Example of Decision Tree Construction

Péter Jeszenszky

March 13, 2019

Entropy, Gini-index, Classification Error (1)

Let $\mathbf{p} = (p_1, \ldots, p_n) \in \mathbb{R}^n$, such that $0 \leq p_i \leq 1$, where $i = 1, \ldots, n$, és $\sum_{i=1}^{n} p_i = 1$. Let's define the following quantities:

$$H(\mathbf{p}) = -\sum_{i=1}^{n} p_i \log_2 p_i \qquad \text{(entropy)}$$

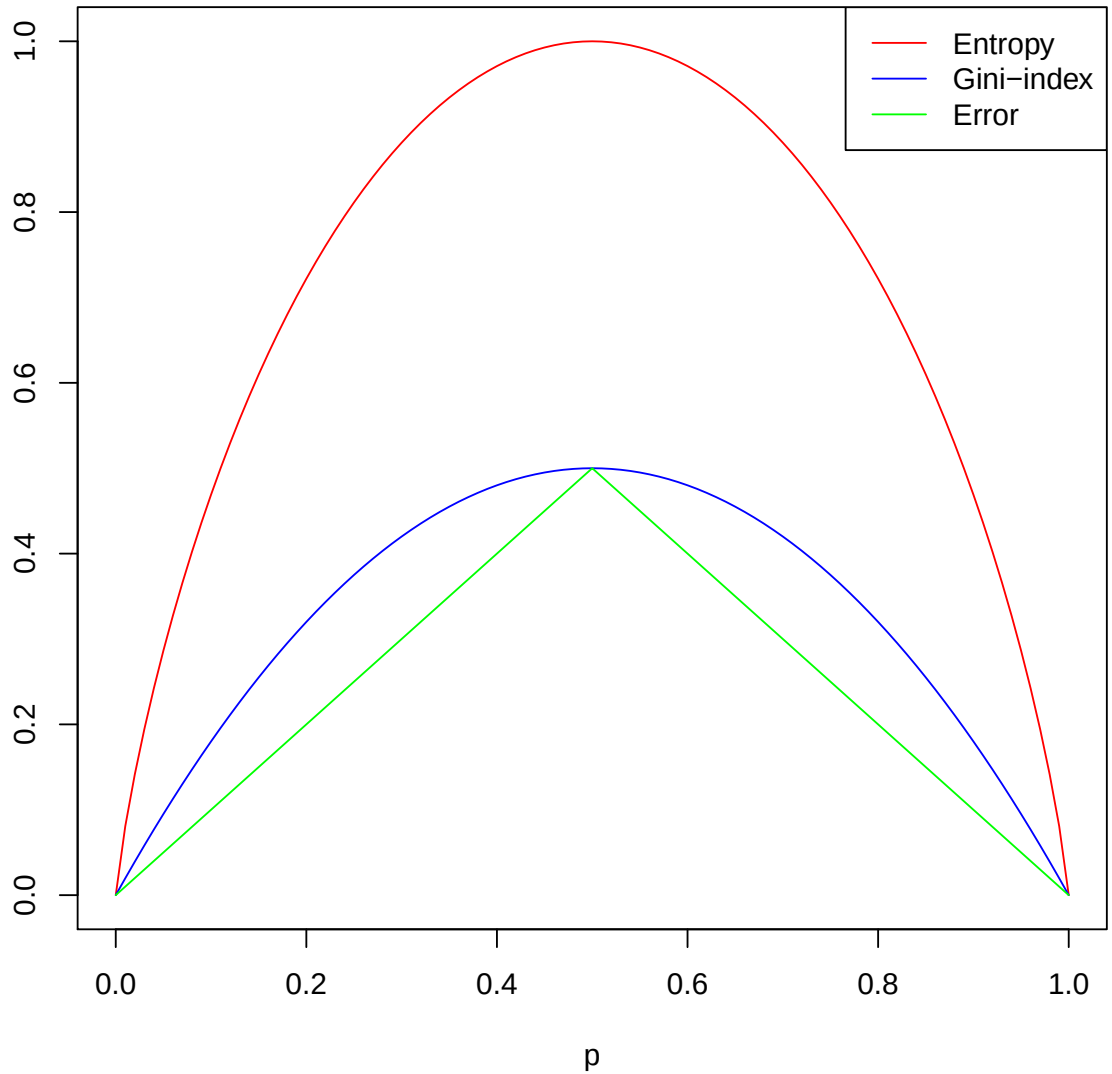
$$G(\mathbf{p}) = 1 - \sum_{i=1}^{n} p_i^2 \qquad \text{(Gini-index)}$$

