Biostat 602 Winter 2017

Lecture Set 2

Principles of Data Reduction

Premise

Reading: CB 6.1–6.2

We assume that the data was generated by a pdf (or pmf) that belongs to a class of pdfs (or pmfs).

$$\mathcal{P} = \{f_X(x|\theta), \theta \in \Omega \subset \mathbb{R}^p\}$$

For example $X \sim \text{Bernoulli}(\theta), \theta \in (0, 1) = \Omega \subset \mathbb{R}$.

We collect data in order to

- Estimate θ (point estimation)
- Perform tests of hypothesis about θ.
- Estimate confidence intervals for θ (interval estimation).
- Make predictions of future data.

Typical Questions

- What is the estimated probability of head given a series of observed coin tosses (H, H, T, T, T)? (**Point Estimation**)
- Given a series of coin tosses, can you tell whether the coin is biased or not? ($\theta = \frac{1}{2}$). (**Test of Hypothesis**)
- What is the plausible range of the true probability of head, given a series of coin tosses? (**Interval Estimation**)
- Given the series of coin tosses, can you predict what the outcome of the next coin toss? (**Prediction**)

Data Reduction

Data; $x_1, \cdots, x_n$: Realization of random variables $X_1, \cdots, X_n$. Often we deal with a random sample whereby $X_1, \cdots, X_n$ is i.i.d.

Define a function of data

$$T(\mathbf{X}) = T(x_1, \cdots, x_n) : \mathbb{R}^n \to \mathbb{R}^d$$

We wish this summary of data to

1. Be simpler than the original data, e.g. $d \leq n$.
2. Keep all the information about θ that is contained in the original data $x_1, \cdots, x_n$.

A **statistic** $T(\mathbf{X}) = T(X_1, \cdots, X_n)$ is a function of random variables $X_1, \cdots, X_n$. Clearly, $T(\mathbf{X})$ itself is a random variable.

Examples

- $T(\mathbf{X}) = \overline{X} = \frac{1}{n} \sum_{i=1}^n X_i$
- $T(\mathbf{X}) = med(X_1, X_2, \ldots, X_n)$
- $T(\mathbf{X}) = s^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \overline{X})^2$
- $T(\mathbf{X}) = \max(X_1, X_2, \ldots, X_n)$

Data Reduction as Partition of Sample Space

Data reduction can be represented as a partition of the sample space $\mathcal{X}$ determined by a statistic $T(\mathbf{X})$

Domain of T: $\mathcal{X}$

Range of T : $\mathcal{T} = \{t : t = T(\mathbf{X}) \text{ for some } \mathbf{x} \in \mathcal{X}\}$

Partition of $\mathcal{X}$: $A_t = \{\mathbf{x} : T(\mathbf{X}) = t, t \in \mathcal{T}\}$

Example

Suppose $X_i \sim$ iid Bernoulli(p) for $i = 1, 2, 3$, and $0 < p < 1$. Define $T(X_1, X_2, X_3) = X_1 + X_2 + X_3$

- What is the domain and range of T?
- How is the sample space partitioned by T?

Partition	X_1	X_2	X_3	$T(\mathbf{X}) = X_1 + X_2 + X_3$
A_0	0	0	0	0
A_1	0	0	1	1
	0	1	0	1
	1	0	0	1
A_2	0	1	1	2
	1	0	1	2
	1	1	0	2
A_3	1	1	1	3

Partition of the sample space based on $T(\mathbf{X})$ is "coarser" than the original sample space.

- There are 8 elements in the sample space $\mathcal{X}$.
- They are partitioned into 4 subsets
- Thus, $T(\mathbf{X})$ is simpler (or coarser) than $\mathbf{X}$.

Therefore, a data reduction can be achieved by $T(\mathbf{X})$.

Sufficiency

- Making original data "simpler" is one goal of ideal data reduction.
- The other goal is to make inference about an underlying parameter θ. Want a statistic that contains all information about θ. (**Sufficient statistic**)
- In the previous example, what is the parameter θ that $T(\mathbf{X})$ is trying to estimate?
- Does the proposed $T(\mathbf{X})$ keep the information about θ contained in $\mathbf{X}$ or not?

