CS207 Estimation

By: Harsh Shah

September 2021

1 Basics

Some definitions:

- Sample: Collection of i.i.d. r.v. from same distribution is called a sample of size N. One set of observed data, is called one instance of the sample.
- Statistic: A function of the sample (say $\mathrm{T}(X_1, X_2 \ldots, X_i)$) is called a statistic. It is also a r.v. and the value it takes for some observed sample is called instance of statistic.
- Statistical Model: A probabilistic description (given by a distribution based on certain parameters) of any real world phenomena.
- Estimation theory: (Inverse of data generation process) Theory that deals with estimation of parameters (in order to estimate a distribution) based on some empirical data. It is often assumed that the data comes from some family of parametric distributions and the parameters required to define that distribution are estimated.
- Estimator: It is a statistic that outputs a deterministic (non-stochastic) value based on an instance of sample. The output is called estimate.

2 Mean, Variance and Bias of estimator

Let T_N=T$(X_1, X_2 \ldots X_n)$ be a statistic.
Some definitions:

- Mean (expectation value) of an estimator: E[T]
- Variance: var(T)=$E[(T - E[T])^2]$
- Bias(T)=E[T]-θ
- Mean Squared Error=MSE(T)=$E[(T - \theta)^2]$
- Unbiased estimator: Bias(T)=0
- Consistent Estimator: $P(|T - \theta| > \epsilon) = 0$ as $n \to \infty$

3 Bias-Variance decomposition of MSE

$$\begin{aligned} MSE(T) &= E[(T-\theta)^2] \\ &= E[T^2 + \theta^2 - 2T\theta] \\ &= E[T^2 - 2E[T]T + (E[T])^2 + \theta^2 - 2T\theta + (E[T])^2 + 2(E[T])^2 - 2E[T]T] \\ &= var(T) + (bias(T))^2 \end{aligned} \tag{1}$$

If two estimators T_1 and T_2 have same MSE then, the one with lower variance has greater bias and vice-versa. Variance can be considered as precision and bias as accuracy.

4 Likelihood function

Let $X_1, X_2 \dots X_n$ be a sample on r.v. X., with PDF/PMF=$P(X;\theta)$. Then
Likelihood function=L$(\theta; X_1, X_2 \dots X_n)$=$\Pi_{i=1}^{n} P(X_i;\theta)$

Important Property:
Let θ_{true} be the parameter value that generated r.v. X_i. If the following assumptions are satisfied:

1. Parameter θ identifies a unique distribution
2. All PMF/PDF have the same support(values of r.v. having non-zero probability) for all θ.
3. $E_{P(\theta_{true};X)}\left[\frac{P(\theta;X)}{P(\theta_{true};X)}\right]$ exists and is finite

Then,

$$lim_{n\to\infty} P(L(\theta_{true}; X_1, X_2 \dots X_n) > L(\theta; X_1, X_2 \dots X_n); \theta_{true}) = 1 \quad \forall \theta \neq \theta_{true}$$

Proof:

$$L(\theta_{true}; X_1, X_2 \dots X_n) > L(\theta; X_1, X_2 \dots X_n) \equiv \frac{1}{N}\sum_{i=i}^{n} log\left[\frac{P(\theta; X_i)}{P(\theta_{true}; X_i)}\right] < 0$$

As $n \to \infty$ we can use law of large numbers on the r.v.'s $log\left[\frac{P(\theta;X_i)}{P(\theta_{true};X_i)}\right]$ resulting in,

$$\frac{1}{N}\sum_{i=i}^{n} log\left[\frac{P(\theta; X_i)}{P(\theta_{true}; X_i)}\right] = E\left[log\left[\frac{P(\theta; X)}{P(\theta_{true}; X)}\right]\right]$$

Using Jensen's inequality,

$$E\left[log\left[\frac{P(\theta; X)}{P(\theta_{true}; X)}\right]\right] < log\left[E\left[\frac{P(\theta; X)}{P(\theta_{true}; X)}\right]\right] = log(1) = 0$$

Hence, we can find the θ value at which the likelihood function is maximum, to get an estimate of θ_{true}.

5 Maximum Likelihood estimator

Maximum likelihood estimator T:=$argmax_\theta L(\theta; X_1, X_2, \dots X_n))$
Properties:

- It is possible ML may not exists or may not be unique.
- If the assumptions of important property[4] are valid, then it can be proved that ML estimator is a consistent estimator, given it exists and is unique.

6 ML calculation for various distributions

6.1 Bernoulli

For Bernoulli distribution, the parameter defining the distribution is probability of success(=θ)

$$L(\theta; X_1, X_2 \dots, X) = \Pi_{i=1}^{n} \theta^{X_i}(1-\theta)^{1-X_i}$$

To get the ML, we need to equate the first derivative of L(.) to zero. Upon simplification, we get:

$$\theta = \frac{X_1 + X_2 \dots X_n}{n}$$

6.2 Binomial

Likelihood function for binomial distribution:

$$L(\theta; X_1, X_2 \dots, X_n) = \Pi_{i=1}^{n} ({}^{m}C_{X_i})\theta^{X_i}(1-\theta)^{n-X_i}$$

ML in this case turns out to be,

$$\theta = \frac{1}{nm}\sum_{i=1}^{n} X_i$$

6.3 Gaussian

ML estimation requires partial differentiation for minimizing the likelihood function. Upon carrying out the procedure(solving two equations to get μ and σ), we get that, μ is sample mean, and σ^2 is sample variance

6.4 Poisson

λ represents arrival rate
ML estimate is sample mean

6.5 Uniform distribution

Let the sorted sample be represented as $\{X_1, X_2 \dots X_n\}$.
The uniform distribution is uniquely determined by parameters a,b (where [a,b] represents the domain of uniform PDF). However, there are also constraints $a < X_1$ and $b > X_n$. This results in ML estimate as $a = X_1$ and $b = X_n$.