

2.3.5 Entropy Measures Homogeneity of Examples

In order to define the gain of information precisely, we begin by defining a measure commonly used in information theory, called entropy, that characterizes the (im)purity of an arbitrary collection of examples. Given a collection S , containing positive and negative examples of some target concept, the entropy of S relative to this boolean classification is:

$$\text{Entropy}(S) = -P_{\oplus} \log_2 P_{\oplus} - P_{\ominus} \log_2 P_{\ominus} \dots \dots \dots (2.1)$$

where p_{\oplus} , is the proportion of positive examples in S and p_{\ominus} , is the proportion of negative examples in S . In all calculations involving entropy we define $0 \log 0$ to be 0.

To illustrate, suppose S is a collection of **14** examples of some Boolean concept, including **9** positive and **5** negative examples (we adopt the notation **[9+, 5-]** to summarize such a sample of data). Then the entropy of S relative to this Boolean classification is:

$$\text{Entropy} ([9+, 5-]) = -(9/14) \log_2 (9/14) - (5/14) \log_2 (5/14) = 0.940 \dots \dots (2.2)$$

Notice that the entropy is **0** if all members of S belong to the same class. For example, if all members are positive ($p_{\oplus} = 1$), then p_{\ominus} , is **0**, and

$$\text{Entropy}(S) = -1 \cdot \log_2(1) - 0 \cdot \log_2 0 = -1 \cdot 0 - 0 \cdot \log_2 0 = 0.$$

Note the entropy is **1** when the collection contains an equal number of positive and negative examples. If the collection contains unequal numbers of positive and negative examples, the entropy is between **0** and **1**. **Figure (2.3)** shows the form of the entropy function relative to a Boolean classification, as p_{\oplus} , varies between **0** and **1**.

One interpretation of entropy from information theory is that it specifies the minimum number of bits of information needed to encode the classification of an arbitrary member of S (i.e., a member of S drawn at random with uniform probability).

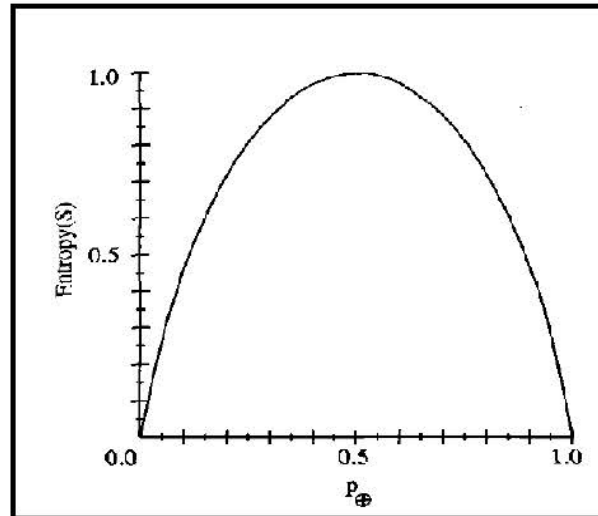


Figure (2.3) The entropy function relative to a Boolean classification, as the proportion, p_+ , of positive examples varies p_+ between 0 and 1.

For example, if $p_+=1$, the receiver knows the drawn example will be positive, so no message need be sent, and the entropy is zero. On the other hand, if $p_+=0.5$, one bit is required to indicate whether the drawn example is positive or negative. If $p_+=0.8$, then a collection of messages can be encoded using on average less than 1 bit per message by assigning shorter codes to collections of positive examples and longer codes to less likely negative examples. Thus, far we have discussed entropy in the special case where the target classification is boolean. More generally, if the target attribute can take on c different values, then the entropy of S relative to this c -wise classification is defined as:

$$Entropy(S) = \sum_{i=1}^c -p_i \log_2 p_i \dots \dots \dots (2.3)$$

where p_i is the proportion of S belonging to class i . Note the logarithm is still base 2 because entropy is a measure of the expected encoding length measured in *bits*. Note also that if the target attribute can take on c possible values, the entropy can be as large as $\log_2 c$.

