Lecture 4

Sufficient Statistics. Factorization Theorem

1 Sufficient statistics

Let $f(x|\theta)$ with $\theta \in \Theta$ be some parametric family. Let $X = (X_1, ..., X_n)$ be a random sample from distribution $f(x|\theta)$. Suppose we would like to learn parameter value θ from our sample. The concept of sufficient statistic allows us to separate information contained in X into two parts. One part contains all the valuable information as long as we are concerned with parameter θ, while the other part contains pure noise in the sense that this part has no valuable information. Thus, we can ignore the latter part.

Definition 1. Statistic $T(X)$ is sufficient for θ if the conditional distribution of X given $T(X)$ does not depend on θ.

Let $T(X)$ be a sufficient statistic. Consider the pair $(X, T(X))$. Obviously, $(X, T(X))$ contains the same information about θ as X alone, since $T(X)$ is a function of X. But if we know $T(X)$, then X itself has no value for us since its conditional distribution given $T(X)$ is independent of θ. Thus, by observing X (in addition to $T(X)$), we cannot say whether one particular value of parameter θ is more likely than another. Therefore, once we know $T(X)$, we can discard X completely.

Example Let $X = (X_1, ..., X_n)$ be a random sample from $N(\mu, \sigma^2)$. Suppose that σ^2 is known. Thus, the only parameter is μ ($\theta = \mu$). We have already seen that $T(X) = \overline{X}_n \sim N(\mu, \sigma^2/n)$. Let us calculate the conditional distribution of X given $T(X) = t$. First, note that

$$\begin{aligned}
\sum_{i=1}^{n}(x_i - \mu)^2 - n(\overline{x}_n - \mu)^2 &= \sum_{i=1}^{n}(x_i - \overline{x}_n + \overline{x}_n - \mu)^2 - n(\overline{x}_n - \mu)^2 \\
&= \sum_{i=1}^{n}(x_i - \overline{x}_n)^2 + 2\sum_{i=1}^{n}(x_i - \overline{x}_n)(\overline{x}_n - \mu) \\
&= \sum_{i=1}^{n}(x_i - \overline{x}_n)^2
\end{aligned}$$

Therefore

$$\begin{aligned} f_{X|T(X)}(x|T(X)=T(x)) &= \frac{f_X(x)}{f_T(T(x))} \\ &= \frac{\exp\{-\sum_{i=1}^n (x_i-\mu)^2/(2\sigma^2)\}/((2\pi)^{n/2}\sigma^n)}{\exp\{-n(\overline{x}_n-\mu)^2/(2\sigma^2)\}/((2\pi)^{1/2}\sigma/n^{1/2})} \\ &= \exp\{-\sum_{i=1}^n (x_i-\overline{x}_n)^2/(2\sigma^2)\}/((2\pi)^{(n-1)/2}\sigma^{n-1}/n^{1/2}) \end{aligned}$$

which is independent of μ. We conclude that $T(X)=\overline{X}_n$ is a sufficient statistic for our parametric family. Note, however, that $\overline{X}_n$ is not sufficient if σ^2 is not known.

Example Let $X=(X_1,...,X_n)$ be a random sample from a Poisson(λ) distribution. From Problem Set 1, we know that $T=\sum_{i=1}^n X_i \sim$ Poisson($n\lambda$). So

$$f_{X|T}(x|T=\sum_{i=1}^n x_i) = \prod_{i=1}^n (e^{-\lambda}\lambda^{x_i}/x_i!)/(e^{-\lambda n}\lambda^{\sum_{i=1}^n x_i}/(\sum_{i=1}^n x_i)!) = (\sum_{i=1}^n x_i)!/\prod_{i=1}^n x_i!$$

which is independent of λ. We conclude that $T=\sum_{i=1}^n X_i$ is a sufficient statistic in this case.

2 Factorization Theorem

The Factorization Theorem gives a general approach for how to find a sufficient statistic:

Theorem 2 (Factorization Theorem)**.** *Let $f(x|\theta)$ be the pdf of X. Then $T(X)$ is a sufficient statistic if and only if there exist functions $g(t|\theta)$ and $h(x)$ such that $f(x|\theta)=g(T(x)|\theta)h(x)$.*

Proof. Let $l(t|\theta)$ be the pdf of $T(X)$.

Suppose $T(X)$ is a sufficient statistic. Then $f_{X|T(X)}(x|T(X)=T(x))=f_X(x|\theta)/l(T(x)|\theta)$ does not depend on θ. Denote it by $h(x)$. Then $f(x|\theta)=l(T(x)|\theta)h(x)$. Denoting l by g yields the result in one direction.

In the other direction we will give a "sloppy" proof. Denote $A(x)=\{y: T(y)=T(x)\}$. Then

$$l(T(x)|\theta) = \int_{A(x)} f(y|\theta)dy = \int_{A(x)} g(T(y)|\theta)h(y)dy = g(T(x)|\theta)\int_{A(x)} h(y)dy.$$

So

$$\begin{aligned} f_{X|T(X)}(x|T(X)=T(x)) &= \frac{f(x|\theta)}{l(T(x)|\theta)} \\ &= \frac{g(T(x)|\theta)h(x)}{g(T(x)|\theta)\int_{A(x)} h(y)dy} \\ &= \frac{h(x)}{\int_{A(x)} h(y)dy}, \end{aligned}$$

which is independent of θ. We conclude that $T(X)$ is a sufficient statistic. $\square$