Sufficiency Principle

If $T(\mathbf{X})$ is sufficient for θ, then any inference about θ should depend on the sample $\mathbf{X}$ only through the value of $T(\mathbf{X})$. Thus, for any two sample points $\mathbf{x}$ and $\mathbf{y}$ such that $T(\mathbf{x}) = T(\mathbf{y})$, then the inference about θ should be the same whether $\mathbf{X} = \mathbf{x}$ or $\mathbf{X} = \mathbf{y}$ is observed.

Definition: A statistic $T(\mathbf{X})$ is sufficient for θ if the conditional distribution of the sample $\mathbf{X}$ given the value of $T(\mathbf{x})$ does not depend on θ.

Example 1: Let $X_1, \cdots, X_n$ be i.i.d. from a pdf f. Then the set of order statistics $T(\mathbf{X}) = (X_{(1)} < X_{(2)} < \cdots < X_{(n)})$ is sufficient since the joint pdf of the random sample can be written as

$$f(\mathbf{x}) = \prod_{i=1}^{n} f(x_i) = \prod_{i=1}^{n} f(x_{(i)}).$$

Theorem 6.2.2: Let $f_{\mathbf{X}}(\mathbf{x}|\theta)$ is a joint pdf or pmf of $\mathbf{X}$. Further let $q(t|\theta)$ be the pdf or pmf of $T(\mathbf{X})$. Then $T(\mathbf{X})$ is a sufficient statistic for θ if, for every $\mathbf{x} \in \mathcal{X}$, the ratio

$$\frac{f_{\mathbf{X}}(\mathbf{x}|\theta)}{q(T(\mathbf{x})|\theta)}$$

is constant as a function of θ.

Proof: (Discrete Case)

Assume that the ratio $f_{\mathbf{X}}(\mathbf{x}|\theta)/q(T(\mathbf{x})|\theta)$ is constant, then

$$\begin{aligned}
\Pr\left(\mathbf{X} = \mathbf{x} | T(\mathbf{X}) = t\right) &= \frac{\Pr\left(\mathbf{X} = \mathbf{x}, T(\mathbf{X}) = t\right)}{\Pr(T(\mathbf{X}) = t)} \\
&= \begin{cases} \dfrac{\Pr(\mathbf{X} = \mathbf{x})}{\Pr(T(\mathbf{X}) = t)} & \text{if } T(\mathbf{x}) = t \\[2ex] 0 & \text{otherwise} \end{cases} \\
&= \begin{cases} \dfrac{f_{\mathbf{X}}(\mathbf{x}|\theta)}{q(T(\mathbf{x})|\theta)} & \text{if } T(\mathbf{x}) = t \\[2ex] 0 & \text{otherwise} \end{cases}
\end{aligned}$$

which does not depend on θ by assumption. Therefore, $T(\mathbf{X})$ is a sufficient statistic for θ.

Example 2: Bernoulli Distribution

Let $X_1, \cdots, X_n \sim$ iid Bernoulli(p), $0 < p < 1$. Show that $T(\mathbf{X}) = \sum_{i=1}^n X_i$ is a sufficient statistic for p.

Proof: Let $x_1, \cdots, x_n$ be the realization corresponding to the random variables $X_1, \cdots, X_n$.

$$
\begin{aligned}
f_{\mathbf{X}}(\mathbf{x}|p) &= p^{x_1}(1-p)^{1-x_1} \times p^{x_2}(1-p)^{1-x_2} \times \cdots \times p^{x_n}(1-p)^{1-x_n} \\
&= p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i} \\
T(\mathbf{X}) &= \sum_{i=1}^n X_i \sim \text{Binomial}(n,p) \\
q(t|p) &= \binom{n}{t} p^t (1-p)^{n-t} \\
\frac{f_{\mathbf{X}}(\mathbf{x}|p)}{q(T(\mathbf{x})|p)} &= \frac{p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i}}{\binom{n}{\sum_{i=1}^n x_i} p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i}} \\
&= \frac{1}{\binom{n}{\sum_{i=1}^n x_i}} = \frac{1}{\binom{n}{T(\mathbf{x})}}
\end{aligned}
$$

By Theorem 6.2.2. $T(\mathbf{X})$ is a sufficient statistic for p.

Example 3: Let Let $X_1, \cdots, X_n \sim$ iid Normal$(\mu, 1)$. Show that the sample mean $\overline{X} = (X_1 + \cdots + X_n)/n$ is sufficient for μ.