$$E(\mathbf{p}) = 1 - \max_{1 \leq i \leq n} p_i \qquad \text{(classification error)}$$

Note

When $n = 2$, it is sufficient to specify only the value of p_1, since $p_2 = 1 - p_1$.

Entropy, Gini-index, Classification Error (2)



Gain

Let

$$\Delta_I = I(v) - \sum_{j=1}^{k} \frac{N(v_j)}{N} I(v_j),$$

where $I(\cdot)$ is any of the entropy, the Gini-index or the classification error, N is the number of records that belong to the parent node v, $N(v_j)$ is the number of records that belong to the jth child node ($j = 1, \ldots, k$).
Select the attribute for testing that maximizes this quantity!
Since the value of $I(v)$ is the same for each attribute, the weighted average of the entropy, the Gini-index or the classification error must be minimized, that are denoted by $\overline{H}$, $\overline{G}$ and $\overline{E}$ on the figures.

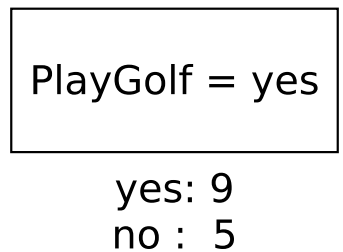
Note

When $I(\cdot)$ is the entropy, the above quantity is called **information gain**.

Data

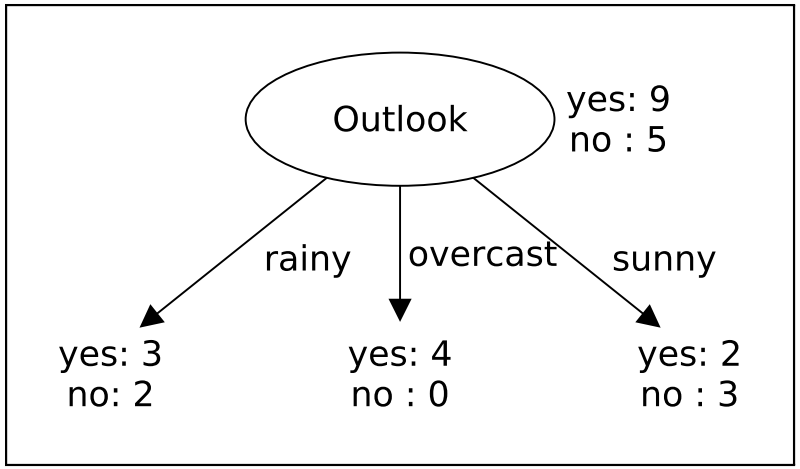
Day	Outlook	Temperature	Humidity	Wind	PlayGolf
1	sunny	hot	high	weak	no
2	sunny	hot	high	strong	no
3	overcast	hot	high	weak	yes
4	rainy	mild	high	weak	yes
5	rainy	cool	normal	weak	yes
6	rainy	cool	normal	strong	no
7	overcast	cool	normal	strong	yes
8	sunny	mild	high	weak	no
9	sunny	cool	normal	weak	yes
10	rainy	mild	normal	weak	yes
11	sunny	mild	normal	strong	yes
12	overcast	mild	high	strong	yes
13	overcast	hot	normal	weak	yes
14	rainy	mild	high	strong	no

Decision Tree (0)

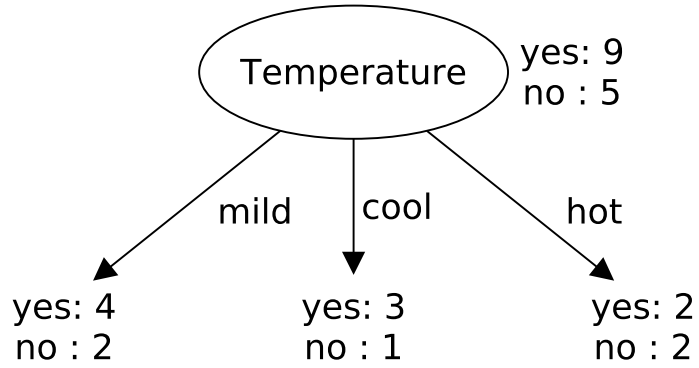


$$H = 0.94, G = 0.459, E = 0.357$$

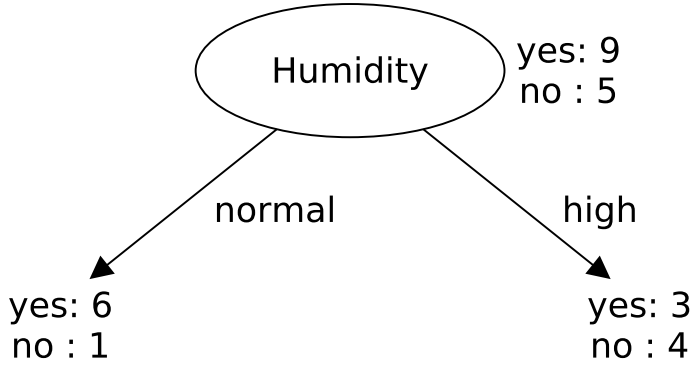
Which attribute to test at the root?



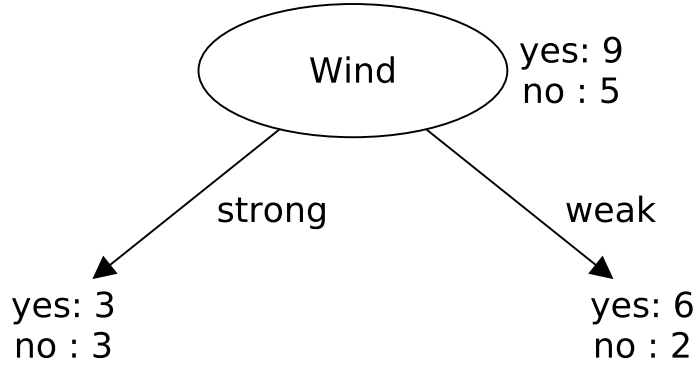
(a) $\overline{H} = 0.694, \overline{G} = 0.343, \overline{E} = 0.286$



(b) $\overline{H} = 0.911, \overline{G} = 0.440, \overline{E} = 0.357$

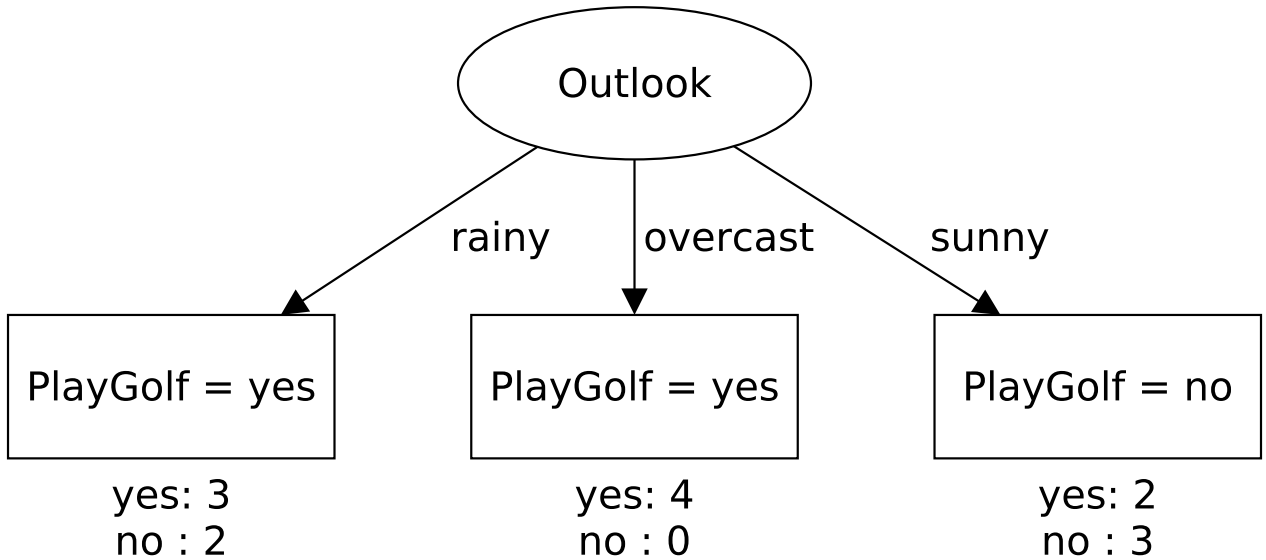


(c) $\overline{H} = 0.788, \overline{G} = 0.367, \overline{E} = 0.286$

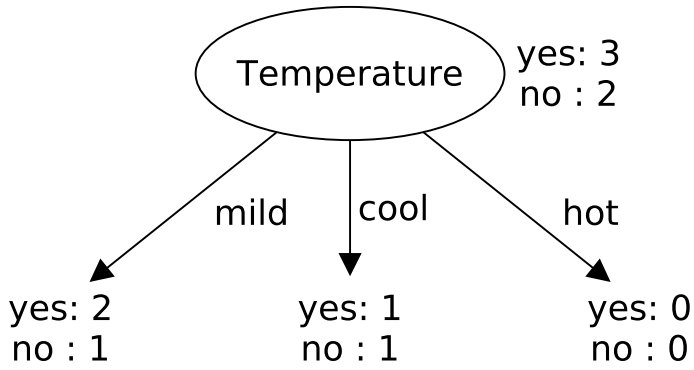


(d) $\overline{H} = 0.892, \overline{G} = 0.429, \overline{E} = 0.357$

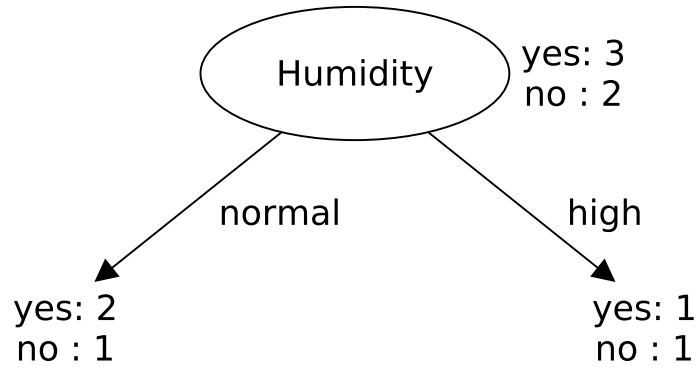
Decision Tree (1)



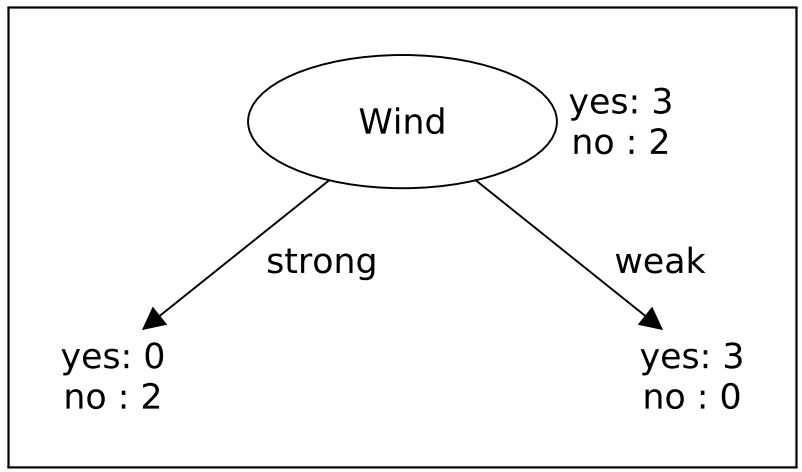
Which attribute to test, when Outlook = rainy?