2.3.6 Information Gain Measures the Expected Reduction in Entropy

Given entropy as a measure of the impurity in a collection of training examples, we can now define a measure of the effectiveness of an attribute in classifying the training data. The measure we will use, called *information gain*, is simply the expected reduction in entropy caused by partitioning the examples according to this attribute. More precisely, the information gain, $Gain(S, A)$ of an attribute A , relative to a collection of examples S , is defined as:

$$Gain(S, A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v) \dots \dots \dots (2.4)$$

where $Values(A)$ is the set of all possible values for attribute A , and S_v is the subset of S for which attribute A has value v (i.e., $S_v = \{s \in S \mid A(s) = v\}$). Note the first term in **Equation (2.4)** is just the entropy of the original collection S , and the second term is the expected value of the entropy after S is partitioned using attribute A . The expected entropy described by this second term is simply the sum of the entropies of each subset S_v , weighted by the fraction of examples $\frac{|S_v|}{|S|}$ that belong to S . $Gain(S, A)$ is therefore the expected reduction in entropy caused by knowing the value of attribute A . Put another way, $Gain(S, A)$ is the information provided about the *target & action value*, given the value of some other attribute A . The value of $Gain(S, A)$ is the number of bits saved when encoding the target value of an arbitrary member of S , by knowing the value of attribute A .

For example, suppose S is a collection of training-example days described by attributes including *Wind*, which can have the values *Weak* or *Strong*. As before, assume S is a collection containing 14 examples, [9+, 5-]. Of these 14 examples, suppose 6 of the positive and 2 of the negative examples have *Wind* = *Weak*, and the remainder have *Wind* = *Strong*. The information gain due to sorting the original 14 examples by the attribute *Wind* may then be calculated as:

$Values(Wind) = Weak, Strong$

$S = [9+, 5-]$

$S_{weak} \leftarrow [6+, 2-]$

$S_{strong} \leftarrow [3+, 3-]$

$$\begin{aligned}
 Gain(S, Wind) &= Entropy(S) - \sum_{v \in \{Weak, Strong\}} \frac{|S_v|}{|S|} Entropy(S_v) \\
 &= Entropy(S) - (8/14) Entropy(S_{weak}) - (6/14) Entropy(S_{strong}) \\
 &= 0.940 - (8/14) * 0.811 - (6/14) * 1.00 = 0.048
 \end{aligned}$$

Information gain is precisely the measure used by ID3 to select the best attribute at each step in growing the tree. The use of information gain to evaluate the relevance of attributes is summarized in **Figure (2.4)**. In this figure the information gain of two different attributes, *Humidity* and *Wind*, is computed in order to determine which is the better attribute for classifying the training examples shown in **Table (2.1)**.

Which attribute is the best classifier?

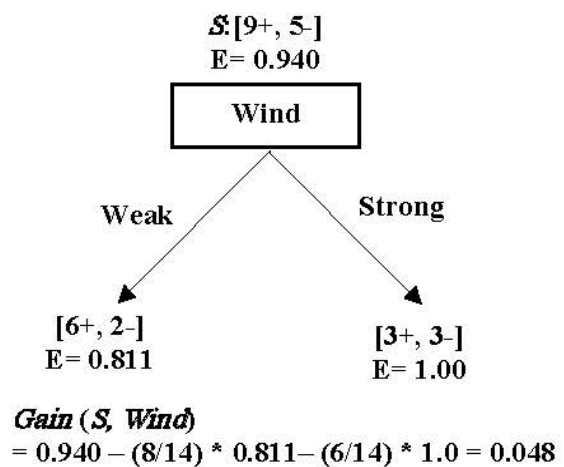
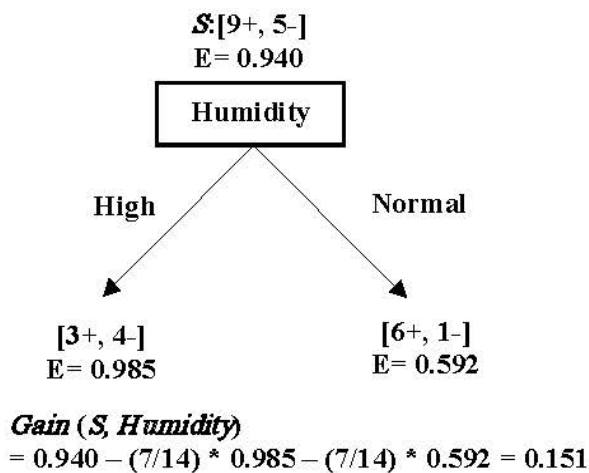


Figure (2.4) *Humidity* provides greater information gain than *Wind*, relative to the target classification. Here, E stands for entropy and S for the original collection of examples. Given an initial collection S of 9 positive and 5 negative examples, $[9+, 5-]$, sorting these by their *Humidity* produces collections of $[3+, 4-]$ (*Humidity*= High) and $[6+, 1-]$ (*Humidity*=Normal). The information gained by this partitioning is **0.151**, compared to a gain of only **0.048** for the attribute *Wind*.

2.3.7 An Illustrative Example of Decision Tree:

To illustrate the operation of ID3, consider the learning task represented by the training examples of **Table (2.1)**. Here the target attribute *PlayTennis*, which can have values *yes* or *no* for different Saturday mornings, is to be predicted based on other attributes of the morning in question. Consider the first step through the algorithm, in which the topmost node of the decision tree is created. Which attribute should be tested first in the tree? ID3 determines the information gain for each candidate attribute (i.e., *Outlook*, *Temperature*, *Humidity*, and *Wind*), then selects the one with highest information gain. The computation of information gain for two of these attributes is shown in **Figure (2.4)**.

TABLE (2.1) Training examples for the target concept *PlayTennis*

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

The information gain values for all four attributes are:

$$\text{Gain}(S, \text{Outlook}) = 0.246$$

$$\text{Gain}(S, \text{Humidity}) = 0.151$$

$$\text{Gain}(S, \text{Wind}) = 0.048$$

$$\text{Gain}(S, \text{Temperature}) = 0.029$$

where S denotes the collection of training examples from **Table (2.1)**.

According to the information gain measure, the **Outlook** attribute provides the best prediction of the target attribute, **PlayTennis**, over the training examples. Therefore, **Outlook** is selected as the decision attribute for the root node, and branches are created below the root for each of its possible values (i.e., **Sunny**, **Overcast**, and **Rain**).

