Entropy notes

We are considering instances that are identified by feature values. For example, the feature can be color, and the feature values can be Red, Green, Blue. Probabilities are associated with the likelihood that an instance has a particular feature value. One can view the probability of a feature value as a measure of (un)certainty that an instance has that value. The goal of Entropy is to associate a measure of uncertainty with a partition of instances according to their feature values.

Let S be a partition of the instances into the subsets $s_1, \ldots, s_n$, according to a feature that can have n distinct values. Set

$$p_i = \mathsf{Prob}(\text{instance is in } s_i) \equiv \mathsf{Prob}(s_i)$$

The p_i values form a probability vector.

Definition: The vector $p = (p_1, p_2, \ldots, p_n)$ is a probability vector if $p_i \geq 0$ and $p_1 + \ldots + p_n = 1$. Let S be a partition.

$$\mathsf{Entropy}(S) = \quad \text{the measure of uncertainty about } S$$

If S is partitioned into $\{s_1, \ldots, s_n\}$ with $p_i = \mathsf{Prob}(s_i)$ then:

$$\mathsf{Entropy}(S) = \sum_{i=1}^{n} p_i \log(1/p_i)$$

We typically use base 2 for the log, and then the entropy is measured in bits. It can be shown that this form of Entropy is, up to a constant, the only one that satisfies a set of axioms that reflect the intuitive understanding of uncertainty. Examples are the following three axioms:

1. $\mathsf{Entropy}(S)$ is a continuous function of the p_i.

2. If $p_1 = p_2 = \ldots = p_n$ then $\mathsf{Entropy}(S)$ is an increasing function of n.

3. If a new partition T is formed from S by subdividing one of the subsets of S then $\mathsf{Entropy}(T) \geq \mathsf{Entropy}(S)$.

Example

Consider an experiment that produces instances with two possible feature values: a and b. We view these feature values to be the "target attributes". Running the experiment 10 times, we get a 6 times, and b 4 times. This enables us to estimate the following probabilities:

$$\text{Case 1:} \qquad \mathsf{Prob}(a) = 0.6, \quad \mathsf{Prob}(b) = 0.4.$$

The value 0.6 is the likelihood that the next experiment will produce a. Thus, we are not completely uncertain about the outcome. We are completely uncertain in Case 2, and there is no uncertainty in Case 3:

$$\text{Case 2:} \qquad \mathsf{Prob}(a) = 0.5, \quad \mathsf{Prob}(b) = 0.5.$$
$$\text{Case 3:} \qquad \mathsf{Prob}(a) = 1, \quad \mathsf{Prob}(b) = 0.$$

Computing the entropy for these cases:

$$\text{Case 1:} \qquad \mathsf{Entropy}(0.6, 0.4) = 0.6 \log(1/0.6) + 0.4 \log(1/0.4) = 0.970951$$
$$\text{Case 2:} \qquad \mathsf{Entropy}(0.5, 0.5) = 0.5 \log(1/0.5) + 0.5 \log(1/0.5) = 1$$
$$\text{Case 3:} \qquad \mathsf{Entropy}(1, 0) = 1 \log(1/1) + 0 \log(1/0) = 0$$

Conditional entropy

The conditional entropy of S assuming that it is know that t happened is:

$$\mathsf{Entropy}(S|t) = \sum_{i=1}^{n} p_i \log(1/p_i)$$

where $p_i = \mathsf{Prob}(s_i|t)$.
Let $S = \{s_1, \ldots, s_n\}$ and $T = \{t_1, \ldots, t_m\}$ be two partitions. The uncertainty about S given that we know the result on T is:

$$\mathsf{Entropy}(S|T) = \sum_{j=1}^{m} \mathsf{Prob}(t_j)\mathsf{Entropy}(S|t_j) = \sum_{j=1}^{m} \mathsf{Prob}(t_j) \sum_{i=1}^{n} p_{ij} \log(1/p_{ij})$$

where $p_{ij} = \mathsf{Prob}(s_i|t_j)$.

Example

Continuing with the previous example assume that we can see the color of each instance, and that the color is one of R,G,B. Here is the table of the experimental results that we observe:

	color	Target attribute
1	R	a
2	R	a
3	R	b
4	R	a
5	G	b
6	G	a
7	G	a
8	G	b
9	B	a
10	B	b

What is the entropy if it is known that the color is R? The new estimates for the probabilities and the entropy are:

$$\mathsf{Prob}(a|R) = 3/4, \quad \mathsf{Prob}(b|R) = 1/4. \quad \mathsf{Entropy}(3/4, 1/4) = 0.811278.$$

What is the entropy if it is known that the color is B?

