Chapter 6: Principles of Data Reduction.

6.1 : Introduction:

$\overline{X} = (X_1, X_2, X_3, \ldots X_n)$

random sample

$\overline{x} = (x_1, x_2, x_3, \ldots x_n)$ is a realization from the random sample.

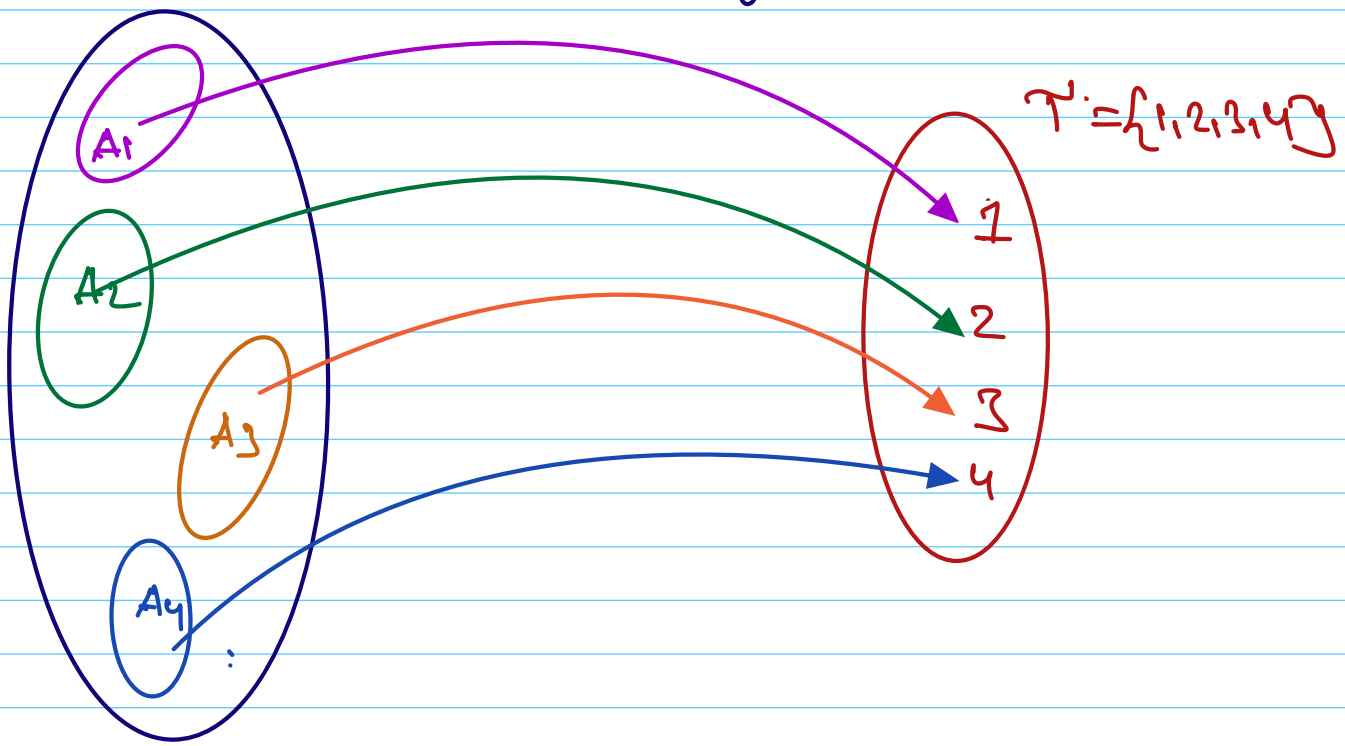
$T(X)$ is a Statistic.

Data reduction in term's of a Particular Statistic can be thought of as a Partition of the sample space.

$\mathcal{X}$: Sample space

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

Example: $\mathcal{X}$: sample space of $\overline{X} = (X_1, X_2, \ldots X_n)$



$T(x)$ partition's the sample space into sets A_t, $t \in \mathcal{T}$, defined by $A_t = \{ x : T(x) = t \}$

* The statistic summarizes the data in that, rather than reporting the entire sample x, id reports only

$T(x) = t$ or equivalently $x \in A_t$.

⇒ we study three Principles of Data reduction.

⇒ interested in data reduction, that do not discard important information about the unknown Parameter θ and methods that successfully discard information that is irrelavant as far as gaining knowledge about θ is concerned.

6.2 The Sufficiency Principle:

* A Sufficient statistic for a Parameter θ is a statistic that, in a certain sense, captures all the information about θ contained in the sample.

* Any additional information in the sample, besides the value of the sufficient statistic

does not contain any more information about θ.

