Statistical Inference

Lecture 03a

ANU - RSFAS

Last Updated: Mon Mar 12 17:12:26 2018

Principles of Data Reduction

- Scientists use information in a sample $X_1, \ldots, X_n$ to infer about an unknown parameter θ (could be a vector).
- The scientist usually wants to summarize a few key features of the data, which is usually done by computing statistics.
- A statistic $T(\boldsymbol{X}) = T(X_1, \ldots, X_n)$ defines a reduction of the data into a summary measure.
- A scientist may just wish to use or store $T(\boldsymbol{x})$ and will treat $\boldsymbol{x}$ and $\boldsymbol{y}$ the same if

$$T(\boldsymbol{x}) = T(\boldsymbol{y})$$

even though the samples may differ in some ways.

- While we typically no longer have need to store reduced versions of the data through statistics, the results can be useful for understanding models.

Sufficiency

Sufficiency Principle: If $T(X_1, \ldots, X_n)$ is a sufficient statistic for θ, then any inference about θ should depend on the sample $\boldsymbol{X}$ only through $T(X_1, \ldots, X_n)$.

Definition 2.5: A statistic $T(X_1, \ldots, X_n)$ is **sufficient** for θ if the conditional distribution of the sample $X_1, \ldots, X_n$ given $T(X_1, \ldots, X_n)$ does not depend on θ.

random

random

Sufficiency

- **Eg.** Let X_1, X_2, X_3 be a sample of size $n = 3$ from a Bernoulli distribution with parameter p (i.e., $P(X_i = 1) = p$).
- Consider the following two statistics:

$$T_1 = X_1 X_2 + X_3$$

$$T_2 = X_1 + X_2 + X_3$$

Want to see the joint dist. of x's given the statistic does not depend on p.

Sufficiency

a statistic

- Suppose that $T_1 = X_1X_2 + X_3 = 0$. This suggests one of the three possible outcomes:

$$\mathcal{X} = \{A = (0,0,0), B = (1,0,0), C = (0,1,0)\}$$

- Let's calculate the conditional distribution:

$$\begin{aligned} P(X_1 = 0, X_2 = 0, X_3 = 0 | T_1 = 0) &= \frac{P(X_1 = 0, X_2 = 0, X_3 = 0, T_1 = 0)}{P(T_1 = 0)} \\ &= \frac{P(X_1 = 0, X_2 = 0, X_3 = 0)}{P(A \text{ or } B \text{ or } C)} \\ &= \frac{(1-p)^3}{(1-p)^3 + 2p(1-p)^2} \\ &= \frac{1-p}{1+p} \end{aligned}$$

$P(A|B) = \frac{P(A \cap B)}{P(B)}$

still has p.

- Conditioning (i.e. knowing) the information from the statistics does not remove the parameter. So knowing the statistic is not enough. It is not sufficient.

Sufficiency

- Suppose that $T_2 = X_1 + X_2 + X_3 = 1$. This suggests one of the three possible outcomes:

$$\mathcal{X} = \{A = (1,0,0), B = (0,1,0), C = (0,0,1)\}$$

- Let's calculate the conditional distribution:
- We can then easily calculate the chance that the actual data set was $(0,1,0)$ as the conditional distribution:

$$\begin{aligned} P(X_1 = 0, X_2 = 1, X_3 = 0 | T_2 = 1) &= \frac{P(X_1 = 0, X_2 = 1, X_3 = 0, T_2 = 1)}{P(T_2 = 1)} \\ &= \frac{P(X_1 = 0, X_2 = 1, X_3 = 0)}{P(A \text{ or } B \text{ or } C)} \\ &= \frac{p(1-p)^2}{3p(1-p)^2} = \frac{1}{3} \end{aligned}$$

- Similar calculations show that for any value $T_2 = t$, the conditional distribution does not depend on p. So the statistic is sufficient.

