We want to use a decision tree to predict if I will play Golf

Let's assume the following features

Features:
- Outlook: Sun, Overcast, Rain
- Temperature: Hot, Mild, Cool
- Humidity: High, Normal, Low
- Wind: Strong, Weak
- Label: +, -

Features are evaluated in mid-morning, and the game is played in the afternoon

Let's look at the **training set**

| | | |
|---|---|---|
| 1. | S H H W | - |
| 2. | S H H S | - |
| 3. | O H H W | + |
| 4. | R M H W | + |
| 5. | R C N W | + |
| 6. | R C N S | - |
| 7. | O C N S | + |
| 8. | S M H W | - |
| 9. | S C N W | + |
| 10. | R M N W | + |
| 11. | S M N S | + |
| 12. | O M H S | + |
| 13. | O H N W | + |
| 14. | R M H S | - |

Outlook: S, O, R

Temp: H, M, C

Humidity: H, N, L

Wind: S, W

9 + 5 - examples

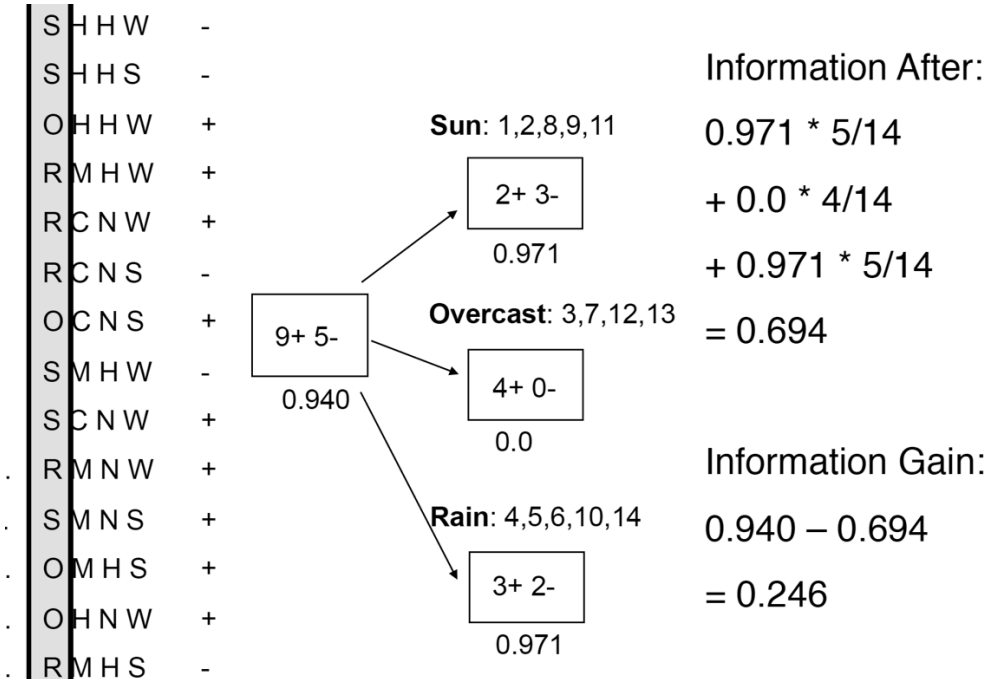
**Current entropy:**

$H(9/14)$

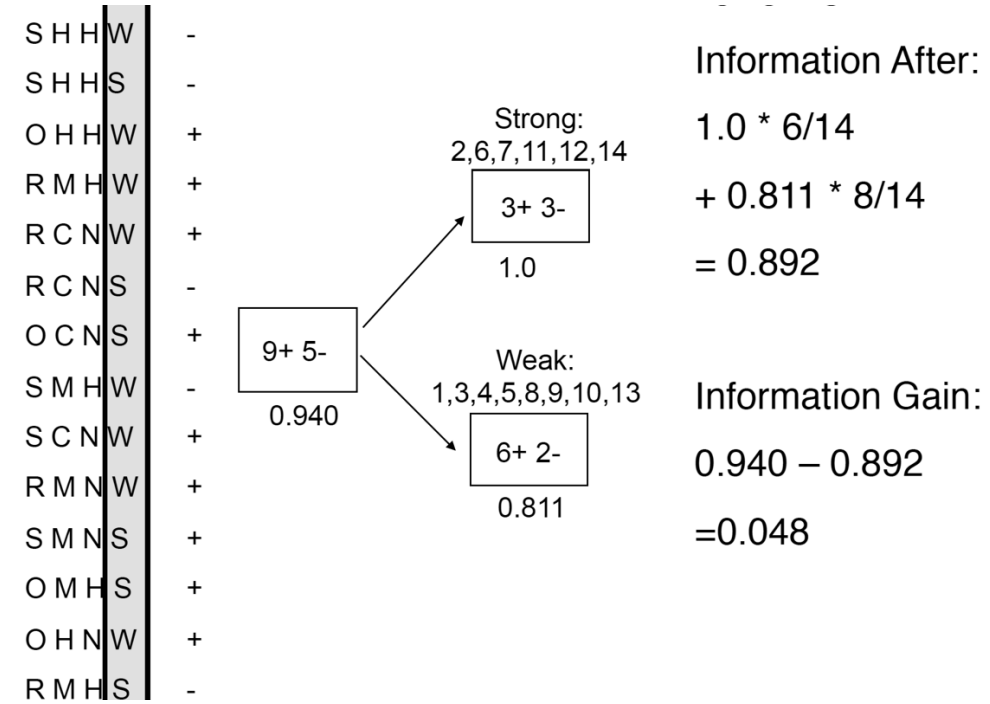
$= -9/14 \log_2(9/14) -5/14 \log_2(5/14)$

$\approx 0.94$

Now let's calculate the **outlook gain**



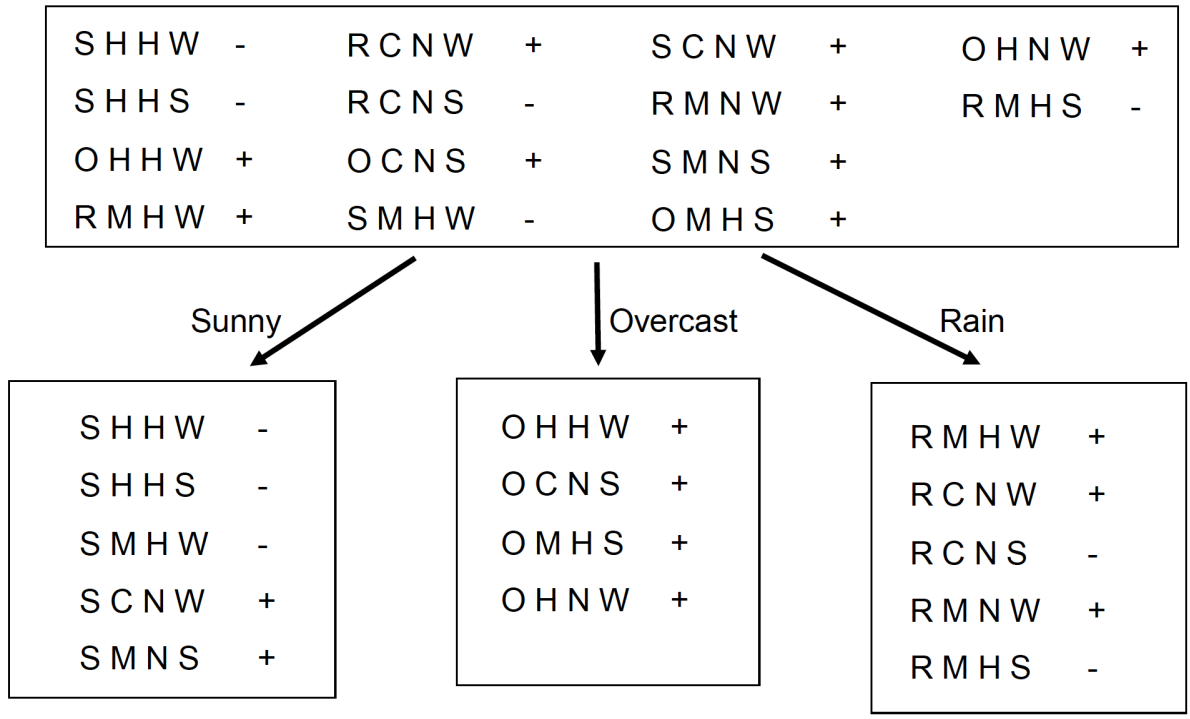

Let's calculate the **wind gain**




We can do the same for **temperature** and **humidity.** In summary:

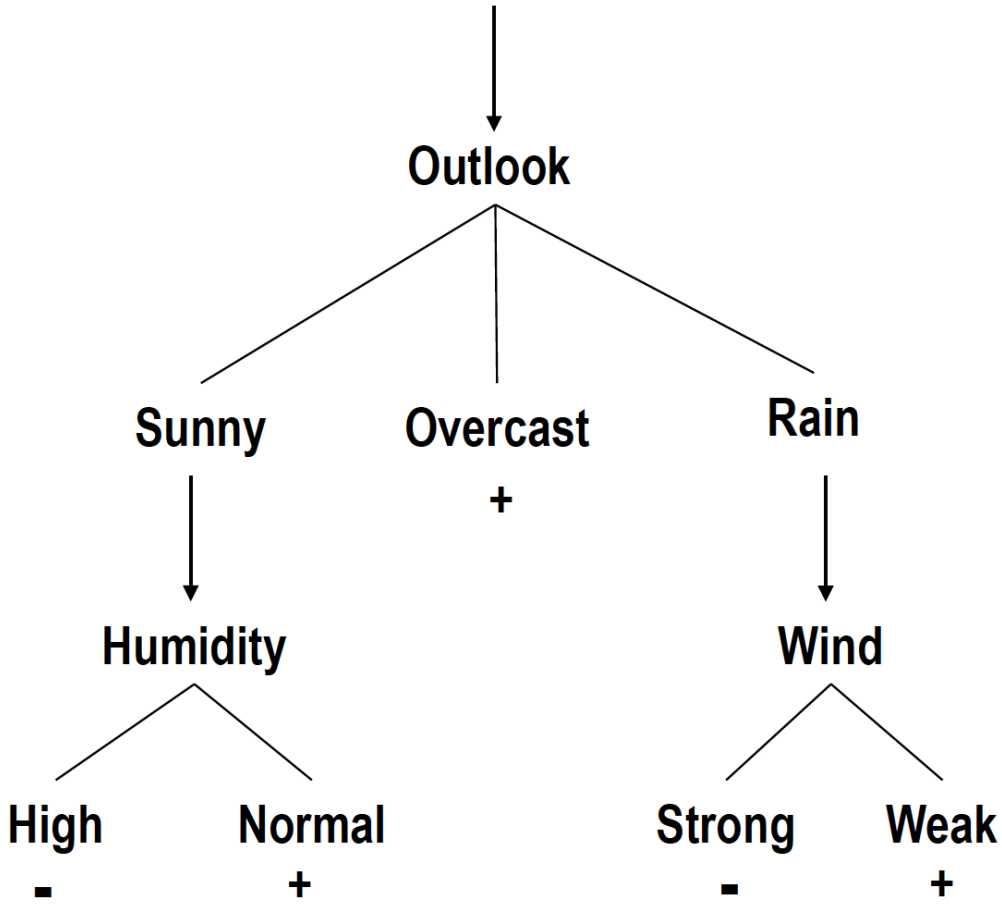
| Variable | Information gain |
|---|---|
| Outlook | 0.25 |
| Temperature | 0.03 |
| Humidity | 0.15 |
| Wind | 0.05 |

Since outlook provides greatest local gain, we use it for **splitting**:




Now **recurse** on each smaller set

What is the final decision tree? (Try to work out for yourself before seeing the answer on the next page)




Advanced questions:

-- Suppose under Sunny we split on Outlook (again) instead of Humidity?

-- What can we say about entropy as we measure additional features?