Sufficient Statistics

Bruce A. Desmarais

Stats I Lab

October 26, 2009

Classical Parametric Statistics

The three steps in classical parametric statistics

1. Assume that the data on hand (X) is a sample from a known parametric family of distributions with one or more unknown parameters to be estimated θ.
2. Identify functions (sufficient statistics) ($T(X)$) to be applied to the data that provide all of the information in the sample about θ.
3. Characterize the likelihood of observing $T(X)$ over all possible values of θ in order to describe the location and certainty of estimation.

The Sample PDF; A Special Case of a Joint Probability

The probability (density in the continuous case) of observing a vector-valued random variable ($X = [x_1, x_2..]$) can always be expressed as a scalar which is equal to the joint PDF of X.

A sample of n observations on a given variable is a vector-valued (multivariate) random variable. If the n observations are assumed to be i.i.d., each from distribution with PDF f, then, using the product rule, the joint PDF of X is

$$\prod_{i=1}^{n} f_X(x_i)$$

The Sufficient Statistic

If the form of the distribution of X is assumed, it is often possible to define a *dimensional reduction* of the data $T(X)$ that captures all of the information in X about the parameter θ.

It is very important to master the concept of sufficiency, since it is used in hypothesis testing and parameter estimation, the two main objects of classical statistics.

If $T(X)$ is sufficient for θ

$$P(X = x | T(X), \theta) = P(X = x | T(X))$$

Fisher-Neyman Factorization Theorem

Denote the PDF or PMF of x given θ $f_\theta(x)$ If $T(X)$ is sufficient for θ

$$f_\theta(x) = h(X)g_\theta(T(X))$$

Where h does not depend on θ and g depends on θ and X only through $T(X)$.

Example: The exponential distribution has PDF

$$f_\lambda(x) = \lambda e^{-\lambda x}$$

What is a sufficient statistic for λ if we have an i.i.d. sample of size n from an exponential distribution?

The Principal of Minimal Sufficiency

The principal of minimal sufficiency says that any sufficient statistic for θ is a one-to-one deterministic function of any other sufficient statistic for θ.

$$S(X) = m(T(X))$$

This is useful when it comes to parameter estimation because if a sufficient statistic ($T(X)$) is known, it is then known that the most efficient estimator of the parameter θ is a deterministic, one-to-one function of $T(X)$.

Example II: The Normal Distribution

The normal PDF is

$$f_{(\mu,\sigma^2)}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right)$$

Find a Sufficient Statistic for μ given that σ^2 is known and we have an i.i.d. sample of size n.

Example III: The Bernoulli Distribution

The PMF of the Bernoulli Distribution is

$$P_\pi(x) = \pi^x(1-\pi)^{1-x}$$

Assuming we have an i.i.d. sample of size n from a Bernoulli Distribution, define a sufficient statistic for π.