

# Notes on Entropy

Tan Yee Fan

2008 November 25

## 1 Entropy

Let  $X$  be a random variable that can take  $n_X$  values. Let the probability of occurrence of  $x \in X$  be  $P(X = x) = p(x)$ .

The *information* provided by observing  $X = x$  is defined to be  $I(x) = \log \frac{1}{p(x)} = -\log_2 p(x)$ . The unit of information is *bits*. From the above definition, we have the following properties:

- If  $p(x) = 1$ , then  $I(x) = 0$ .
- If  $0 \leq p(x) \leq 1$ , then  $I(x) \geq 0$ .
- If  $p(x) < p(x')$ , then  $I(x) > I(x')$ .

The *entropy* or *information content* of  $X$  is the expected information gained from observing its value, i.e.,  $H(X) = E_X(I(x)) = -\sum_{x \in X} p(x) \log_2 p(x)$ , where  $0 \log_2 0 = 0$ . Therefore, entropy measures the uncertainty of the random variable  $X$ . We have the following properties:

- The entropy  $H(X)$  is bounded by  $0 \leq H(X) \leq \log_2 n_X$ .
- $H(X) = 0$  if and only if  $p(x) = 1$  for some  $x \in X$ .
- $H(X) = \log_2 n_X$  if and only if  $p(x) = \frac{1}{n_X}$  for all  $x \in X$ .

For two variables  $X$  and  $Y$ , their *joint entropy* is defined as  $H(X, Y) = -\sum_{x \in X} \sum_{y \in Y} p(x, y) \log_2 p(x, y)$ . We can also compute the entropy of  $Y$  when  $X$  takes on the value  $x$ , denoted by  $H(Y|X = x)$ . Then the *conditional entropy* of  $Y$  given  $X$  is defined as  $H(Y|X) = E_X(H(Y|X = x)) = \sum_{x \in X} p(x) H(Y|X = x)$ , with the property that  $0 \leq H(Y|X) \leq H(Y)$ . Joint entropy and conditional entropy is related by  $H(X, Y) = H(Y|X) + H(X)$ .

## 2 Mutual Information

The *mutual information* between  $X$  and  $Y$  measures the difference in uncertainty for  $X$  before and after observing  $Y$ , defined as  $I(X, Y) = H(X) - H(X|Y) = \sum_{x \in X} \sum_{y \in Y} p(x, y) \log_2 \frac{p(x, y)}{p(x)p(y)}$ . Note that  $H(X) = I(X, X)$ , and we have the following properties:

- $I(X, Y) = I(Y, X)$ .
- $I(X, Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$ .
- $I(X, Y) \geq 0$ .

In the ID3 algorithm, suppose we want to use the attribute  $X$  to predict the class  $Y$ . The uncertainty of  $Y$  before observing  $X$  is  $H(Y)$ . The expected uncertainty of  $Y$  after observing  $X$  is  $H(Y|X)$ . Then the *information gain* from observing  $X$  is defined to be  $IG(X) = H(Y) - H(Y|X)$ . As it turns out, it is exactly the same as the mutual information between  $X$  and  $Y$ .

The *gain ratio* of  $X$  is defined to be  $GR(X) = \frac{IG(X)}{H(X)}$ . It is used in the C4.5 learning algorithm.

### 3 Kullback-Leibler Divergence

Let  $p(x)$  and  $q(x)$  be two different probability distributions on  $X$ . The *Kullback-Leibler divergence* between  $p(x)$  and  $q(x)$  is defined to be  $D_{KL}(p||q) = \sum_{x \in X} p(x) \log_2 \frac{p(x)}{q(x)}$ . We note the following:

- $D_{KL}(p||q) \neq D_{KL}(q||p)$  in general.
- $D_{KL}(p||q) \geq 0$ .
- $D_{KL}(p||q) = 0$  if and only if  $p(x)$  and  $q(x)$  are identical distributions.

The *Jensen-Shannon divergence* is defined as  $D_{JS}(p||q) = \frac{1}{2}D_{KL}(p||m) + \frac{1}{2}D_{KL}(q||m)$ , where  $m(x) = \frac{1}{2}p(x) + \frac{1}{2}q(x)$  for all  $x \in X$ .

### References

- [Haykin, 1998] Haykin, S. (1998). *Neural Networks: A Comprehensive Foundation*. Prentice Hall, second edition.
- [Russell and Norvig, 2002] Russell, S. and Norvig, P. (2002). *Artificial Intelligence: A Modern Approach*. Prentice Hall, second edition.