Proof:

Factorization Teorem – Theorem 6.2.6

Let $f_{\mathbf{X}}(\mathbf{x}|\theta)$ denote the joint pdf or pmf of a sample $\mathbf{X}$. A statistic $T(\mathbf{X})$ is sufficient for θ, if and only if there exist functions $g(t|\theta)$ and $h(\mathbf{x})$ such that, for all sample points $\mathbf{x}$, and for all parameter points θ,

$$f_{\mathbf{X}}(\mathbf{x}|\theta) = g(T(\mathbf{x})|\theta)h(\mathbf{x})$$

Remarks

- θ can be vector valued and so can be T
- g is a function of $T(\mathbf{x})$ as well as of θ.
- h is a function of $\mathbf{x}$, but must be free of θ.

Proof for Discrete Distributions

only if part : sufficient $\Longrightarrow$ factorization

Suppose that $T(\mathbf{X})$ is a sufficient statistic

$$\begin{aligned}
f_{\mathbf{X}}(\mathbf{x}|\theta) &= \Pr(\mathbf{X} = \mathbf{x}|\theta) \\
&= \Pr(\mathbf{X} = \mathbf{x} \text{ and } T(\mathbf{X}) = T(\mathbf{x})|\theta) \\
&= \Pr(T(\mathbf{X}) = T(\mathbf{x})|\theta)\Pr(\mathbf{X} = \mathbf{x}|T(\mathbf{X}) = T(\mathbf{x}), \theta) \\
&= \Pr(T(\mathbf{X}) = T(\mathbf{x})|\theta)\Pr(\mathbf{X} = \mathbf{x}|T(\mathbf{X}) = T(\mathbf{x}))
\end{aligned}$$

Choose $g(t|\theta) = \Pr(T(\mathbf{X}) = t|\theta)$, and $h(\mathbf{x}) = \Pr\left(\mathbf{X} = \mathbf{x}|T(\mathbf{X}) = T(\mathbf{x})\right)$, then

$$f_{\mathbf{X}}(\mathbf{x}|\theta) = g(T(\mathbf{x})|\theta)h(\mathbf{x})$$

if part : factorization $\Longrightarrow$ sufficient

Assume that the factorization $f_{\mathbf{X}}(\mathbf{x}|\theta) = g(T(\mathbf{x})|\theta)h(\mathbf{x})$ holds and let $q(t|\theta)$ be the pmf of $T(\mathbf{X})$. Define $A_t = \{\mathbf{y} : T(\mathbf{y}) = t\}$. Then

$$q(t|\theta) = \Pr(T(\mathbf{X}) = t|\theta) = \sum_{\mathbf{y}\in A_t} f_{\mathbf{X}}(\mathbf{y}|\theta)$$

$$\begin{aligned}
\frac{f_{\mathbf{X}}(\mathbf{x}|\theta)}{q(T(\mathbf{x})|\theta)} &= \frac{g(T(\mathbf{x})|\theta)h(\mathbf{x})}{q(T(\mathbf{x})|\theta)} \\
&= \frac{g(T(\mathbf{x})|\theta)h(\mathbf{x})}{\sum_{\mathbf{y}\in A_{T(\mathbf{x})}} f_{\mathbf{X}}(\mathbf{y}|\theta)} \\
&= \frac{g(T(\mathbf{x})|\theta)h(\mathbf{x})}{\sum_{\mathbf{y}\in A_{T(\mathbf{x})}} g(T(\mathbf{y})|\theta)h(\mathbf{y})} \\
&= \frac{g(T(\mathbf{x})|\theta)h(\mathbf{x})}{g(T(\mathbf{x})|\theta)\sum_{\mathbf{y}\in A_{T(\mathbf{x})}} h(\mathbf{y})} \\
&= \frac{h(\mathbf{x})}{\sum_{\mathbf{y}\in A_{T(\mathbf{x})}} h(\mathbf{y})}
\end{aligned}$$

which is free of θ and hence by Theorem 6.2.2, $T(\mathbf{X})$ is sufficient for θ.