Sufficiency

a statistic of sum of variables

- Generally if we have:
  $X_1, \ldots, X_n \sim \text{Bernoulli}(\theta)$ and $T(\boldsymbol{X}) = X_1 + \cdots + X_n$, then:

$$
\begin{aligned}
P(\boldsymbol{X} = \boldsymbol{x} | T(\boldsymbol{X}) = t(\boldsymbol{x})) &= \frac{P(\boldsymbol{X} = \boldsymbol{x} \text{ and } T(\boldsymbol{X}) = t(\boldsymbol{x}))}{P(T(\boldsymbol{X}) = t(\boldsymbol{x}))} \\
&= \frac{P(\boldsymbol{X} = \boldsymbol{x})}{P(T(\boldsymbol{X}) = t(\boldsymbol{x}))} \\
&= \frac{p(\boldsymbol{x}; \theta)}{q(T(\boldsymbol{x}); \theta)} \\
&= \frac{\prod \theta^{x_i}(1-\theta)^{1-x_i}}{\binom{n}{t}\theta^t(1-\theta)^{n-t}} \\
&= \frac{\theta^{\sum x_i}(1-\theta)^{n-\sum x_i}}{\binom{n}{t}\theta^t(1-\theta)^{n-t}} \\
&= \frac{\theta^t(1-\theta)^{n-t}}{\binom{n}{t}\theta^t(1-\theta)^{n-t}} = \frac{1}{\binom{n}{t}}
\end{aligned}
$$

some specific value

remove it.

Joint density

b/c indep.

binomial dist.

$t = \sum x_i$

$Y = \sum x_i$

$Y \sim$ binomial(n, p)

Say $n = 10$, $t = 3$

10 tosses, $t = 3$ tails.

$\binom{10}{3}$ different cases.

- The conditional distribution does not depend on θ, thus $T(\boldsymbol{X})$ is sufficient.

Sufficiency

Theorem 2.1 (the factorization theorem/criterion): Suppose $X_1, \ldots, X_n$, form a random sample from $f(x;\theta)$. Then $T(\boldsymbol{X})$ is a sufficiant statistic for θ **if and only if** there exists two non-negative functions K_1 and K_2, such that the likelihood $L(\theta; \boldsymbol{x})$ can be written:

$$f(\vec{x};\theta) = L(\theta; \boldsymbol{x}) = K_1\left[t(\boldsymbol{x}); \theta\right] \; K_2\left[\boldsymbol{x}\right]$$

Sufficiency

Proof (based on discrete distributions):

1. Suppose $T(\boldsymbol{X})$ is a sufficient statistic.

why? b/c $T(X)$ is sufficient, P_θ does not depend on it.

joint

working backwards

$$\begin{aligned} L(\theta; \boldsymbol{x}) &= P_\theta(\boldsymbol{X} = \boldsymbol{x}) \\ &= P_\theta(\boldsymbol{X} = \boldsymbol{x} \text{ and } T(\boldsymbol{X}) = t(\boldsymbol{x})) \\ &= P(\boldsymbol{X} = \boldsymbol{x} | T(\boldsymbol{X}) = t(\boldsymbol{x})) P_\theta(T(\boldsymbol{X}) = t(\boldsymbol{x})) \\ &= g(\boldsymbol{x}) h(t(\boldsymbol{x}); \theta) \end{aligned}$$

$$K_1 = h(t(\boldsymbol{x}); \theta) \quad K_2 = g(\boldsymbol{x})$$

Sufficiency

Proof (based on discrete distributions):

2. Assume that a factorization exists. Then we can write the marginal ~~distirbution~~ of $T(\boldsymbol{x})$ as:
distribution

$$
\begin{aligned}
f_{T(\boldsymbol{x})}(t) &= \sum_{\{\boldsymbol{x}: T(\boldsymbol{x})=t(\boldsymbol{x})\}} K_2(\boldsymbol{x}) K_1(t(\boldsymbol{x}); \theta) \\
&= K_1(t(\boldsymbol{x}); \theta) \sum_{\{\boldsymbol{x}: T(\boldsymbol{x})=t(\boldsymbol{x})\}} K_2(\boldsymbol{x})
\end{aligned}
$$

pull it out

Sufficiency

assume factorization exists

$$
\begin{aligned}
P(\boldsymbol{X} = \boldsymbol{x} | T(\boldsymbol{X}) = t) = \frac{L(\theta; \boldsymbol{x})}{f(t(\boldsymbol{x}); \theta)} &= \frac{K_2(\boldsymbol{x}) K_1(t; \theta)}{f(t; \theta)} \\
&= \frac{K_2(\boldsymbol{x}) K_1(t; \theta)}{K_1(t; \theta) \sum_{\{\boldsymbol{x}: T(\boldsymbol{x}) = t\}} K_2(\boldsymbol{x})} \\
&= \boxed{\frac{K_2(\boldsymbol{x})}{\sum_{\{\boldsymbol{x}: T(\boldsymbol{x}) = t\}} K_2(\boldsymbol{x})}}
\end{aligned}
$$

something does not depend on θ.

Sufficiency

- Example: Normally distributed data.

$$X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$$

1. What are the sufficient statistic(s) when μ is unknown and σ^2 is known?
2. What are the sufficient statistic(s) when μ and σ^2 is uknown?
3. What are the sufficient statistic(s) when μ is known and σ^2 is unknown?

① μ unknown, σ^2 known.

$$f(x;\theta) = f(x;\mu,\sigma^2)$$

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

$$f(x_1,\dots,x_n) \overset{(iid)}{=} f(x_1)f(x_2)\dots f(x_n)$$

$$= \prod_{i=1}^{n} (2\pi\sigma^2)^{-\frac{1}{2}} \exp\left(-\frac{1}{2\sigma^2}(x_i-\mu)^2\right)$$

$$= (2\pi\sigma^2)^{-\frac{1}{2}n} \exp\left(-\frac{1}{2\sigma^2}\sum \underline{(x_i-\mu)^2}\right)$$

$$= (2\pi\sigma^2)^{-\frac{1}{2}n} \exp\left(-\frac{1}{2\sigma^2}\sum\left[x_i^2 - 2x_i\mu + \mu^2\right]\right)$$

$$= (2\pi\sigma^2)^{-\frac{1}{2}n} \exp\left(-\frac{1}{2\sigma^2}\sum x_i^2 + \frac{\mu}{\sigma^2}\sum x_i - \frac{n}{2\sigma^2}\mu^2\right)$$

worried about μ

$$= \underbrace{(2\pi\sigma^2)^{-\frac{1}{2}n} \exp\left(-\frac{1}{2\sigma^2}\sum x_i^2\right)}_{\text{one quantity ①}} \cdot \underbrace{\exp\left(\frac{\mu}{\sigma^2}\sum x_i - \frac{n}{2\sigma^2}\mu^2\right)}_{\text{another quantity ②}}$$

K_2 K_1

$t = \sum x_i$

quantity ②: $\exp\left(\frac{\mu}{\sigma^2} t - \frac{n}{2\sigma^2}\mu^2\right)$

$= K_2 \cdot K_1$ where K_1 is quantity ②, sufficient statistic

② μ & σ^2 both unknown.

Starting from:

$$= (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2}\sum x_i^2 + \frac{\mu}{\sigma^2}\sum x_i - \frac{n}{2\sigma^2}\mu^2\right)$$

$K_1 = f(\vec{x}; \mu, \sigma^2) = L(\mu, \sigma^2; \vec{x})$

$K_2 = 1$

$t_1 = \sum x_i^2$, $t_2 = \sum x_i$ are sufficient statistics

$\downarrow S^2$ $\quad\quad$ $\downarrow \bar{X}$ also sufficient statistics