Homework 1

Zhong Yun 2016K8009915009

September 12, 2018

Theorem 1 *Information Gain* $Gain(X,Y) \geq 0$

Proof 1 The definition of Information Gain is as follows:

$$Gain(X,Y) = H(X) - H(X|Y) \qquad (1)$$

According to *Jensen Inequality*: if function f(x) is convex, g(x) is an arbitrary function about f(x), p(x)≥ 0, then

$$\frac{\int_a^b f(g(x))p(x)dx}{\int_a^b p(x)dx} \geq f\left(\frac{\int_a^b g(x)p(x)dx}{\int_a^b p(x)dx}\right)$$

Meanwhile, according to the definition of $Kullback-Leibler\ divergence$,

$$D(P\|Q) = \sum_{x\in X} p(x)log\left(\frac{p(x)}{q(x)}\right) = \int_{x\in X} p(x)log\left(\frac{p(x)}{q(x)}\right)dx$$

Among them,$p(x)$ indicates the probability of x, $q(x)$ indicates the probability density of x.
Then, we derive (1)

$$\begin{aligned}
Gain(X,Y) = H(X) - H(X|Y)& \\
&= -\sum_{x\in X} p(x)log(p(x)) + \sum_{y\in Y} p(y)\sum_{x\in X} p(x|y)log(p(x|y)) \\
&= \sum_{x\in X}\sum_{y\in Y} p(x,y)(\log(p(x)) - log(p(y)) + log(p(x,y))) \\
&= \sum_{x\in X}\sum_{y\in Y} p(x,y)log\left(\frac{p(x,y)}{p(x)p(y)}\right) \\
&= D(P(X,Y)\|P(X)P(Y))
\end{aligned}$$

Then we have

$$Gain(X,Y) = D(P(X,Y)\|P(X)P(Y))$$

Therefore, we only need to prove $Kullback-Leibler\ divergence\ formula \geq 0$. The proof is as follows:

$$\begin{aligned}
D(P\|Q) &= \int_{x\in X} p(x)log\left(\frac{p(x)}{q(x)}\right) \\
&= \int_{x\in X} -log\left(\frac{q(x)}{p(x)}\right)p(x)dx
\end{aligned}$$

Here, we regard $-log(x)$ as f(x), regard $\frac{q(x)}{p(x)}$ as g(x), and there is $\int_{x\in X} p(x)dx = 1$, according to

Jensen Inquality Formula,

$$\begin{aligned}\int_{x\in X} -log\left(\frac{q(x)}{p(x)}\right)p(x)dx \geq & -log\left(\int_{x\in X}\frac{q(x)}{p(x)}p(x)dx\right) \\ &= -log\left(\int_{x\in X} q(x)dx\right) \\ &= -log(1) \\ &= 0\end{aligned}$$

Then, we have $D(P\|Q) \geq 0$, therefore, $Gain(X,Y) \geq 0$, Therom 1 is proved.