# Decision Trees with Numerical Attributes

# Dealing with continuous-valued attributes

Initial definition of ID3 is restricted in dealing with discrete sets of values.

It handles symbolic attribute effectively.

However, we have to extend it sphere to continuous-valued attributes(numeric attribute) to fit the real world scenario.

| No. | outlook | temperature | humidity | windy | play |
|---|---|---|---|---|---|
| 1 | sunny | 85 | 85 | FALSE | no |
| 2 | sunny | 80 | 90 | TRUE | no |
| 3 | overcast | 83 | 86 | FALSE | yes |
| 4 | rainy | 70 | 96 | FALSE | yes |
| 5 | rainy | 68 | 80 | FALSE | yes |
| 6 | rainy | 65 | 70 | TRUE | no |
| 7 | overcast | 64 | 65 | TRUE | yes |
| 8 | sunny | 72 | 95 | FALSE | no |
| 9 | sunny | 69 | 70 | FALSE | yes |
| 10 | rainy | 75 | 80 | FALSE | yes |
| 11 | sunny | 75 | 70 | TRUE | yes |
| 12 | overcast | 72 | 90 | TRUE | yes |
| 13 | overcast | 81 | 75 | FALSE | yes |
| 14 | rainy | 71 | 91 | TRUE | no |

weather.numeric.csv

What we have done is to define new discrete valued attributes that partition the continuous-valued attribute into symbolic attribute again.

For a numeric attribute A, we need to create a new boolean value that is true when **$A \leq c$** and false otherwise

The only thing left is to compute the best threshold **c**.

For humidity attribute

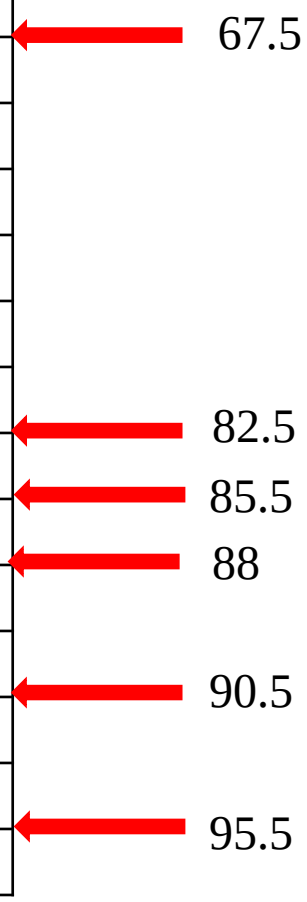
| humidity | play |
|---|---|
| 85 | no |
| 90 | no |
| 86 | yes |
| 96 | yes |
| 80 | yes |
| 70 | no |
| 65 | yes |
| 95 | no |
| 70 | yes |
| 80 | yes |
| 70 | yes |
| 90 | yes |
| 75 | yes |
| 91 | no |

First we need to sort the data

| humidity | play |
|---|---|
| 65 | yes |
| 70 | no |
| 70 | yes |
| 70 | yes |
| 75 | yes |
| 80 | yes |
| 80 | yes |
| 85 | no |
| 86 | yes |
| 90 | no |
| 90 | yes |
| 91 | no |
| 95 | no |
| 96 | yes |

We need a threshold that produces the greatest information gain

| humidity | play |
|---|---|
| 65 | yes |
| 70 | no |
| 70 | yes |
| 70 | yes |
| 75 | yes |
| 80 | yes |
| 80 | yes |
| 85 | no |
| 86 | yes |
| 90 | no |
| 90 | yes |
| 91 | no |
| 95 | no |
| 96 | yes |




Once sorting the numeric attribute values, then identifying adjacent examples that differ in their target classification

We can generate a set of candidate threshold

Then we compute **information gain** for each candidate and find the best one for splitting

$$Gain(S, A) = Entropy(S) - \sum_{v \in values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

$$Entropy\ (S) = -(p_+ \log_2 p_+) - (p_- \log_2 p_-)$$

$$Entropy\ (S) = -(p_+ \log_2 p_+) - (p_- \log_2 p_-) =$$
$$= -(9/14)\log_2(9/14) - (5/14)\log_2(5/14) = 0.94028$$

| humidity | play |
|---|---|
| 65 | yes |
| 70 | no |
| 70 | yes |
| 70 | yes |
| 75 | yes |
| 80 | yes |
| 80 | yes |
| 85 | no |
| 86 | yes |
| 90 | no |
| 90 | yes |
| 91 | no |
| 95 | no |
| 96 | yes |

