|  |  |  |
| --- | --- | --- |
|  |  |  |
|  |  |  |
| https://www.saedsayad.com/images/Decision_Tree_1.png |  |  |
|  |  |  |
|  |  |  |
| **Entropy** |  |  |
| A decision tree is built top-down from a root node and involves partitioning the data into subsets that contain instances with similar values (homogenous). ID3 algorithm uses entropy to calculate the homogeneity of a sample. If the sample is completely homogeneous the entropy is zero and if the sample is an equally divided it has entropy of one. |  |  |
| https://www.saedsayad.com/images/Entropy.png |  |  |
|  |  |  |
| To build a decision tree, we need to calculate two types of entropy using frequency tables as follows: |  |  |
|  |  |  |
| a) Entropy using the frequency table of one attribute: |  |  |
| https://www.saedsayad.com/images/Entropy_3.png |  |  |
| b) Entropy using the frequency table of two attributes: |  |  |
| https://www.saedsayad.com/images/Entropy_2.png |  |  |

|  |  |  |
| --- | --- | --- |
| **Information Gain** |  |  |
| The information gain is based on the decrease in entropy after a dataset is split on an attribute. Constructing a decision tree is all about finding attribute that returns the highest information gain (i.e., the most homogeneous branches). |  |  |
|  |  |  |
| *Step 1*: Calculate entropy of the target. |  |  |
| https://www.saedsayad.com/images/Entropy_target.png |  |  |
| *Step 2*: The dataset is then split on the different attributes. The entropy for each branch is calculated. Then it is added proportionally, to get total entropy for the split. The resulting entropy is subtracted from the entropy before the split. The result is the Information Gain, or decrease in entropy. |  |  |
| https://www.saedsayad.com/images/Entropy_attributes.png |  |  |
| https://www.saedsayad.com/images/Entropy_gain.png |  |  |
| *Step 3*: Choose attribute with the largest information gain as the decision node, divide the dataset by its branches and repeat the same process on every branch. |  |  |
| https://www.saedsayad.com/images/Entropy_attribute_best.png |  |  |
| https://www.saedsayad.com/images/decision_tree_slices.png |  |  |
| *Step 4a*: A branch with entropy of 0 is a leaf node. |  |  |
| https://www.saedsayad.com/images/Entropy_overcast.png |  |  |
| *Step 4b*: A branch with entropy more than 0 needs further splitting. |  |  |
| https://www.saedsayad.com/images/Entropy_Sunny.png |  |  |
| *Step 5*: The ID3 algorithm is run recursively on the non-leaf branches, until all data is classified. |  |  |