MACHINE LEARNING WITH PYTHON

DECISION TREES

Themistoklis Diamantopoulos

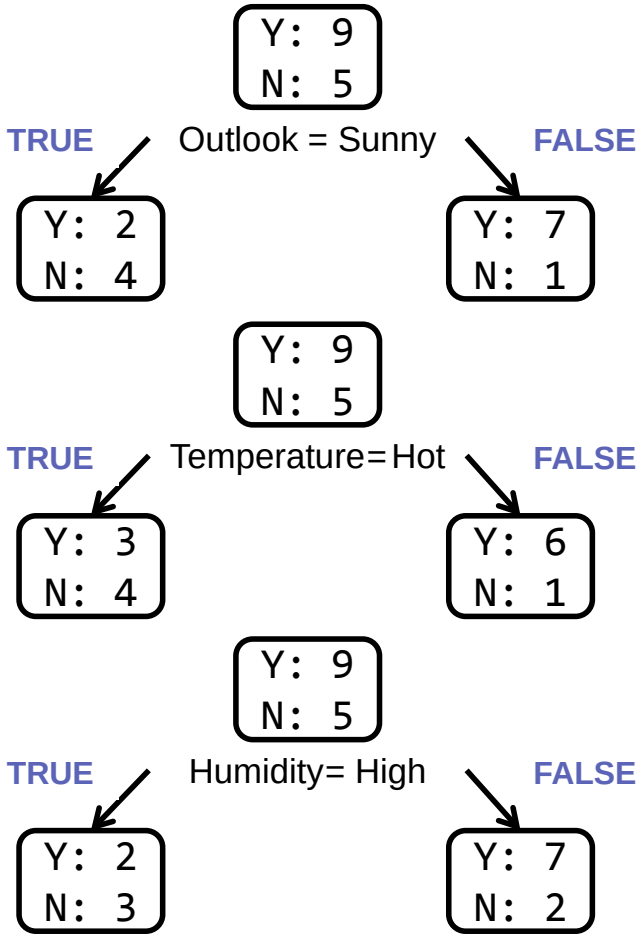
Categorical Classification

- Decide whether to play tennis based on outlook, temperature, and humidity

Outlook	Temperature	Humidity	Play
Sunny	Hot	High	No
Sunny	Hot	Low	No
Rainy	Hot	Low	Yes
Rainy	Cool	High	Yes
Rainy	Cool	Low	Yes
Rainy	Hot	Low	No
Rainy	Cool	Low	Yes
Sunny	Hot	High	No
Sunny	Cool	Low	Yes
Rainy	Hot	Low	Yes
Sunny	Cool	Low	Yes
Rainy	Hot	High	Yes
Rainy	Cool	Low	Yes
Sunny	Cool	High	No

Decision Tree Learning

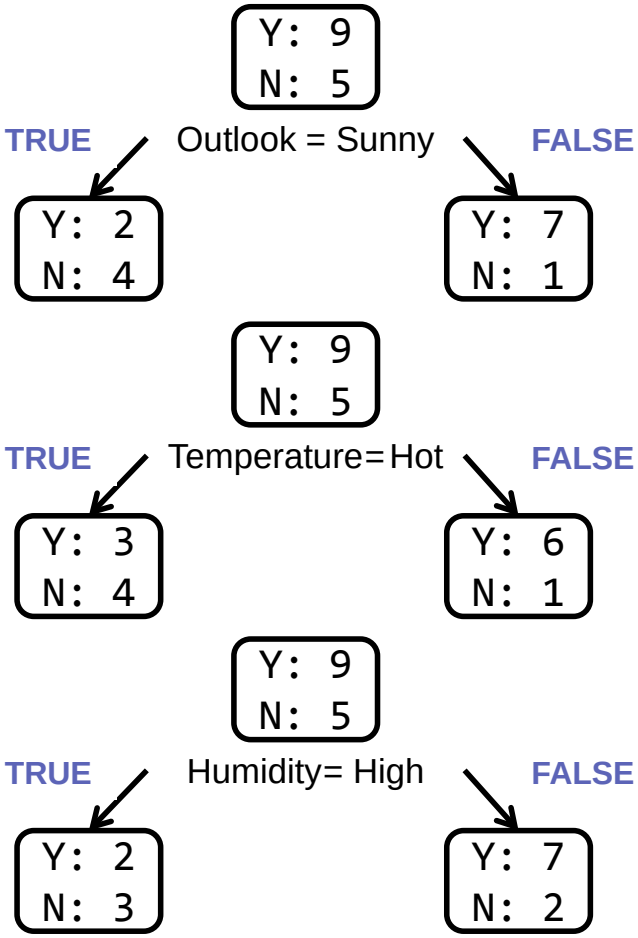
- Decide where to split



Outlook	Temperature	Humidity	Play
Sunny	Hot	High	No
Sunny	Hot	Low	No
Rainy	Hot	Low	Yes
Rainy	Cool	High	Yes
Rainy	Cool	Low	Yes
Rainy	Hot	Low	No
Rainy	Cool	Low	Yes
Sunny	Hot	High	No
Sunny	Cool	Low	Yes
Rainy	Hot	Low	Yes
Sunny	Cool	Low	Yes
Rainy	Hot	High	Yes
Rainy	Cool	Low	Yes
Sunny	Cool	High	No

Splitting Criteria

- Decide where to split



GINI Index

$$GINI(t) = 1 - \sum_j [p(j \mid t)]^2$$

$$GINI_{split} = \sum_{i=1}^{k} \frac{n_i}{n} GINI(i)$$

Information Gain

$$Entropy(t) = -\sum_j p(j \mid t) \log p(j \mid t)$$

$$GAIN_{split} = Entropy(p) - \left(\sum_{i=1}^{k} \frac{n_i}{n} Entropy(i) \right)$$

Example using GINI Index

GINI(Sunny) = 1 – (4/6)² – (2/6)² = 0.444
GINI(Rainy) = 1 – (1/8)² – (7/8)² = 0.219
$GINI_{Outlook}$ = 0.315 ⟸ **BEST SPLIT**
…
$GINI_{Temperature}$ = 0.367
…
$GINI_{Humidity}$ = 0.394

Feature Representation

- Required for applying algorithms in scikit-learn

Outlook	Temperature	Humidity	Play
0	0	0	0
0	0	1	0
1	0	1	1
1	1	0	1
1	1	1	1
1	0	1	0
1	1	1	1
0	0	0	0
0	1	1	1
1	0	1	1
0	1	1	1
1	0	0	1
1	1	1	1
0	1	0	0

Decision Tree Building