$$\mathsf{Prob}(a|B) = 1/2, \quad \mathsf{Prob}(b|B) = 1/2. \quad \mathsf{Entropy}(1/2, 1/2) = 1.$$

Observe that the conditional entropy can be smaller or larger than the unconditional entropy. Now what is the entropy if we know the color? It is the weighted sum of the conditional entropies:

$$\begin{aligned}
\mathsf{Entropy}(\text{Target}|\text{color}) &= \mathsf{Prob}(R)\mathsf{Entropy}(\text{Target}|R) \\
&\quad + \mathsf{Prob}(G)\mathsf{Entropy}(\text{Target}|G) \\
&\quad + \mathsf{Prob}(B)\mathsf{Entropy}(\text{Target}|B) \\
&= 0.4 \times 0.811278 + 0.4 \times \mathsf{Entropy}(0.5, 0.5) + 0.2 \times \mathsf{Entropy}(0.5, 0.5) \\
&= 0.4 \times 0.811278 + 0.4 \times 1 + 0.2 \times 1 = 0.924511
\end{aligned}$$

It can be shown that this conditional entropy always decreases. Specifically, for any partitions S, T:

$$\mathsf{Entropy}(S) \geq \mathsf{Entropy}(S|T), \quad \mathsf{Gain}(S, T) \equiv \mathsf{Entropy}(S) - \mathsf{Entropy}(S|T) \geq 0$$

Application to feature selection

Let S be a partition of the data. Given several features A_j to choose from, we choose the feature A_j with the minimum $\mathsf{Entropy}(S|A_j)$. The "gain" in entropy when using a feature A_j is defined to be:

$$\mathsf{Gain}(S, A_j) = \mathsf{Entropy}(S) - \mathsf{Entropy}(S|A_j)$$

The feature A to be selected is the one that minimizes $\mathsf{Entropy}(S|A_j)$, or, equivalently, the one that maximizes $\mathsf{Gain}(S, A_j)$.

ID3

Input: A dataset S.

Output: A decision tree. Intermediate nodes in the tree are assigned an attribute (feature), with branches for each possible attribute value. Leaves in the tree are subsets of uniform values for the target-atribute.

1. If all the instances in S have the same value for the target attribute, return.

2. Compute Gain values for all candidate-attributes and select an attribute with the largest gain.

(Equivalently, compute the entropy conditioned by each one of the candidate-attributes and select one with the lowest value.)

3. Create an intermediate node for that attribute and compute its children (leaves).

4. Apply the algorithm (recursively) to each leaf.

Example:

Consider the following data:

	MOTOR	Wheels	Doors	Size	Efficiency
0	no	two	none	small	good
1	no	three	none	small	bad
2	yes	two	none	small	good
3	yes	four	two	small	bad
4	yes	four	three	medium	good
5	yes	four	four	medium	good
6	yes	four	four	large	bad

The partition of the 7 instances according to Efficiency is: (4,3). The entropy is: 0.985228.
The MOTOR feature has two values. The value "no" creates a partition of (1,1) with entropy of 1; the value "yes" creates a partition of (3,2) with an entropy if 0.970951. The probability of "no" MOTOR is 2/7, and the probability of "yes" MOTOR is 5/7. Therefore:

$$\mathsf{Entropy}(\text{Efficiency}|\text{MOTOR}) = 2/7 \times 1 + 5/7 \times 0.970951 = 0.97925$$

$$\mathsf{Gain}(\text{Efficiency}, \text{MOTOR}) = 0.985228 - 0.97925 = 0.00597758$$

For Wheels we have 3 partitions: { "two", (2,0) }, { "three", (0,1) }, { "four, (2,2) }

$$\mathsf{Entropy}(\text{Efficiency}|\text{Wheels}) = 2/7 \times 0 + 1/7 \times 0 + 4/7 \times 1 = 0.571429.$$

For Doors we have 4 partitions: { “none, (2,1) }, { “two”, (1,0) }, { “three, (1,0) }, { “four, (1,1) }

$$\mathsf{Entropy}(\text{Efficiency}|\text{Doors}) = 3/7 \times 0.918296 + 1/7 \times 0 + 1/7 \times 0 + 2/7 \times 1 = 0.67927$$

For Size we have 3 partitions: { “small”, (2,2) }, { “medium, (2,0) }, { “large”, (0,1) }

$$\mathsf{Entropy}(\text{Efficiency}|\text{Size}) = 4/7 \times 1 + 2/7 \times 0 + 1/7 \times 0 = 0.571429$$

Therefore we should choose Wheels or Size as our first attribute.

Choosing Wheels as the first attribute splits the data into three classes. Two of them need not be partitioned further. The one that does is composed of instances {3,4,5,6}. Proceeding as above we see that the next attribute is Size.