Decision Trees Unplugged: Solutions & Data

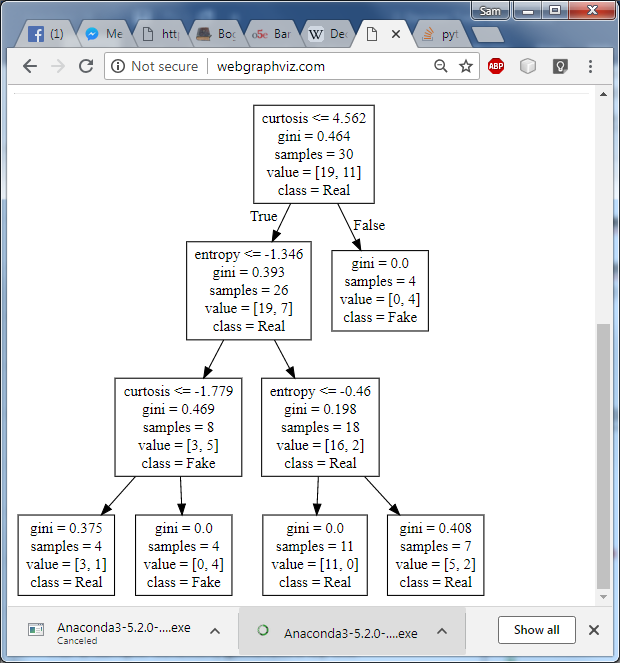
© Sam Scott, Mohawk College, 2019

# The Configuration

An sklearn tree.DecisionTreeClassifier object was created with the min\_samples\_leaf parameter set to 4. Then it was fitted to the data shown on the original handout.

# The Tree (Visual)

The raw GraphViz output (shown on the next page) was pasted into the text window at webgraphviz.com to produce the graph below. The “gini” field is a measure of purity at each node. The “value” field shows how many samples fell into each class.



# The Tree (Raw Graphviz Output)

digraph Tree {

node [shape=box] ;

0 [label="curtosis <= 4.562\ngini = 0.464\nsamples = 30\nvalue = [19, 11]\nclass = Real"] ;

1 [label="entropy <= -1.346\ngini = 0.393\nsamples = 26\nvalue = [19, 7]\nclass = Real"] ;

0 -> 1 [labeldistance=2.5, labelangle=45, headlabel="True"] ;

2 [label="curtosis <= -1.779\ngini = 0.469\nsamples = 8\nvalue = [3, 5]\nclass = Fake"] ;

1 -> 2 ;

3 [label="gini = 0.375\nsamples = 4\nvalue = [3, 1]\nclass = Real"] ;

2 -> 3 ;

4 [label="gini = 0.0\nsamples = 4\nvalue = [0, 4]\nclass = Fake"] ;

2 -> 4 ;

5 [label="entropy <= -0.46\ngini = 0.198\nsamples = 18\nvalue = [16, 2]\nclass = Real"] ;

1 -> 5 ;

6 [label="gini = 0.0\nsamples = 11\nvalue = [11, 0]\nclass = Real"] ;

5 -> 6 ;

7 [label="gini = 0.408\nsamples = 7\nvalue = [5, 2]\nclass = Real"] ;

5 -> 7 ;

8 [label="gini = 0.0\nsamples = 4\nvalue = [0, 4]\nclass = Fake"] ;

0 -> 8 [labeldistance=2.5, labelangle=-45, headlabel="False"] ;

}

# Iris Raw Data

Left table is sorted by Sepal Length, Right table sorted by Sepal Width

|  |  |  |
| --- | --- | --- |
| **Sepal Length** | **Sepal Width** | **Classification** |
| 4.3 | 3 | Setosa |
| 4.6 | 3.1 | Setosa |
| 4.7 | 3.2 | Setosa |
| 4.7 | 3.2 | Setosa |
| 4.8 | 3.1 | Setosa |
| 4.8 | 3.4 | Setosa |
| 4.8 | 3.4 | Setosa |
| 5 | 3 | Setosa |
| 5 | 3.2 | Setosa |
| 5.1 | 3.5 | Setosa |
| 5.1 | 3.3 | Setosa |
| 5.1 | 3.7 | Setosa |
| 5.1 | 3.8 | Setosa |
| 5.2 | 3.5 | Setosa |
| 5.2 | 3.4 | Setosa |
| 5.4 | 3 | Versicolor |
| 5.4 | 3.9 | Setosa |
| 5.4 | 3.9 | Setosa |
| 5.7 | 2.8 | Versicolor |
| 5.7 | 2.6 | Versicolor |
| 5.8 | 4 | Setosa |
| 5.8 | 2.6 | Versicolor |
| 5.9 | 3 | Versicolor |
| 6 | 2.7 | Versicolor |
| 6 | 2.2 | Versicolor |
| 6.1 | 2.8 | Versicolor |
| 6.3 | 3.3 | Versicolor |
| 6.4 | 2.9 | Versicolor |
| 6.7 | 3 | Versicolor |
| 6.8 | 2.8 | Versicolor |