Example 4 (Bernoulli): Let $X_1, \cdots, X_n \sim$ iid Bernoulli(p), $0 < p < 1$.

$$\begin{aligned}
f_{\mathbf{X}}(\mathbf{x}|p) &= p^{x_1}(1-p)^{1-x_1}\cdots p^{x_n}(1-p)^{1-x_n} \\
&= p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i} \\
&= p^{T(\mathbf{x})}(1-p)^{n-T(\mathbf{x})} \;=\; g(T(\mathbf{x})|p)h(\mathbf{x}),
\end{aligned}$$

where $g(t|p) = p^t(1-p)^{n-t}, h(\mathbf{x}) = 1$. Then by Factorization Theorem $T(\mathbf{X}) = \sum_{i=1}^n X_i$ is a sufficient statistic for p.

Example 5: Normal Distribution with known variance
Let $X_1, \cdots, X_n$ iid $\mathcal{N}(\mu, \sigma^2)$ with σ^2 known.

$$f_{\mathbf{X}}(\mathbf{x}|\mu) = (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \overline{x})^2 - \frac{n}{2\sigma^2}(\overline{x} - \mu)^2\right) \tag{1}$$

Take

$$h(\mathbf{x}) \;=\; (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{2\sigma^2}\right)$$

and

$$g(t|\mu) \;=\; \Pr(T(\mathbf{X}) = t|\mu) = \exp\left(-\frac{n(t-\mu)^2}{2\sigma^2}\right)$$

Then $f_{\mathbf{X}}(\mathbf{x}|\mu) = h(\mathbf{x})g(T(\mathbf{x})|\mu)$ holds, and $T(\mathbf{X}) = \overline{X}$ is a sufficient statistic for μ.

Example 6: Normal Distribution with both parameters unknown

Both μ and σ^2 are unknown. The parameter is a vector : $\boldsymbol{\theta} = (\mu, \sigma^2)$. The problem is to use the Factorization Theorem to find a sufficient statistic for $\boldsymbol{\theta}$.

Since the parameter is two-dimensional it is natural to assume that the sufficient statistic is also two dimensional. Consider

$$\mathbf{T}(\mathbf{X}) = (T_1(\mathbf{X}), T_2(\mathbf{X})) \equiv \big(\frac{1}{n}\sum_{i=1}^{n} X_i, \sum_{i=1}^{n}(X_i - \overline{X})^2\big).$$

Take

$$\begin{aligned} h(\mathbf{x}) &= 1 \\ g(t_1, t_2|\mu, \sigma^2) &= (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}t_2 - \frac{n}{2\sigma^2}(t_1 - \mu)^2\right) \end{aligned}$$

Then, in view of (1)

$$f_{\mathbf{X}}(\mathbf{x}|\mu, \sigma^2) = g(T_1(\mathbf{x}), T_2(\mathbf{x})|\mu, \sigma^2)h(\mathbf{x})$$

Thus, $\mathbf{T}(\mathbf{X}) = (T_1(\mathbf{x}), T_2(\mathbf{x})) = \left(\overline{x}, \sum_{i=1}^n (x_i - \overline{x})^2\right)$ is sufficient for $\boldsymbol{\theta} = (\mu, \sigma^2)$.

Equivalently, $(\overline{x}, s^2)$ is also sufficient for $\boldsymbol{\theta}$, where $s^2 = (n-1)^{-1}T_2$ is the sample variance.

Example 7 (Discrete Uniform) Let $X_1, \cdots, X_n$ be iid observations uniformly drawn from $\{1, \cdots, \theta\}$, where θ is a positive integer. Find a sufficient statistic for θ.

The pmf of discrete uniform is given by

$$f_X(x|\theta) = \begin{cases} 1/\theta & x = 1, 2, \cdots, \theta \\ \\ 0 & \text{otherwise} \end{cases}$$

The joint pmf of $X_1, \cdots, X_n$ is

$$f_{\mathbf{X}}(\mathbf{x}|\theta) = \begin{cases} \theta^{-n} & x_i \in \{1, 2, \cdots, \theta\}, \quad i = 1, 2, \ldots, n \\ \\ 0 & \text{otherwise} \end{cases}$$

Question: How can you implement factorization theorem here?

Example 8: Assume $X_1, \cdots, X_n$ iid Uniform$(\theta, \theta+1)$, $-\infty < \theta < \infty$. Find a sufficient statistic for θ.

Proof: