Brief Introduction to Entropy

1 Origin of Information Theory

The field of Information Theory originated with the, now classical, paper by Shanon and Weaver (1949): *The mathematical theory of communication*. University of Illinois Press, Urbana.

To introduce the basic ideas let us consider the following example:

Example 1 *Suppose that we toss a coin and we are interested in the question* "Will it come up heads?".

Suppose one knows already something about the coin, e.g. $P(H) = 0.9$*. Then, in betting on a coin toss, the bet will be on* H*. Suppose one bets $1.00.*

The expected value for the bet, calculated according to the usual formula for the expectation of a random variable (probability distribution) is

$$EV(bet) = 1 \times 0.9 + (-1) \times 0.1 = 0.9 - 0.1 = 0.8$$

This means that one should be willing to pay at most 0.2 *for advance information on the actual outcome. In general, if* p *is the probability (estimated, guessed) for* H *and the bet is on* H *then*

$$EV_{(p)}(bet) = p - 1(1 - p) = 2p - 1 \tag{1}$$

Therefore, $EV_{(p)}(bet) > 0 \; iff \; 2p - 1 > 0 \; (or, \; p > 1/2)$ *and the amount to pay to obtain information on the outcome of the coin toss is at most* $1-(2p-1) = 2(1-p)$.

For a fair coin, $p = 1/2$, $EV_{(1/2)}(bet) = 0$ *and one would be willing to pay up to $(1-0)=$1 for advance information. If* $A(p) = 2(1 - p)$ *then the graph of* A *is shown in* **??**.

That is, the smaller p *is, the more one should be willing to pay. Of the values of* p *one should really exclude* $p < 1/2$ *because for such values the expected value of the bet,* $EV_{(p)}(bet) < 0$ *(one would have to pay for the information more than the amount won by the bet).*

In information theory the ideas are similar, but the currency is not $; instead it is **bits**: one bit of information is enough to answer yes/no about which one has no idea-such as the flip of a fair coin.

In general, let a_i denote the possible answer to a question and $p_i = p(a_i)$ denote the probability of a_i, $i = 1, ..., n$, $0 \leq p_i \leq 1$, $\sum_{i=1}^{n} p_i = 1$. Then the entropy is defined as

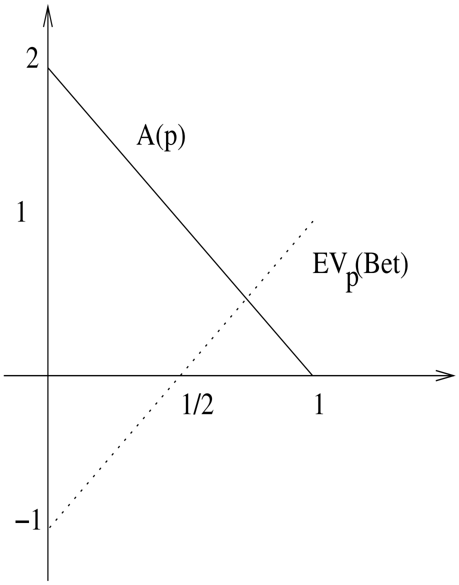

Figure 1: Amount willing to pay $A(p)$ and the expected value of the bet $EV_p(Bet)$ to obtain information about p

$$H(p_1, ..., p_n) = -\sum_{i=1} p_i \log_2 p_i \tag{2}$$

From probability theory, we see that $H(p_1, ..., p_n)$ is the expected value of the variable X when $p_i = P(X = -\log_2 p_i)$.

Example 2 *Calculate the entropy for the uniform distribution:*

1. $n = 2,\ p_1 = p_2 = \frac{1}{2}$

   $$H\left(\frac{1}{2}, \frac{1}{2}\right) = -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2} = -\log_2\frac{1}{2} = -\log_2 2^{-1} = 1\times\log_2 2 = 1$$

2. $p_1 = ... = p_n = \frac{1}{n}$

   $$H\left(\frac{1}{n}, ..., \frac{1}{n}\right) = -\sum_{i=1}^{n}\frac{1}{n}\log_2\frac{1}{n} = -\frac{1}{n}\log_2\frac{1}{n} - \frac{1}{n}\log_2\frac{1}{n} - ... - \frac{1}{n}\log_2\frac{1}{n} = -\frac{1}{n}(\log_2\frac{1}{n} + ... + \log_2\frac{1}{n}) = -n\times$$

   This is just the maximum value of $H(p_1, ..., p_n)$. Thus the uniform distribution has highest entropy and, as there are more vaues (a_i) and therefore $p_i's$, the entropy increases!

In connection with learning, e.g. using decision trees, one has to decide the correct classification for an example. Usually, attribute values need to be tested in order to answer this.

But, before any such test, the answer, an estimate of the probabilities to be in a class, is given by the proportion of positive(+) and negative (-) examples.