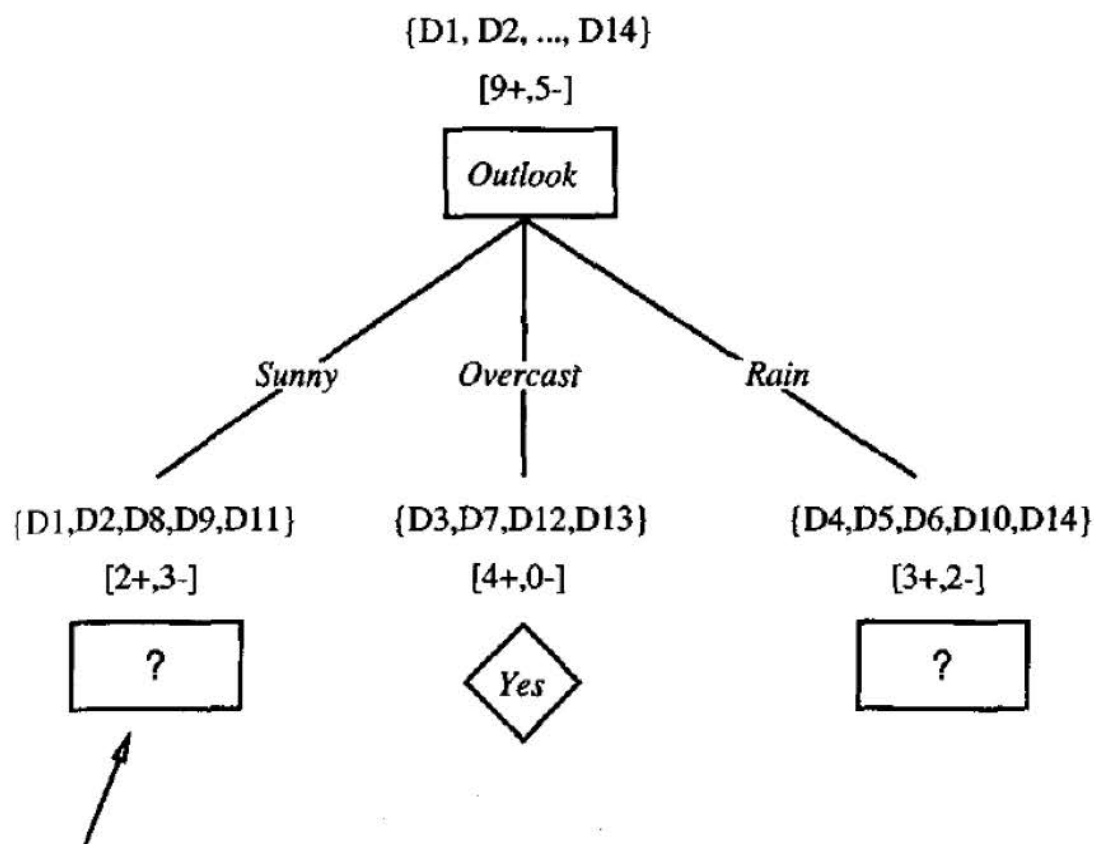
The resulting partial decision tree is shown in **Figure (2.4)**, along with the training examples sorted to each new descendant node.

Note that every example for which **Outlook** = **Overcast** is also a positive example of **PlayTennis**. Therefore, this node of the tree becomes a leaf node with the classification **PlayTennis** = **Yes**. In contrast, the descendants corresponding to **Outlook** = **Sunny** and **Outlook** = **Rain** still have nonzero entropy, and the decision tree will be further elaborated below these nodes.

The process of selecting a new attribute and partitioning the training examples is now repeated for each nonterminal descendant node, this time using only the training examples associated with that node. Attributes that have been incorporated higher in the tree are excluded, so that any given attribute can appear at most once along any path through the tree. This process continues for each new leaf node until either of two conditions is met:

- (1) every attribute has already been included along this path through the tree;
- (2) the training examples associated with this leaf node all have the same target attribute value (i.e., their entropy is zero).

Figure (2.4) illustrates the computations of information gain for the next step in growing the decision tree. The final decision tree learned by ID3 from the 14 training examples of Table (2.1) is shown in Figure (2.1).



$$S_{\text{sunny}} = \{D1, D2, D8, D9, D11\}$$

$$\text{Gain}(S_{\text{sunny}}, \text{Humidity}) = 0.970 - (3/5) * 0.0 - (2/5) * 0.0 = \mathbf{0.970}$$

$$\text{Gain}(S_{\text{sunny}}, \text{Temperature}) = 0.970 - (2/5) * 0.0 - (2/5) * 1.0 - (1/5) * 0.0 = \mathbf{0.570}$$

$$\text{Gain}(S_{\text{sunny}}, \text{Wind}) = 0.970 - (2/5) * 1.0 - (3/5) * 0.918 = \mathbf{0.019}$$

Figure (2.4) The partially learned decision tree resulting from the first step of ID3. The training examples are sorted to the corresponding descendant nodes. The **Overcast** descendant has only positive examples and therefore becomes a leaf node with classification **Yes**. The other two nodes will be further expanded, by selecting the attribute with highest information gain relative to the new subsets of examples.