. $\mu \sim U(0, 1)$.

Another example:

$$\begin{split} &X \sim N(1,\sigma^2) \\ &\sigma^2 \sim U(0,1) \text{ - there is no preference of } \sigma^2 \\ \text{Reparameterization: } &\tau = \frac{1}{\sigma} \text{ and the distribution becomes } X \sim N(1,\frac{1}{\tau^2}) \\ &p(\sigma^2) = 1 \\ &p(\tau^2) = \frac{1}{\sigma^2} \end{split}$$

This reparameterization shows that a flat prior can be flat for one parameterization but not for another.

2.2 Jeffreys prior

Jeffreys prior is invariant to reparameterization (Jeffreys, 1961). Jeffreys prior works well for a single parameter, but multi-parameter situations may have inappropriate aspects accumulate across dimensions to detrimental effect.

 $X \sim N(1, \sigma^2)$ $p(\sigma^2) \propto \sqrt{I(\sigma^2)}$

Fisher's information is how much information the likelihood tells you about the parameter and is defined as: $\int \left[\left(dL(\sigma^2 | X) \right)^2 \right]$

$$I(\sigma^2) = E \left[\left(\frac{dL(\sigma^2|X)}{d\sigma^2} \right)^2 \right]$$

Let's reparameterize to show that Jeffrey's prior is invariant to reparameterization:

$$\tau^2 = \frac{1}{\sigma^2} \Rightarrow \sigma^2 = \frac{1}{\tau^2} = f(\tau^2)$$

$$p(\tau^2) = p(\sigma^2) \cdot \left| \frac{d\sigma^2}{d\tau^2} \right|$$

= $\propto \sqrt{I(\sigma^2) \left(\frac{d\sigma^2}{d\tau^2}\right)^2}$
= $\sqrt{E\left[\left(\frac{dL(\sigma^2|X)}{d\sigma^2}\right)^2 \right] \cdot \left(\frac{d\sigma^2}{d\tau^2}\right)}$ where $\frac{d\sigma^2}{d\tau^2}$ is a constant

$$= \sqrt{E\left[\left(\frac{dL(\sigma^2|X)}{d\sigma^2} \cdot \frac{d\sigma^2}{d\tau^2}\right)^2\right]} = E\left[\left(\frac{dL(\tau^2|X)}{d\tau^2}\right)^2\right] = \sqrt{I(\tau^2)} \text{ same form!}$$

This demonstrates that Jeffrey's prior is invariant to the parameter.

2.3 Principle of maximum entropy

The principle states that, subject to precisely stated prior data, the probability distribution that best represents the current state of knowledge is the one with the largest entropy (adding noise to the likelihood). This is not often seen, and this is what typically is meant by uninformative priors.

2.4 Conjugate prior

A conjugate prior is not objective, but convenient. The aim is to have the posterior in the same distribution family as the prior distribution.

Example:

 $X \sim binomial(n, p)$ where n is fixed, estimate p

conjugate prior of $p: p \sim Beta(a, b)$ with support $p \in (0, 1)$

likelihood: $p(x|p) = \binom{n}{x} p^x (1-p)^{n-x}$ describes the distribution of x given p

prior: $p(p) \propto p^a (1-p)^b$ describes the distribution of p

posterior: $p(p|x) \propto p(p) \cdot p(x|p) \propto p^{x+a}(1-p)^{n-x+b}$ is the posterior distribution after combining terms and has the form $p|x \sim Beta(x+a, n-x+b)$

This shows that the beta distribution is the conjugate prior of binomial. Conjugate priors are often used Bayesian statistical modeling in practice.