In the following we use the notation:

- E(examples): training set
- $E_+ = \{e \in E;\ e\ is\ a\ positive\ example\}$
- $E_- = \{e \in E;\ e\ is\ a\ negative\ example\}$
- $n = |E_-|;\ \ p = E_+$ such that $|E| = n + p$
- $H_E\left(\frac{p}{n+p}, \frac{n}{n+p}\right)$: estimate of the information in a correct answer

Example 3 *(***Saturday Morning***)*

E contains 14 examples of which 9 are "+" and 5 are "-". Therefore,

$$H_E\left(\frac{9}{14}, \frac{5}{14}\right) = -\frac{9}{14}\log_2\frac{9}{14} - \frac{5}{14}\log_2\frac{5}{14}$$

In general, if $n = 0$ or $p = 0$ $H_E(0,1) = H_E(1,0) = 0$

$H_E(0,1) = -0\log_2 0 - 1\log_2 1 = 0 - 0 = 0$
(l'Hospital rule to show that $\lim_{\epsilon\to 0}\epsilon\log_2\epsilon = 0$)

In general

$$f(p) = H(p, 1-p) = -p\log_2 p - (1-p)\log_2(1-p)$$

as a function of p has the graph shown in Figure 2.

It is easy to verify that

$$f(1) = 0,\ f(0) = 0$$

$$f'(p) = -\log_2 p - p\frac{1}{p}c - (-\log_2(1-p) - c) = -\log_2 p - c + \log_2(1-p) + c = \log_2\frac{1-p}{p}$$

$$f'(p) = 0$$

$$\log_2\frac{1-p}{p} = 0 \Leftrightarrow \frac{1-p}{p} = 1 \Leftrightarrow p = \frac{1}{2};\ f\prime\prime(p) < 0$$

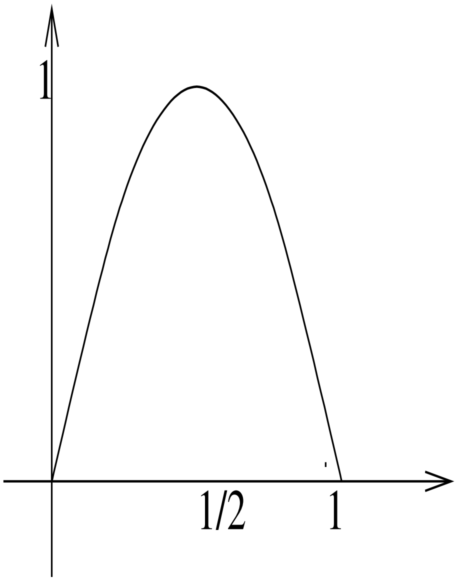

Figure 2: Amount willing to pay $A(p)$ and the expected value of the bet $EV_p(Bet)$ to obtain information about p

2 Information gain

The information gain measures the change of the entropy when the probability distribution changes.

For example,

- Initially: $p = \frac{1}{2}; q = \frac{1}{2}; \Rightarrow H(\frac{1}{2}, \frac{1}{2}) = 1$. One bit of information is needed ...

- Assume the probability changes to $p, 1-p$. Then $H(p, 1-p) = -p \log_2 p - (1-p) \log_2 (1-p)$, and therefore, the change in entropy, or the informatins gain IG, is

$$IG(p) = 1 + p \log_2 p + (1-p) \log_2 (1-p)$$

$IG(p) = 1 + p \log_2 p + (1-p) \log_2 (1-p)$
$IG(0) = IG(1) = 1$
$IG(\frac{1}{2}) = 1 + \frac{1}{2} \log_2 \frac{1}{2} + \frac{1}{2} \log_2 \frac{1}{2} = 0$
$IG'(p) = \log_2 p + p\frac{1}{p}c + (-1) \log_2 (1-p) - (1-p)\frac{1}{1-p}c = \log_2 \frac{p}{(1-p)}$
$IG'(p) = 0 \Rightarrow \frac{p}{(1-p)} = 1 \ \Rightarrow \ p = 1-p \ \Rightarrow \ p = \frac{1}{2}$
$IG''(p) = \frac{c}{p(1-p)} > 0 \ \Rightarrow \ p < 1$

3 Entropy, Information gain and learning

Question: How does all this relate to the problem of learning?

Answer: In attribute/feature selection: it helps selecting the next attribute to test.

Suppose that somehow a concept has been learned and that we want to classify a new instance. One way to do this is to test the value for each attribute of the new instance (against those representing the learned concepts).

Before testing, the probability distribution of the positive /negative examples determines the entropy (the information content): $H(\frac{p}{n+p}, \frac{n}{n+p})$.

Now let us look at one attribute, A. A fair question concerns the impact that this attribute makes on the ocncept, or **how much it contributes to the classification of an instance**. It turns out that it will be as much as it will actually change the probability of classification, that is, **as much as its information gain will show**.

A divides E into $E_{(1)}, ..., E_{(n)}$ according to the values $a_{(1)}, ..., a_{(n)}$ that A can take. Witout loss of generality we can assume $n = 2$ and therefore E splits into $E_{(1)}$, $E_{(2)}$.