Sufficiency Principle:

If $T(X)$ is a sufficient statistic for θ, then any inference about θ should depend on the sample X only through $T(x)$

* if x, y are two sample's such that $T(x) = T(y)$, then inference about θ should be same wheather $X = x$ or $X = y$ is observed.

Definition 6.2.1 :-

A statistic $T(X)$ is a sufficient statistic for θ if the conditional distribution of the sample X given the value of $T(x)$ does not depend on θ

$$\mathbb{P}(\bar{X} = \bar{x} \mid T(\bar{X}) = t)$$

$$= \mathbb{P}(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n \mid T(x_1, x_2, \ldots, x_n) = t)$$

does not depend on θ.

what happen's after conditioning on $T(X)$?

when we condition the sample x on the value of $T(X)$:

$$\mathbb{P}(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n \mid T(\bar{X}) = t)$$

the Parameter θ is no longer relevant. This is because.

(1) $T(\bar{X})$ already captured everything about θ.

(2) The remaining data X (after

knowing $T(x)$) contains no extra information about θ.

$\Rightarrow$ once we know $T(x)$, we have already extracted everything relevant to θ.

$\Rightarrow$ The remaining variability in X (after knowing $T(x)$) is unrelated to θ. it only depends on the internal structure of x, like how the data can be rearranged while keeping $T(x)$ constant

key takeaway:

The conditional distribution of x, given $T(x)$, does not depend on θ because $T(x)$ already contains all the information about θ.

$$P_\theta(X=x \mid T(X)=T(x))$$

$$= P_\theta(X_1=x_1, X_2=x_2, \ldots X_n=x_n \mid T(x_1,x_2,\ldots x_n)=T(x))$$

$$= \frac{P_\theta(X=x \text{ and } T(X)=T(x))}{P_\theta(T(X)=T(x))}$$

$$= \frac{P_\theta(x \mid \theta)}{q(T(x) \mid \theta)}$$

$$\Rightarrow \frac{P_\theta(X_1=x_1, X_2=x_2, \ldots, X_n=x_n \mid \theta)}{P_\theta(T(X)=T(x) \mid \theta)}$$

* $P(x|\theta)$ is the joint PMF of the sample X
* $q(t|\theta)$ is the PMF of $T(X)$

$\Rightarrow$ $T(X)$ is a sufficient statistic for θ

$\Updownarrow$

$\forall x \in X \in \mathbb{R}^n$, the above ratio of Pmf's is constant as a function of θ.

Theorem 6.2.2 :

If $P(X|\theta)$ is the joint pdf or pmf of X and $q(t|\theta)$ is the pdf or pmf of $T(x)$ is a sufficient statistic for θ if, for every x in the sample space, the ratio $\frac{P(X|\theta)}{q(t|\theta)}$ is constant as a function of θ.

EX 6.2.3 : (Binomial sufficient statistic)

$X_1, X_2, \ldots X_n$ iid $\text{Ber}(\theta)$

$0 < \theta < 1$

we need to show $T(x) = x_1 + x_2 + \cdots + x_n$ is a sufficient statistic of θ.

$$\frac{P(x|\theta)}{q(T(x)|\theta)} = \frac{\prod \theta^{x_i} (1-\theta)^{1-x_i}}{\binom{n}{t} \theta^t (1-\theta)^{n-t}}$$

$$= \frac{\theta^{\sum x_i} (1-\theta)^{n - \sum x_i}}{\binom{n}{t} \theta^t (1-\theta)^{n-t}}$$

$$= \frac{\theta^t \cdot (1-\theta)^{n-t}}{\binom{n}{t} \theta^t (1-\theta)^{n-t}}$$

$$= \frac{1}{\binom{n}{\sum x_i}}$$

$\Rightarrow$ does not depend on θ

$\Rightarrow$ $T(x)$ is sufficient statistic of θ.

Ex 6.2.4

(Normal statistic)

$X_1, X_2, \cdots X_n$ iid $N(\mu, \sigma^2)$

σ^2 is known.

