IE605: Engineering Statistics

Lecture 09

Manjesh K. Hanawal

Previous Lecture:

- Generating Random Samples

This Lecture:

- Data reduction
- Sufficiency principle
- Sufficient statistics
- factorization theorem

Data Reduction

- A random sample $x := (x_1, x_2, \ldots, x_n)$ is a long list of numbers hard to interpret
- We may wish to summarize the information in it by determining a few key features
  - sample mean
  - sample variance
  - smallest value
  - largest value
- Any statistic $T(\cdot)$ defines a data reduction or data summary
- When we have statistics $T(x)$ we treat sample x and y the same as long as $T(x) = T(y)$

- $T(x) = \sum_{i=1}^{n} x_i$ gives the sum of the sample x. There could be many other samples with the same sum of sample space
- Data reduction can be though of partition
- For any t, $A_t = \{x : T(x) = t\}$
- What are advantages and consequences of data reduction?

Statistic are associated with parameters.

Principles of data reduction:

- Sufficiency Principle: Parameter information by summarization of data
- Likelihood Principle: function of parameter determined by observed sample
- Equivariance Principle: preserves important features of model

Sufficiency Principle

- A sufficient statistic for a parameter θ, captures all information about θ contained in the sample
- knowledge of individual samples does not contain any more information about θ
- $x = (x_1, x_2, \dots, x_n)$ and $T(x)$ is a statics. $T(x)$ contains all information about θ. Further knowing x_3 or x_4 or any other component provides no additional information

> Sufficency 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 and y are such that $T(x) = T(y)$, then the inference about θ should be the same whether $X = x$ or $X = y$ is observed.

Characterization of Sufficient Statistic

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

- $P_\theta(X = x | T(X) = t)$ is independent of θ if $T(X)$ is sufficient
- Sufficient statistic captures all the information about θ

Example:

- Statistician 1: Generate sample x and computes $T(x)$
- Statistician 2: is informed about $T(x)$ only not x
- if T is a sufficient statistic for parameter θ, both the statistician have same amount of information about θ.

Sufficient Statistic contd..

Let $T(X)$ is a sufficient Statistic for parameter θ

- $P_\theta(X = x|T(X) = t)$ doesn't depend on θ for all t
- Probability is non zero for sample $T(x) = t$
- We consider $P_\theta(X = x|T(X) = T(x))$. Note $\{X = x\} \subset \{y : T(y) = T(x)\}$. Then

$$\begin{aligned} P_\theta\left(X = x|T(X) = T(x)\right) &= \frac{P_\theta(X = x, T(X) = T(x))}{P_\theta(T(X) = T(x))} \\ &= \frac{P_\theta(X = x)}{P_\theta(T(X) = T(x))} \\ &= \frac{p(x|\theta)}{q(T(x)|\theta)} \end{aligned}$$

- $p(x|\theta)$ is the joint pmf of sample X and $q(t|\theta)$ is the distribution of statistic $T(X)$.

Alternate definition of sufficient statistic

> If $p(x|\theta)$ is a joint pdf/pmf of X and $q(t|\theta)$ is pdf/pmf of $T(X)$, then $T(X)$ is a sufficient statistic for θ if for every x, the ratio $p(x|\theta)/q(T(x)|\theta)$ does not depend on θ.

Example 1: Bernoulli sufficient statistics Let $X = (X_1, X_2, \ldots, X_n)$ be iid $\sim Ber(\theta)$. $T(X) = \sum_{i=1}^{n} X_i$ is a sufficient statistic for θ. For $x = (x_1, x_2, \ldots, x_n)$, define $t = \sum_{i=1}^{n} x_i$

$$\begin{aligned}
\frac{p(x|\theta)}{q(T(x)|\theta)} &= \frac{\prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i}}{\binom{n}{t}\theta^t(1-\theta)^{n-t}} \\
&= \frac{\theta^{\sum x_i}(1-\theta)^{\sum(1-x_i)}}{\binom{n}{t}\theta^t(1-\theta)^{n-t}} \\
&= \frac{\theta^t(1-\theta)^{n-t}}{\binom{n}{t}\theta^t(1-\theta)^{n-t}} \\
&= \frac{1}{\binom{n}{t}} = \frac{1}{\binom{n}{\sum x_i}} \quad \text{(not a function of } \theta\text{)}
\end{aligned}$$

Example 2: Normal sufficient statistics $X = (X_1, X_2, \ldots, X_n)$ is iid $\sim \mathcal{N}(\mu, \sigma^2)$. Sample mean $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, is a sufficient statistic for μ.

$$p(x|\mu) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-(x_i - \mu)^2/2\sigma^2\}$$

We know $\bar{X} \sim \mathcal{N}(\mu, \sigma^2/n)$

$$q(\bar{X}|\mu) = \frac{\sqrt{n}}{\sqrt{2\pi\sigma^2}} \exp\{-n(\bar{x} - \mu)^2/2\sigma^2\}$$

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

(does not depend on μ)

Factorization theorem

How to come up with a sufficient statistic for a parameter

- guess a statistics (required good intuition)
- find its pdf/pmf (expression can be tedious)
- find the ration to acertain

Factorization Theorem: For a random sample X with pdf/pmf $f(x|\theta)$, let $T(X)$ is a statistic for θ. Then $T(X)$ is sufficient statistic **if and only if** if there exists functions $g(t|\theta)$ and $h(x)$ such that, for all x and parameters points θ

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

Proof of factorization theorem for discrete case

Assume $T(X)$ is a sufficient statistic ($\implies$).

- $P_\theta(X = x | T(X) = T(x)$ does not depend on θ
- Set $h(x) = P_\theta(X = x | T(X) = T(x))$ (valid!)
- Set $g(t|\theta) = P_\theta(T(X) = t)$ (valid!)

$$\begin{aligned} f(x|\theta) &= P_\theta(X = x)) \\ &= P_\theta(X = x, T(X) = T(x)) \\ &= P_\theta(X = x | T(X) = T(x)) P_\theta(T(X) = T(x)) \\ &= h(x) g(T(x)|\theta) \end{aligned}$$

Proof of factorization theorem for discrete case contd..

Assume the factorization holds ($\Longleftarrow$)

- We show that ratio $f(x|\theta)/q(T(x)|\theta)$ does not depend on θ
- Choose x. Let $t = T(x)$ and $A_t = \{y : T(y) = t\}$

$$
\begin{aligned}
\frac{f(x|\theta)}{q(T(x)|\theta)} &= \frac{g(T(x)|\theta)h(x)}{q(T(x)|\theta)} \\
&= \frac{g(T(x)|\theta)h(x)}{\sum_{y\in A_t} f(y|\theta)} \\
&= \frac{g(T(x)|\theta)h(x)}{\sum_{y\in A_t} g(T(y)|\theta)h(y)} \\
&= \frac{g(T(x)|\theta)h(x)}{g(T(x)|\theta)\sum_{y\in A_t} h(y)} \\
&= \frac{h(x)}{\sum_{y\in A_t} h(y)} \quad \text{does not depend on } \theta
\end{aligned}
$$

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