Homework 10

Li Yunzhi

November 15, 2017

1 Question 1

We use Shannon Entropy to prove this property
In this proof, we use $P(y_i), (i = 1, 2, ..., n)$ to represent the probability distribution of random variable Y, $P(x_j), (j = 1, 2, ..., m)$ to represent the probability distribution of random variable X, $P(y_i|x_j), (i = 1, 2, ..., n, j = 1, 2, ..., m)$ to represent the conditional probability of $P(Y|X)$
First, the definition of Shannon Entropy is $H(Y) = \sum_{i=1}^{n} P(y_i) \log_2 \frac{1}{P(y_i)}$
And the information gain is $Gain(X, Y) = H(Y) - H(Y|X)$.
Then

$$
\begin{aligned}
Gain(X, Y) &= H(Y) - (H(Y|X)) \\
&= \sum_{i=1}^{n} P(y_i) \log_2 \frac{1}{P(y_i)} - \sum_{j=1}^{m} P(x_j) H(Y|x_j) \\
&= \sum_{i=1}^{n} \sum_{j=1}^{m} P(y_i|x_j) \log_2 \frac{1}{P(y_i)} - \sum_{j=1}^{m} P(x_j) \sum_{i=1}^{n} P(y_i|x_j) \log_2 \frac{1}{P(y_i|x_j)} \\
&= \sum_{j=1}^{m} P(x_j) \sum_{i=1}^{n} P(y_i|x_j) \log_2 P(y_i|x_j) - \sum_{i=1}^{n} \sum_{j=1}^{m} P(y_i|x_j) \log_2 P(y_i) \\
&= \sum_{i=1}^{n} \sum_{j=1}^{m} P(x_j) P(y_i|x_j) \log_2 P(y_i|x_j) - \sum_{i=1}^{n} \sum_{j=1}^{m} P(y_i|x_j) \log_2 P(y_i) \\
&= \sum_{i=1}^{n} \sum_{j=1}^{m} P(x_j) P(y_i|x_j) \log_2 \frac{P(y_i|x_j)}{P(y_i)} \\
&= \frac{1}{\ln 2} \sum_{i=1}^{n} \sum_{j=1}^{m} P(x_j) P(y_i|x_j) \ln \frac{P(y_i|x_j)}{P(y_i)} \\
&= -\frac{1}{\ln 2} \sum_{i=1}^{n} \sum_{j=1}^{m} P(x_j) P(y_i|x_j) \ln \frac{P(y_i)}{P(y_i|x_j)} \\
&\geqslant \frac{1}{\ln 2} \sum_{i=1}^{n} \sum_{j=1}^{m} P(x_j) P(y_i|x_j) (\frac{P(y_i)}{P(y_i|x_j)} - 1) \\
&= \frac{1}{\ln 2} \sum_{j=1}^{m} P(x_j) \sum_{i=1}^{n} (P(y_i) - P(y_i|x_j)) \\
&= \frac{1}{\ln 2} (1 - 1) \\
&= 0
\end{aligned}
$$

2 Question2

I don't think there exist some properties that make the information gain is always positive. In other words, there must exist some conditions that lead the information gain to be zero.
For example:
First, the definition of information entropy is $H(Y) : Y \in y \rightarrow \mathcal{R}$, the conditional information entropy is $H(Y|X) : Y, X \in y \rightarrow \mathcal{R}$
So if we know the distribution of vatiable Y, we can calculate its entropy. So, now let' see when the information gain is zero.
Suppose that D(Y) represent the distribution of random variable Y, and$D(Y|X_i)$ represent the distribution of random variable Y under the condition X_i.
Then

$$\begin{aligned} Gain(X,Y) &= H(D(Y)) - (H(D(Y|X))) \\ &= H(D(Y)) - \sum_{j=1}^{m} P(x_j)H(D(Y|x_j)) \end{aligned}$$

so if $\forall i \in [1,m]$, $D(Y) = D(Y|x_j)$
Then we have Gain(X,Y) = 0
So if the distribution of variable Y under the condition of X is different from its distribution in the whole scale, we can get a positive information gain.
We can think about it in another way, if the distribution of variable Y is the same in any different condition of X_i, of course we cannot gain more information than we observe it in the whole scale.