(a) $\overline{H} = 0.951, \overline{G} = 0.467, \overline{E} = 0.4$

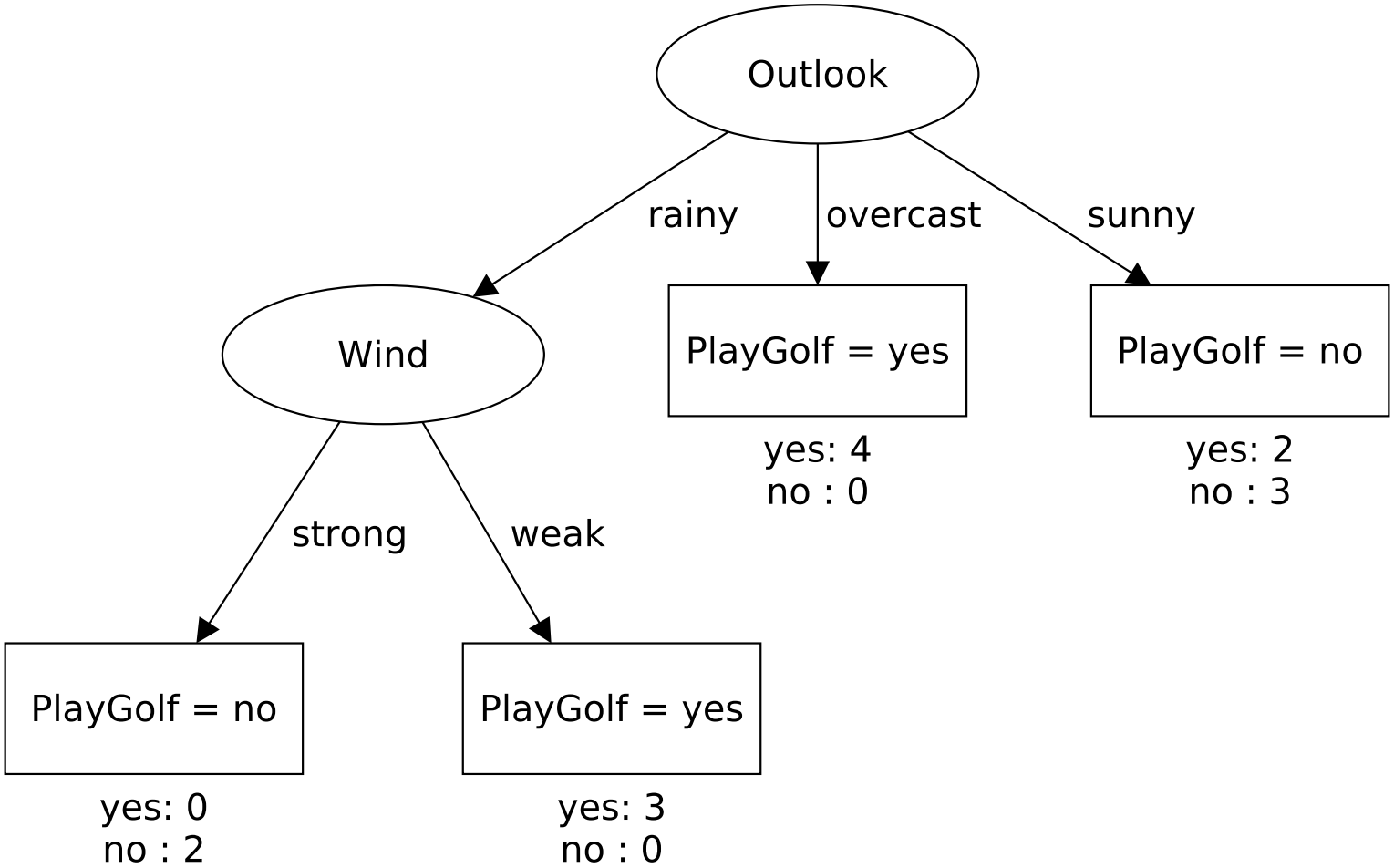


(b) $\overline{H} = 0.951, \overline{G} = 0.467, \overline{E} = 0.4$

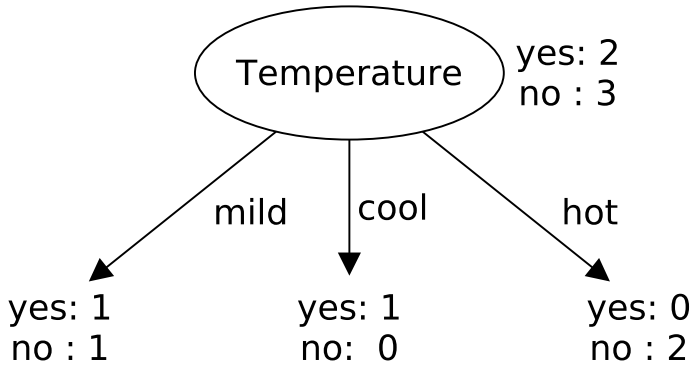


(c) $\overline{H} = 0, \overline{G} = 0, \overline{E} = 0$

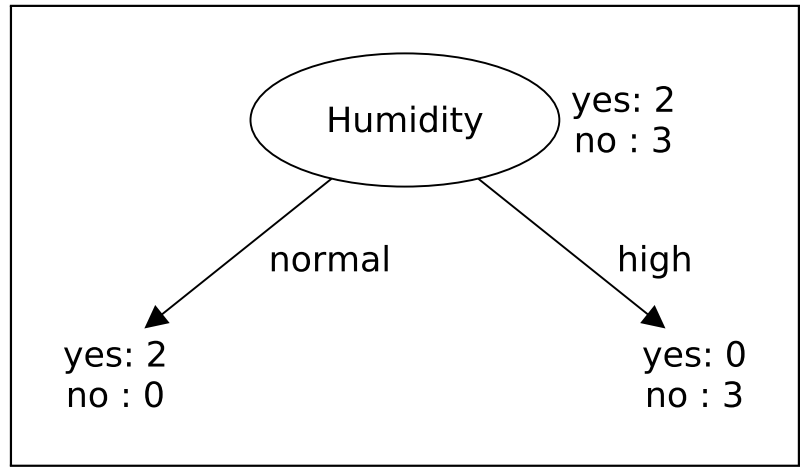
Decision Tree (2)



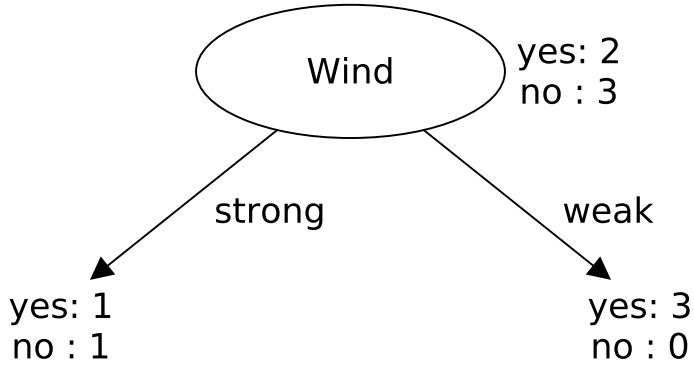
Which attribute to test, when Outlook = sunny?



(a) $\overline{H} = 0.4, \overline{G} = 0.2, \overline{E} = 0.2$



(b) $\overline{H} = 0, \overline{G} = 0, \overline{E} = 0$



(c) $\overline{H} = 0.4, \overline{G} = 0.2, \overline{E} = 0.2$

Decision Tree (3)