$T(X) = \bar{X} = \frac{1}{n}(X_1 + X_2 + \cdots + X_n)$

is sufficient statistic of μ.

$$f(X_1, X_2, \cdots X_n \mid T(X) = t) = \frac{f(X_1, X_2, \cdots, X_n, T(X) = t \mid \mu)}{f(T(X) = t \mid \mu)}$$

$$f(x \mid \mu) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2\right)$$

$$\bar{X} = \frac{1}{n}(X_1 + X_2 + \cdots + X_n) = T(X)$$

$$T(X) \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

$$q(T(x)\,|\,\mu) = \frac{1}{(2\pi\sigma^2)^{1/2}} \cdot \exp\left(\frac{-1}{2\frac{\sigma^2}{n}}(\bar{x}-\mu)^2\right)$$

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

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

$$\frac{f(x|\mu)}{q(T(x)|\mu)} = \frac{\frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(\frac{-1}{2\sigma^2}\left(\sum x_i^2\right)\right) \exp\left(\frac{-1}{2\frac{\sigma^2}{n}}\left(\mu^2 - 2\mu\bar{x}\right)\right)}{\frac{1}{\left(2\pi\frac{\sigma^2}{n}\right)^{1/2}} \exp\left(-\frac{1}{2\frac{\sigma^2}{n}}\left(\bar{x}^2 + \mu^2 - 2\mu\bar{x}\right)\right)}$$

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

$$= \frac{1}{\sqrt{n}} \cdot \frac{1}{(2\pi\sigma^2)^{\frac{n-1}{2}}} \exp\left(\frac{-1}{2\sigma^2}\left(\sum_{i=1}^{n}(x_i^2 - \bar{x}^2)\right)\right)$$

this does not depend on μ

$\Rightarrow$ $\bar{X}$ is sufficient statistic.

To use the definition of sufficient statistic, we must guess a statistic $T(x)$ to be sufficient, find the pdf or pmf of $T(x)$ and then check the ratio of pdf's or pm's does not depend on θ.

$\Rightarrow$ 1st step requires a good deal of intuition

$\Rightarrow$ 2nd step requires a tedious analysis.

Theorem 6.2.6 (Factorization theorem)

Let $f(x|\theta)$ denote the joint pdf or pmf of a sample X.

A statistic $T(X)$ is a sufficient statistic for θ

$\Leftrightarrow$

$\exists$ functions $g(t|\theta)$ and $h(x)$ such that, $\forall$ sample points x and all parameter's points θ

$$f(x|\theta) = g(T(x)|\theta)\, h(x)$$

Proof: Proof only for discrete distributions.

one direction. ($\Rightarrow$)

Suppose $T(X)$ is sufficient statistic.

Choose $g(t|\theta) = \mathbb{P}_\theta(T(X)=t)$

$h(x) = \mathbb{P}(X=x \mid T(X)=T(x))$

because $T(x)$ is sufficient $h(x)$ is independent of θ.

$$f(X=x) = P_\theta(X=x)$$

$$= P_\theta(X=x \text{ and } T(X)=x)$$

$$= P_\theta(X=x \mid T(x)=T(x)) P(T(X)=T(x))$$

$$= h(x)\, g(T(x)\mid\theta)$$

if $T(X)$ is sufficient statistic $\Longrightarrow$ factorization exists

($\Longleftarrow$) other direction

Suppose the $f(x|\theta) = g(T(x)|\theta)\, h(x)$ factorization exists.

Let $q(t|\theta)$ is pmf of $T(X)$

To show that $T(X)$ is sufficient statistic we examine the ratio

$$\frac{f(x|\theta)}{q(T(x)|\theta)}$$

Suppose $A_{T(x)} = \{y : T(y)=T(x)\}$

all the elements in X s.t $T(y) = T(x)$

$$\frac{f(x|\theta)}{q(T(x)|\theta)} = \frac{g(T(x)|\theta)\, h(x)}{q(T(x)|\theta)}$$

$$= \frac{g(T(x)|\theta)\, h(x)}{\sum_{T(x)} f(y|\theta)}$$

$$= \frac{g(T(x)|\theta)\, h(x)}{\sum_{T(x)} g(T(y)|\theta)\, h(y)}$$

$$= \frac{g(T(x)|\theta)\, h(x)}{g(T(y)|\theta) \sum_{T(x)} h(y)}$$

$$= \frac{h(x)}{\sum_{T(x)} h(y)}$$

The ratio doesnot depend on $\theta \Rightarrow$ factorization exists $\Rightarrow T(x)$ is sufficient

EX 6.2.7 (Continuation of Example 6.2.4)

for the normal distribution

$X_1, X_2, \ldots, X_n$ iid $N(\mu, \sigma^2)$

σ^2 = known

μ = unknown

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

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

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

$$= h(x) \cdot g(\textstyle\sum x_i \mid \mu)$$

$\Rightarrow$ $T(x) = \sum x_i$ is a sufficient statistic.

