Sufficient Statistics

Guy Lebanon

May 2, 2006

A sufficient statistics with respect to θ is a statistic $T(X_1,\ldots,X_n)$ that contains all the information that is useful for the estimation of θ. It is useful a data reduction tool, and studying its properties leads to other useful results.

Definition 1. *A statistic T is sufficient for θ if $p(x_1,\ldots,x_n|T(x_1,\ldots,x_n))$ is not a function of θ.*

A useful way to visualize it is as a Markov chain $\theta \to T(X_1,\ldots,X_n) \to \{X_1,\ldots,X_n\}$ (although in classical statistics θ is not a random variable but a specific value). Conditioned on the middle part of the chain, the front and back are independent.

As mentioned above, the intuition behind the sufficient statistic concept is that it contains all the information necessary for estimating θ. Therefore if one is interested in estimating θ, it is perfectly fine to 'get rid' of the original data while keeping only the value of the sufficient statistic. The motivation connects to the formal definition by considering the concept of sampling a ghost sample: Consider a statistician who erased the original data, but kept the sufficient statistic. Since $p(x_1,\ldots,x_n|T(x_1,\ldots,x_n))$ is not a function of θ (which is unknown), we assume that it is a known distribution. The statistician can then sample $x'_1,\ldots,x'_n$ from that conditional distribution, and that ghost sample can be used in lieu of the original data that was thrown away.

The definition of sufficient statistic is very hard to verify. A much easier way to find sufficient statistics is through the factorization theorem.

Definition 2. *Let $X_1,\ldots,X_n$ be iid RVs whose distribution is the pdf f_{X_i} or the pmf p_{X_i}. The likelihood function is the product of the pdfs or pmfs*

$$L(x_1,\ldots,x_n|\theta) = \begin{cases} \prod_{i=1}^n f_{X_i}(x_i) & X_i \text{ is a continuous RV} \\ \prod_{i=1}^n p_{X_i}(x_i) & X_i \text{ is a discrete RV} \end{cases}.$$

The likelihood function is sometimes viewed as a function of $x_1,\ldots,x_n$ (fixing θ) and sometimes as a function of θ (fixing $x_1,\ldots,x_n$). In the latter case, the likelihood is sometimes denoted $L(\theta)$.

Theorem 1 (Factorization Theorem)**.** *T is a sufficient statistic for θ if the likelihood factorizes into the following form*

$$L(x_1,\ldots,x_n|\theta) = g(\theta, T(x_1,\ldots,x_n)) \cdot h(x_1,\ldots,x_n)$$

for some functions g, h.

Proof. We prove the theorem only for the discrete case (the continuous case requires different techniques). First assume the likelihood factorizes as above. Then

$$p(x_1,\ldots,x_n|T(x_1,\ldots,x_n)) = \frac{p(x_1,\ldots,x_n,T(x_1,\ldots,x_n))}{p(T(x_1,\ldots,x_n))} = \frac{p(x_1,\ldots,x_n)}{\sum_{y:T(y)=T(x)} p(y_1,\ldots,y_n)} = \frac{h(x_1,\ldots,x_n)}{\sum_{y:T(y)=T(x)} h(y_1,\ldots,y_n)}$$

which is not a function of θ. Conversely, assume that T is a sufficient statistic for θ. Then

$$L(x_1,\ldots,x_n|\theta) = p(x_1,\ldots,x_n|T(x_1,\ldots,x_n),\theta)p(T(x_1,\ldots,x_n)|\theta) = h(x_1,\ldots,x_n)g(T(x_1,\ldots,x_n),\theta).$$

□

Example: A sufficient statistic for Ber(θ) is $\sum X_i$ since

$$L(x_1,\ldots,x_n|\theta) = \prod_i \theta^{x_i}(1-\theta)^{1-x_i} = \theta^{\sum x_i}(1-\theta)^{n-\sum x_i} = g\left(\theta, \sum x_i\right)\cdot 1.$$

Example: A sufficient statistic for the uniform distribution $U([0,\theta])$ is $\max(X_1,\ldots,X_n)$ since

$$L(x_1,\ldots,x_n|\theta) = \prod_i \frac{1}{\theta}\cdot 1_{\{0\le x_i\le\theta\}} = \theta^{-n}\cdot 1_{\{\max(x_1,\ldots,x_n)\le\theta\}}\cdot 1_{\{\min(x_1,\ldots,x_n)\ge 0\}} = g(\theta,\max(x_1,\ldots,x_n))h(x_1,\ldots,x_n).$$

In the case that θ is a vector rather than a scalar, the sufficient statistic may be a vector as well. In this case we say that the sufficient statistic vector is jointly sufficient for the parameter vector θ. The definitions and factorization theorem carry over with little change.

Example: $T = (\sum X_i, \sum X_i^2)$ are jointly sufficient statistics for $\theta = (\mu,\sigma^2)$ for normally distributed data $X_1,\ldots,X_n \sim N(\mu,\sigma^2)$:

$$\begin{aligned} L(x_1,\ldots,x_n|\theta) &= \prod_i \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x_i-\mu)^2/(2\sigma^2)} = (2\pi\sigma^2)^{-n/2} e^{-\sum_i (x_i-\mu)^2/(2\sigma^2)} \\ &= (2\pi\sigma^2)^{-n/2} e^{-\sum_i x_i^2/(2\sigma^2)+2\mu\sum_i x_i/(2\sigma^2)-\mu^2/(2\sigma^2)} = g(\theta,T)\cdot 1 = g(\theta,T)\cdot h(x_1,\ldots,x_n) \end{aligned}$$

Clearly, sufficient statistics are not unique. From the factorization theorem it is easy to see that (i) the identity function $T(x_1,\ldots,x_n) = (x_1,\ldots,x_n)$ is a sufficient statistic vector and (ii) if T is a sufficient statistic for θ then so is any 1-1 function of T. A function that is not 1-1 of a sufficient statistic may or may not be a sufficient statistic. This leads to the notion of a minimal sufficient statistic.

Definition 3. *A statistic that is a sufficient statistic and that is a function of all other sufficient statistics is called a minimal sufficient statistic.*

In a sense, a minimal sufficient statistic is the smallest sufficient statistic and therefore it represents the ultimate data reduction with respect to estimating θ. In general, it may or may not exists.

Example: Since $T = (\sum X_i, \sum X_i^2)$ are jointly sufficient statistics for $\theta = (\mu,\sigma^2)$ for normally distributed data $X_1,\ldots,X_n \sim N(\mu,\sigma^2)$, then so are $(\bar{X}, S^2)$ which are a 1-1 function of $(\sum X_i, \sum X_i^2)$.

The following theorem provides a way of verifying that a sufficient statistic is minimal.

Theorem 2. *T is a minimal sufficient statistics if*

$$\frac{L(x_1,\ldots,x_n|\theta)}{L(y_1,\ldots,y_n|\theta)} \text{ is not a function of } \theta \quad \Leftrightarrow \quad T(x_1,\ldots,x_n) = T(y_1,\ldots,y_n).$$

Proof. First we show that T is a sufficient statistic. For each element in the range of T, fix a sample $y_1^t,\ldots,y_n^t$. For arbitrary $x_1,\ldots,x_n$ denote $T(x_1,\ldots,x_n) = t$ and

$$L(x_1,\ldots,x_n|\theta) = \frac{L(x_1,\ldots,x_n|\theta)}{L(y_1^t,\ldots,y_n^t|\theta)} L(y_1^t,\ldots,y_n^t|\theta) = h(x_1,\ldots,x_n)g(T(x_1,\ldots,x_n),\theta).$$

We show that T is a function of some other arbitrary sufficient statistic T'. Let x, y be such that $T'(x_1,\ldots,x_n) = T'(y_1,\ldots,y_n)$. Since

$$\frac{L(x_1,\ldots,x_n|\theta)}{L(y_1,\ldots,y_n|\theta)} = \frac{g'(T'(x_1,\ldots,x_n),\theta)h'(x_1,\ldots,x_n)}{g'(T'(y_1,\ldots,y_n),\theta)h'(y_1,\ldots,y_n)} = \frac{h'(x_1,\ldots,x_n)}{h'(y_1,\ldots,y_n)}$$

is independent of θ, $T(x_1,\ldots,x_n) = T(y_1,\ldots,y_n)$ and T is a 1-1 function of T'.

□

Example: $T = (\sum X_i, \sum X_i^2)$ is a minimal sufficient statistic for the Normal distribution since the likelihood ratio is not a function of θ iff $T(x) = T(y)$

$$\frac{L(x_1,\ldots,x_n|\theta)}{L(y_1,\ldots,y_n|\theta)} = e^{-\frac{1}{2\sigma^2}\sum (x_i-\mu)^2-(y_i-\mu)^2} = e^{-\frac{1}{2\sigma^2}(\sum x_i^2-\sum y_i^2)+\frac{\mu}{\sigma^2}(\sum x_i-\sum y_i)}.$$

Since $(\bar{X}, S^2)$ is a function of T, it is minimal sufficient statistic as well.