Now each $E_{(1)}$, $E_{(2)}$ will contain some positive examples and some negative examples, which as before we can denote by E_i^+ and E_i^-, $i = 1, 2$ and $p_i = |E_i^+|$; $n_i = |E_i^-|$.

Therefore $|E_i| = p_i + n_i$ and $|E| = \sum_{i=1}^{2}(p_i + n_i)$.

If $A = a_1$ in the instance to be classified, then before any other test the information content of $A = a_1$ is

$$H_1 = H\left(\frac{p_1}{p_1 + n_1}, \frac{n_1}{p_1 + n_1}\right)$$

Similarlly, if $A = a_2$ the information content of $A = a_1$ is

$$H_2 = H\left(\frac{p_2}{p_2 + n_2}, \frac{n_2}{p_2 + n_2}\right)$$

Thus, the information content of a is either H_1or H_2 depending on whether $A = a_1$ or $A = a_2$.

Now, given the example set E some examples will correspond to $A = a_1$, others to $A = a_2$ and the probability that an example drawn at random from E has $A = a_i$, i=1,2 is given by

$$\frac{|E_i|}{|E|} = \frac{n_i + p_i}{n + p}$$

Therefore the average information contents of A is

$$EH_A = \sum_{i=1}^{2} \frac{n_i + p_i}{n + p} H\left(\frac{p_i}{p_i + n_i}, \frac{n_i}{p_i + n_i}\right)$$

Finally, the information gain of a is

$$IG(A) = H\left(\frac{p}{p+n}, \frac{n}{p+n}\right) - EH(A) = H\left(\frac{p}{p+n}, \frac{n}{p+n}\right) - \sum_{i=1}^{2}(\frac{n_i+p_i}{p+n})H\left(\frac{p_i}{p_i+n_i}, \frac{n_i}{p_i+n_i}\right)$$

The magnitude of $IG(A)$ conveys the effectiveness of the attribute A. More specifically, an attribute is effective if its information gain is high.

More on Entropy

I include in this section some important properties of the entropy which are useful in ML. Recall that,

Decomposability of the entropy

- $H(X) \geq 0$ with equality if and only if there exits i such that $p_i = 1$ (and all other P's are 0).
- Entropy is maximized if $p = (p_1, \ldots, p_n)$ is uniform. That is $p_i = \frac{1}{n}$
- If we define the concept of joint entropy as

  $$H(X,Y) = -\sum_{(x,y)\in\mathcal{A}_X,\mathcal{A}_Y} p(x, y \log p(x,y),$$

  where $\mathcal{A}_X$ denotes the domain of the variable X, then

  $$H(X,Y) = H(X) + H(Y) \iff P(x,y) = P(x)P(y), \text{ that is, iff } X \text{ and } Y \text{ are independent.}$$

- Recursive property: For $p = (p_1, p_2, \ldots, p_n)$

  $$H(p) = H(p_1, 1-p_1) + (1-p_1)H\left(\frac{p_2}{1-p_1}, \frac{p_3}{1-p_1}, \cdots, \frac{p_n}{1-p_1}\right)$$

  For example, $p = (1/2, 1/4, 1/4)$

  $$H(p) = -(1/2)\log_2(1/2) - 2(1/4)\log_2(1/4) = 1/2 + 1 = 3/2$$

  Also,

  $$H(1/2,1/2) + (1/2)H\left(\frac{1/4}{1/2}, \frac{1/4}{1/2}\right) = H(1/2,1/2) + (1/2)H\left(\frac{1}{2}, \frac{1}{2}\right) = 1 + (1/2)(1) = 3/2$$

- Further generalization of the recursive formula is as follows: $p = (p_1, p_2, \ldots, p_n)$. Divide p into $q_1 = p_1 + \ldots + p_m$, and $q_2 = p_{m+1} + \ldots + p_n$. Note that $q_i \geq 0$ and $q_1 + q_2 = 1$. Thus, we can talk about the entropy in the distribution $q = (q_1, q_2)$. Then we have

  $$H(p) = H(q) + q_1 H\left(\frac{p_1}{q_1}, \cdots, \frac{p_m}{q_1}\right) + q_2 H\left(\frac{p_{m+1}}{q_2}, \cdots, \frac{p_n}{q_2}\right)$$

- Relative Entropy (Kullback-Leibler Divergence) X and Y random variables with distributions P an d Q over the same domain $\mathcal{A}$. The KL divergence is defined as

  $$D_{KL}(P||Q) = \sum_x P(x) \frac{P(x)}{Q(x)}$$

  The Gibs inequality, VERY IMPORTANT in ML, is given by:

  $$D_{KL}(P||Q) \geq 0, \text{ with equality when } P = Q.$$

- Jensen's Inequality: f is a convex function if:

  $$f \text{ is a convex function if: } f(\lambda x_1 + (1-\lambda)x_2) \leq \lambda f(x_1) + (1-\lambda) f(x_2)$$

  $$\text{Jensen's inequality: for } f \text{ convex } E[f(X)] \geq f(E[X]),$$

  where $E[X]$ denotes the expected value of X.