|  |  |  |
| --- | --- | --- |
| **Sepal Length** | **Sepal Width** | **Classification** |
| 6 | 2.2 | Versicolor |
| 5.7 | 2.6 | Versicolor |
| 5.8 | 2.6 | Versicolor |
| 6 | 2.7 | Versicolor |
| 5.7 | 2.8 | Versicolor |
| 6.1 | 2.8 | Versicolor |
| 6.8 | 2.8 | Versicolor |
| 6.4 | 2.9 | Versicolor |
| 4.3 | 3 | Setosa |
| 5 | 3 | Setosa |
| 5.4 | 3 | Versicolor |
| 5.9 | 3 | Versicolor |
| 6.7 | 3 | Versicolor |
| 4.6 | 3.1 | Setosa |
| 4.8 | 3.1 | Setosa |
| 4.7 | 3.2 | Setosa |
| 4.7 | 3.2 | Setosa |
| 5 | 3.2 | Setosa |
| 5.1 | 3.3 | Setosa |
| 6.3 | 3.3 | Versicolor |
| 4.8 | 3.4 | Setosa |
| 4.8 | 3.4 | Setosa |
| 5.2 | 3.4 | Setosa |
| 5.1 | 3.5 | Setosa |
| 5.2 | 3.5 | Setosa |
| 5.1 | 3.7 | Setosa |
| 5.1 | 3.8 | Setosa |
| 5.4 | 3.9 | Setosa |
| 5.4 | 3.9 | Setosa |
| 5.8 | 4 | Setosa |

# Banknotes Example

Left table is sorted by Curtosis, right table sorted by Entropy

|  |  |  |
| --- | --- | --- |
| **Curtosis** | **Entropy** | **Classification** |
| -4.413 | -4.0211 | Real |
| -4.1722 | -4.7582 | Real |
| -4.1594 | -1.9379 | Fake |
| -3.3034 | -1.0509 | Real |
| -3.2846 | -1.1608 | Real |
| -3.2794 | -1.2009 | Real |
| -3.1123 | -2.7164 | Real |
| -2.9024 | -1.0379 | Real |
| -2.6848 | -0.92544 | Real |
| -2.6256 | -1.0341 | Real |
| -2.4774 | -0.50648 | Real |
| -1.8785 | 1.3258 | Fake |
| -0.44499 | -1.4905 | Fake |
| -0.0834 | -6.4172 | Fake |
| 0.16594 | -4.5396 | Fake |
| 0.20792 | 0.33662 | Real |
| 0.52581 | -7.0107 | Fake |
| 0.65005 | -0.92544 | Real |
| 0.77344 | 1.2095 | Real |
| 0.9885 | -0.87371 | Real |
| 1.5454 | -0.26079 | Real |
| 1.7785 | -0.47156 | Real |
| 1.9833 | -0.44829 | Fake |
| 2.0416 | 1.1319 | Real |
| 2.1341 | 0.3211 | Real |
| 3.0895 | -0.9849 | Real |
| 6.0344 | -0.20777 | Fake |
| 6.2169 | -0.62285 | Fake |
| 8.6521 | 1.8198 | Fake |
| 10.2184 | -1.0043 | Fake |

|  |  |  |
| --- | --- | --- |
| **Curtosis** | **Entropy** | **Classification** |
| 0.52581 | -7.0107 | Fake |
| -0.0834 | -6.4172 | Fake |
| -4.1722 | -4.7582 | Real |
| 0.16594 | -4.5396 | Fake |
| -4.413 | -4.0211 | Real |
| -3.1123 | -2.7164 | Real |
| -4.1594 | -1.9379 | Fake |
| -0.44499 | -1.4905 | Fake |
| -3.2794 | -1.2009 | Real |
| -3.2846 | -1.1608 | Real |
| -3.3034 | -1.0509 | Real |
| -2.9024 | -1.0379 | Real |
| -2.6256 | -1.0341 | Real |
| 10.2184 | -1.0043 | Fake |
| 3.0895 | -0.9849 | Real |
| -2.6848 | -0.92544 | Real |
| 0.65005 | -0.92544 | Real |
| 0.9885 | -0.87371 | Real |
| 6.2169 | -0.62285 | Fake |
| -2.4774 | -0.50648 | Real |
| 1.7785 | -0.47156 | Real |
| 1.9833 | -0.44829 | Fake |
| 1.5454 | -0.26079 | Real |
| 6.0344 | -0.20777 | Fake |
| 2.1341 | 0.3211 | Real |
| 0.20792 | 0.33662 | Real |
| 2.0416 | 1.1319 | Real |
| 0.77344 | 1.2095 | Real |
| -1.8785 | 1.3258 | Fake |
| 8.6521 | 1.8198 | Fake |