Independence of Two Random Variables

Definition. We define two random variables X and Y to be independent if

$$\Pr(X = x, Y = y) = \Pr(X = x)\Pr(Y = y) \quad \text{for all} \quad x \in \mathcal{X} \quad \text{and} \quad y \in \mathcal{Y} \tag{1.1}$$

where $\mathcal{X}$ and $\mathcal{Y}$ are the alphabet of the random variable X and that of Y respectively.

Example) Let $X \sim p(x)$ for $\mathcal{X} = \{0,1\}$ and $Y \sim p(y)$ for $\mathcal{Y} = \{0,1\}$. Let $p(x=0) = 0.4$ and $p(x=1) = 0.6$. Let $p(y=0) = 0.3$ and $p(y=1) = 0.7$. Let the joint distribution be

$p(x,y)$	$y=0$	$y=1$
$x=0$	0.12	0.28
$x=1$	0.18	0.42

(1.2)

Q1. Are X and Y independent?

We note that the joint distribution $p(x,y)$ is equal to the product distribution $p(x)p(y)$ for all $x \in \mathcal{X}$ and $y \in \mathcal{Y}$. Thus, the two are independent.

Q2. What is the mutual information $I(X;Y)$?

Since the mutual information is the relative distance $D(p(x,y) \,||\, p(x)p(y))$, it is equal to zero.

Independence of Two Random Variables iff Zero Mutual Information

Now consider two different random variables U and V. We know that $I(U;V) = 0$.

Q3. Can we say that the two random variables U and V are independent with each other?

The answer is YES.

Proof: Let us use the definition of mutual information, i.e.,

$$\begin{aligned} I(U;V) &:= \sum_{u\in\mathcal{U}}\sum_{v\in\mathcal{V}} p(u,v)\log\frac{p(u,v)}{p(u)\,p(v)} \\ &= 0.0 \end{aligned} \tag{1.3}$$

It is not obvious to say that $p(u,v) = p(u)\,p(v)$ for all $u \in \mathcal{U}$ and $v \in \mathcal{V}$. Why?

Now let us try to use the Jenson's inequality:

$$\begin{aligned} -I(U;V) &= \sum_{u\in\mathcal{U}}\sum_{v\in\mathcal{V}} p(u,v)\log\frac{p(u)\,p(v)}{p(u,v)} \\ &\le \log\left(\sum_{u\in\mathcal{U}}\sum_{v\in\mathcal{V}} p(u,v)\frac{p(u)\,p(v)}{p(u,v)}\right) \\ &= \log\left(\sum_{u\in\mathcal{U}}\sum_{v\in\mathcal{V}} p(u)\,p(v)\right) \\ &= \log(1.0) \\ &= 0.0 \end{aligned} \tag{1.4}$$

For the second line, we have used the Jensen's inequality. Note that the logarithm is a *strictly concave* function. For the strictly concave functions, the only way the equality is met is the case when the function is evaluated at a single fixed point.

The proof for the equality part is left to the readers, not explicitly given in the text by Cover and Thomas. See Theorem 2.6.2 of Cover and Thomas. The equality at the second line holds only if $\frac{p(u)\,p(v)}{p(u,v)}$ is a constant for all for all $u \in \mathcal{U}$ and $v \in \mathcal{V}$. But the constant has to be 1.0, since otherwise the result in the fourth line cannot be met.

Thus, we have proved the following theorem.

Theorem. $I(U;V) = 0.0$ if and only if U and V are independent with each other.

Jensen's Inequality

Definition. Jensen's inequality is one of the fundamental tools very frequently used in information theory.

A function $f(x)$ is said to be *convex* over an interval (a,b), if for every $x_1,\ x_2 \in (a,b)$ and $0.0 \le \lambda \le 1.0$,

$$f\left(\lambda x_1 + (1-\lambda)x_2\right) \le \lambda f(x_1) + (1-\lambda) f(x_2). \tag{1.5}$$

It is said *strictly convex* if the equality holds only if $\lambda = 1.0$ or $\lambda = 0.0$.

A function f is concave if $-f$ is convex.

Examples of convex functions include x^2 and $|x|$.

Examples of concave functions include $\log(x), \sqrt{x}$ over $x \ge 0$.

Theorem. (Jensen's Inequality) If f is a convex function and X is a random variable,

$$\mathbb{E}f(X) \ge f(\mathbb{E}X). \tag{1.6}$$

In addition, if f is strictly convex, the equality in (1.6) implies that X is *degenerate* such that $X = \mathbb{E}(X)$ with probability 1.

Proof. The inequality part is easy. Just apply the definition of convex function (1.5) repeatedly. That's the approach taken in the proof of Theorem 2.6.2 by Cover and Thomas. But the equality part was not given there. It was left as a reader's exercise.

Proof of the equality part or the meaning of it. Take a look at the convex function again. When the function is *strictly convex*, the equality part shall be removed. That is, for a *strictly convex* function, $f\left(\lambda x_1 + (1-\lambda)x_2\right) < \lambda f(x_1) + (1-\lambda) f(x_2)$ for every $x_1,\ x_2 \in (a,b)$. The equality part holds only if $X = \mathbb{E}(X)$ is a constant, i.e., $\Pr\left(X = \mathbb{E}(X)\right) = 1.0$.

Typographical Correction IN HW#1

1. In HW#1, the last problem. θ_1 should be e_1. θ_2 should be e_2.

Example on Sufficient Statistics

Setting: Consider a sequence of coin tosses X_1, X_2, …, X_n, independent and identically distributed (i.i.d.) with $X_i \in \{0,1\}$ with an unknown parameter $\theta =: \Pr\{X_i = 1\}$.