Ex : 6.2.8 (uniform sufficient statistic)

$X_1, X_2, \cdots X_n$ iid Discrete uniform $(1, 2, \cdots \theta)$

where θ is unknown parameter.

$$f(x|\theta) = \begin{cases} \frac{1}{\theta} & \text{if } x = 1, 2, 3, \cdots \theta \\ 0 & \text{o.w} \end{cases}$$

$$f(x_1, x_2, \cdots x_n|\theta) = \begin{cases} \theta^{-n} & x_i \in \{1, 2, \cdots \theta\} \text{ for } i = 1, 2, \cdots n \\ 0 & \text{o.w} \end{cases}$$

$$= \begin{cases} \theta^{-n} & x_i \in \{1, 2, \cdots\} \text{ for } i = 1, 2, \cdots n \\ & \max_i x_i \leq \theta \end{cases}$$

$$T(x) = \max_i x_i$$

$$h(x) = \begin{cases} 1 & x_i \in \{1, 2, \cdots\} \text{ for } i = 1, 2, \cdots n \\ 0 & \text{o.w} \end{cases}$$

$$g(t|\theta) = \begin{cases} \frac{1}{\theta^n} & t \le \theta \\ 0 & o.w \end{cases}$$

$$\Rightarrow \quad f(x|\theta) = g(T(x)|\theta)\, h(x)$$

another method using Indicator function's.

<u>in dicator function:</u>

$$I_A(x) = \begin{cases} 1 & \text{if } x \in A \\ 0 & o.w \end{cases}$$

$$\mathbb{P}_{I_A}(x=1) = \mathbb{P}(A)$$

$$\mathbb{P}_{I_A}(x=0) = \mathbb{P}(\Omega \setminus A)$$

Let $\mathbb{N} = \{1, 2, \dots\}$

$\mathbb{N}_\theta = \{1, 2, \dots \theta\}$

Joint pdf of $x_1, x_2, \ldots x_n$

$$f(x|\theta) = \prod_{i=1}^{n} \frac{1}{\theta} I_{N_\theta}(x_i)$$

$$= \theta^{-n} \prod_{i=1}^{n} I_{N_\theta}(x_i)$$

define $T(X) = \max_i x_i$

$$\prod_{i=1}^{n} I_{N_\theta}(x_i) = \left[\prod_{i=1}^{n} I_N(x_i) \right] I_{N_\theta}(T(x))$$

$$f(x|\theta) = \theta^{-n} I_{N_\theta}(T(x)) \left(\prod_{i=1}^{n} I_N(x_i) \right)$$

The sufficient statistic can be vector also

$$T(X) = (T_1(X), T_2(X) \cdots, T_n(X))$$

when Parameter is also vector

$$\theta = (\theta_1, \theta_2, \ldots \theta_n)$$

Ex:- 6.2.9 (Normal Sufficient statistic, both Parameter's unknown)

$$X_1, X_2, X_3, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$$

Both μ, σ^2 are unknown

$$\Rightarrow \theta = (\mu, \sigma^2)$$

$$f(x|\mu,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(\frac{-1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right)$$

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

$$\sum_{i=1}^{n}x_i^2 + n\mu^2 - 2\mu\, n\bar{x}$$

$$= (n-1)S^2 + n\bar{x}^2 + n\mu^2 - 2\mu n\bar{x}$$

$$= (n-1)S^2 + n(\bar{x}^2 + \mu^2 - 2\mu\bar{x})$$

$$= (n-1)s^2 + n(\bar{x}-\mu)^2$$

$$\Rightarrow f(x|\mu,\sigma^2)$$

$$= \frac{1}{(2\pi\sigma^2)^{n/2}} \cdot \exp\left(\frac{-1}{2\sigma^2}\left((n-1)s^2 + n(\bar{x}-\mu)^2\right)\right)$$

$$= \qquad T_1(x) = \bar{x} \quad , T_2(x) = s^2$$

$$\Rightarrow$$

$$f(x|\mu,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \cdot \exp\left(\frac{-1}{2\sigma^2}\left((n-1)t_2 + n(t_1-\mu)^2\right)\right)$$

$$= g(T_1(x), T_2(x) \mid \mu,\sigma^2) \cdot h(x)$$

$$h(x) = 1$$

Therefore $(\bar{x}, s^2)$ is a sufficient statistic for (μ,σ^2)

Theorem 6.2.10 :- Let $x_1, x_2, \ldots, x_n \overset{iid}{\sim} f(x|\theta)$

such that

$$f(x|\theta) = h(x)\, c(\theta) \exp\left(\sum_{i=1}^{k} w_i(\theta)\, t_i(x) \right)$$

where $\theta = (\theta_1, \theta_2, \ldots \theta_d)$ $d \leq k$ Then

$$T(x) = \left(\sum_{j=1}^{n} t_1(x_j), \ldots, \sum_{j=1}^{n} t_k(x_j) \right)$$

is a sufficient statistic for θ