Example Let us show how to use the factorization theorem in practice. Let $X_1, ..., X_n$ be a random sample from $N(\mu, \sigma^2)$ where both μ and σ^2 are unknown, i.e. $\theta = (\mu, \sigma^2)$. Then

$$\begin{aligned} f(x|\theta) &= \exp\{-\sum_{i=1}^{n}(x_i-\mu)^2/(2\sigma^2)\}/((2\pi)^{n/2}\sigma^n) \\ &= \exp\{-[\sum_{i=1}^{n}x_i^2 - 2\mu\sum_{i=1}^{n}x_i + n\mu^2]/(2\sigma^2)\}/((2\pi)^{n/2}\sigma^n). \end{aligned}$$

Thus, $T(X) = (\sum_{i=1}^n X_i^2, \sum_{i=1}^n X_i)$ is a sufficient statistic (here $h(x) = 1$ and g is the whole thing). Note that in this example we actually have a pair of sufficient statistics. In addition, as we have seen before,

$$\begin{aligned} f(x|\theta) &= \exp\{-[\sum_{i=1}^{n}(x_i-\overline{x}_n)^2 + n(\overline{x}_n-\mu)^2]/(2\sigma^2)\}/((2\pi)^{n/2}\sigma^n) \\ &= \exp\{-[(n-1)s_n^2 + n(\overline{x}_n-\mu)^2]/(2\sigma^2)\}/((2\pi)^{n/2}\sigma^n). \end{aligned}$$

Thus, $T(X) = (\overline{X}_n, s_n^2)$ is another sufficient statistic. Yet another sufficient statistic is $T(X) = (X_1, ..., X_n)$. Note that $\overline{X}_n$ is not sufficient in this example.

Example A less trivial example: let $X_1, ..., X_n$ be a random sample from $U[\theta, 1+\theta]$. Then $f(x|\theta) = 1$ if $\theta \le \min_i X_i \le \max_i X_i \le 1+\theta$ and 0 otherwise. In other words, $f(x|\theta) = I\{\theta \le X_{(1)}\}I\{1+\theta \ge X_{(n)}\}$. So $T(X) = (X_{(1)}, X_{(n)})$ is sufficient.

3 Minimal Sufficient Statistics

Could we reduce sufficient statistic $T(X)$ in the previous example even more? Suppose we have two statistics, say, $T(X)$ and $T^\star(X)$. We say that $T^\star$ is not bigger than T if there exists some function r such that $T^\star(X) = r(T(X))$. In other words, we can calculate $T^\star(X)$ whenever we know $T(X)$. In this case when T^* changes its value, statistic T must change its value as well. In this sense T^* does not give less of an information reduction than T.

Definition 3. A sufficient statistic $T^\star(X)$ is called *minimal* if for any sufficient statistic $T(X)$ there exists some function r such that $T^\star(X) = r(T(X))$.

Thus, in some sense, the minimal sufficient statistic gives us the greatest data reduction without a loss of information about parameters. The following theorem gives a characterization of minimal sufficient statistics:

Theorem 4. *Let $f(x|\theta)$ be the pdf of X and $T(X)$ be such that, for any x, y, statement $\{f(x|\theta)/f(y|\theta)$ does not depend on $\theta\}$ is equivalent to statement $\{T(x) = T(y)\}$. Then $T(X)$ is minimal sufficient.*

Proof. This is a "sloppy" proof again.

First, we show that the statistic described above is in fact a sufficient statistic. Let $\mathcal{X}$ be the space of possible values of x. We divide it into equivalence classes A_x. For any x, let $A_x = \{y \in \mathcal{X} : T(y) = T(x)\}$. Then, for any x, y, A_x either coincides with A_y or A_x and A_y have no common elements. Thus, we can choose subset X of $\mathcal{X}$ such that $\cup_{x \in X} A_x = \mathcal{X}$ and A_x has no common elements with A_y for any $x, y \in X$. Then $T(x) \neq T(y)$ if $x, y \in X$ and $x \neq y$. Then there is a function g of $x \in X$ and $\theta \in \Theta$ such that $f(x|\theta) = g(T(x)|\theta)$. Fix some $\theta \in \Theta$. For any $x' \in \mathcal{X}$, there is $x(x') \in X$ such that $x' \in A_x$, i.e. $T(x) = T(x')$. Denote $h(x') = f(x'|\theta)/f(x(x')|\theta)$. Then for any $\theta' \in \Theta$, $f(x'|\theta')/f(x(x')|\theta') = h(x')$. So $f(x'|\theta') = g(T(x')|\theta')h(x')$. Thus, $T(X)$ is sufficient.

Let us now show that $T(X)$ is actually minimal sufficient in the sense of Definition 3. Take any other sufficient statistic, $T^\star(X)$. Then there exist functions $g^\star$ and $h^\star$ such that $f(x|\theta) = g^\star(T^\star(x)|\theta)h^\star(x)$. If $T^\star(x) = T^\star(y)$ for some x, y, then

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

which is independent of θ. Thus $T(x) = T(y)$ as well. So we can define a function r such that $T(X) = r(T^\star(X))$. □

Example Let us now go back to the example with $X_1, ..., X_n \sim U[\theta, 1+\theta]$. Ratio $f(x|\theta)/f(y|\theta)$ is independent of θ if and only if $x_{(1)} = y_{(1)}$ and $x_{(n)} = y_{(n)}$ which is the case if and only if $T(x) = T(y)$. Therefore $T(X) = (X_{(1)}, X_{(n)})$ is minimal sufficient.

Example Let $X_1, ..., X_n$ be a random sample from the Cauchy distribution with parameter θ, i.e. the distribution with the pdf $f(x|\theta) = 1/(\pi(x-\theta)^2)$. Then $f(x_1, ..., x_n|\theta) = 1/(\pi^n \prod_{i=1}^n (x_i - \theta)^2)$. By the theorem above, $T(X) = (X_{(1)}, ..., X_{(n)})$ is minimal sufficient.