Statement: Given the total number n of coin tosses, the number of 1s in *n*-tosses is a *sufficient statistic* for inference on θ.

Is the statement True or False?

Sketch of Proof:

Let $T(X_1, X_2, ..., X_n) := \sum_{i=1}^{n} X_i$.

To show that T is a sufficient statistic for θ, we need to show $I(\theta; X) = I(\theta; T)$.

It can be also proven by showing $\theta \to T \to X$ MC. This relation $\theta \to T \to X$ can be proven by showing that the following holds:

$$\Pr\left(X = x \mid T = t, \theta = c\right) = \Pr(X = x \mid T = t) \tag{1.7}$$

where $x := (x_1, x_2, ..., x_n)$, $x_i \in \{0,1\}$, $t \in \{0,1,2,...,n\}$ and $c \in (0,1]$.

There are two parts:

1. Show $\theta \to X \to T$

   To show this, we will show $\Pr\{T \mid X, \theta\} = \Pr(T \mid X)$

2. Show $\theta \to T \to X$

   To show this, we will show $\Pr\{X \mid T, \theta\} = \Pr(X \mid T)$

The first part: Show $\theta \to X \to T$

$$
\begin{aligned}
\Pr\{T=t \mid X=x\} &= \int_0^1 \Pr\{T=t, \theta \in (c, c+d\theta] \mid X=x\} d\theta \\
&= \int_0^1 \Pr\{T=t \mid \theta \in (c, c+d\theta], X=x\} \Pr\{\theta \in (c, c+d\theta] \mid X=x\} d\theta \\
&= \int_0^1 \Pr\left\{\sum_{i=1}^n x_i = t \mid \theta \in (c, c+d\theta], X=x\right\} \Pr\{\theta \in (c, c+d\theta] \mid X=x\} d\theta \\
&= \Pr\left\{\sum_{i=1}^n x_i = t \mid \theta \in (c, c+d\theta], X=x\right\} \int_0^1 \Pr\{\theta \in (c, c+d\theta] \mid X=x\} d\theta \qquad (1.8) \\
&= \Pr\left\{\sum_{i=1}^n x_i = t \mid \theta = c, X=x\right\} \\
&= \begin{cases} 1.0 & \text{if } t = \sum_{i=1}^n x_i \\ 0.0 & \text{otherwise} \end{cases}
\end{aligned}
$$

Q.E.D.

The second part: Show $\theta \to T \to X$:

First, note that

$$
\begin{aligned}
&\Pr\left\{X = (x_1, x_2, ..., x_n) \mid \theta = c\right\} \\
&\quad = \left(c^{x_1}(1-c)^{(1-x_1)}\right)\left(c^{x_2}(1-c)^{(1-x_2)}\right)\cdots\left(c^{x_n}(1-c)^{(1-x_n)}\right) \qquad (1.9) \\
&\quad = c^{\sum_{i=1}^n x_i}(1-c)^{n-\sum_{i=1}^n x_i}
\end{aligned}
$$

Second, note that

$$
\Pr\left\{X = \underbrace{(x_1 = 0, x_2 = 1, ..., x_n = 0)}_{t \text{ 1s}} \mid \theta = c\right\} = c^t(1-c)^{n-t} \qquad (1.10)
$$

Third, note that

$$\Pr\left\{X=\left(x_1,x_2,...,x_n\right)\mid\theta=c,T=t\right\}=\frac{\Pr\left\{X=\left(x_1,x_2,...,x_n\right),T=t\mid\theta=c\right\}}{\Pr\{T=t\mid\theta=c\}}$$

$$=\frac{\Pr\left\{X=\left(x_1,x_2,...,x_n\right),T=t\mid\theta=c\right\}}{\sum_x\Pr\{X=\left(x_1,x_2,...,x_n\right),T=t\mid\theta=c\}}$$

$$=\begin{cases}\dfrac{c^t(1-c)^{n-t}}{\binom{n}{t}c^t(1-c)^{n-t}} & \text{if } \sum_{i=1}^n x_i=t \\ 0.0 & \text{o.w.}\end{cases} \qquad (1.11)$$

$$=\begin{cases}\dfrac{1}{\binom{n}{t}} & \text{if } \sum_{i=1}^n x_i=t \\ 0.0 & \text{o.w.}\end{cases}$$

Fourth, we consider $\Pr\left\{X=\left(x_1,x_2,...,x_n\right)\mid T=t\right\}$ but this will result in the same. What matters the case where the sum satisfies $\sum_{i=1}^n x_i=t$. Thus, we focus on such event. Namely, we have

$$\Pr\left\{X=\underbrace{\left(x_1,x_2,...,x_n\right)}_{t\ 1\text{s}}\mid T=t\right\}=\int_0^1\Pr\left\{X=\underbrace{\left(x_1,x_2,...,x_n\right)}_{t\ 1\text{s}},\theta\in(c,c+d\theta]\mid T=t\right\}d\theta$$

$$=\int_0^1\Pr\left\{X=\underbrace{\left(x_1,x_2,...,x_n\right)}_{t\ 1\text{s}}\mid\theta\in(c,c+d\theta],T=t\right\}\Pr\left\{\theta\in(c,c+d\theta]\mid T=t\right\}d\theta$$

$$=\frac{1}{\binom{n}{t}}\int_0^1\Pr\left\{\theta\in(c,c+d\theta]\mid T=t\right\}d\theta \qquad (1.12)$$

$$=\frac{1}{\binom{n}{t}}$$

Q.E.D.