67.5

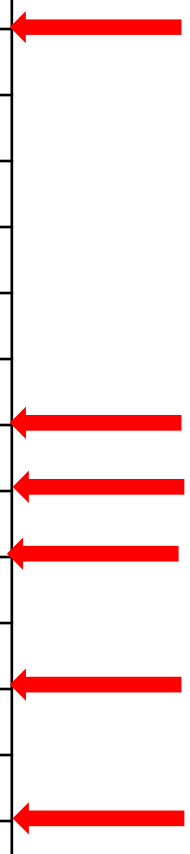
$$Entropy\ (humidity \le 67.5) = 0$$
$$Entropy\ (humidity > 67.5) =$$
$$= -(8/13)\log_2(8/13) - (5/13)\log_2(5/13) = 0.9612$$

$$Gain(S, humidity) = Entropy(S) - \sum_{v \in values(humidity)} \frac{|S_{humidity=v}|}{|S|} Entropy(S_{humidity=v})$$

$$Gain(S, humidity) = Entropy(S) - \frac{|S_{humidity \le 67.5}|}{|S|}(0) - \frac{|S_{humidity > 67.5}|}{|S|}(0.9612)$$

$$Gain(S, humidity) = 0.9402 - (1/14)(0) - (13/14)(0.9612) = 0.0477$$

| humidity | play |
|---|---|
| 65 | yes |
| 70 | no |
| 70 | yes |
| 70 | yes |
| 75 | yes |
| 80 | yes |
| 80 | yes |
| 85 | no |
| 86 | yes |
| 90 | no |
| 90 | yes |
| 91 | no |
| 95 | no |
| 96 | yes |



67.5 *Gain*(67.5) = 0.0477

82.5 *Gain*(82.5) = 0.1518

85.5 *Gain*(85.5) = 0.04812

88 *Gain*(88) = 0.1022

90.5 *Gain*(90.5) = 0.0793

93 *Gain*(93) = 0.0477

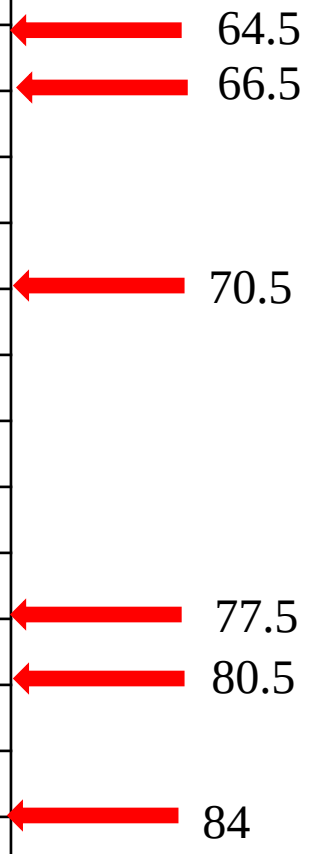
The maximum gain is 0.1518 so the chosen threshold should be 82.5

| temperature | play |
|---|---|
| 85 | no |
| 80 | no |
| 83 | yes |
| 70 | yes |
| 68 | yes |
| 65 | no |
| 64 | yes |
| 72 | no |
| 69 | yes |
| 75 | yes |
| 75 | yes |
| 72 | yes |
| 81 | yes |
| 71 | no |

First we need to sort the data

| temperature | play |
|---|---|
| 64 | yes |
| 65 | no |
| 68 | yes |
| 69 | yes |
| 70 | yes |
| 71 | no |
| 72 | no |
| 72 | yes |
| 75 | yes |
| 75 | yes |
| 80 | no |
| 81 | yes |
| 83 | yes |
| 85 | no |

| temperature | play |
|---|---|
| 64 | yes |
| 65 | no |
| 68 | yes |
| 69 | yes |
| 70 | yes |
| 71 | no |
| 72 | no |
| 72 | yes |
| 75 | yes |
| 75 | yes |
| 80 | no |
| 81 | yes |
| 83 | yes |
| 85 | no |




*Gain*(64.5) = 0.0477

*Gain*(66.5) = 0.0103

*Gain*(70.5) = 0.06455

*Gain*(77) = 0.0.00048

*Gain*(80.5) = 0.0103

*Gain*(84) = 0.0419

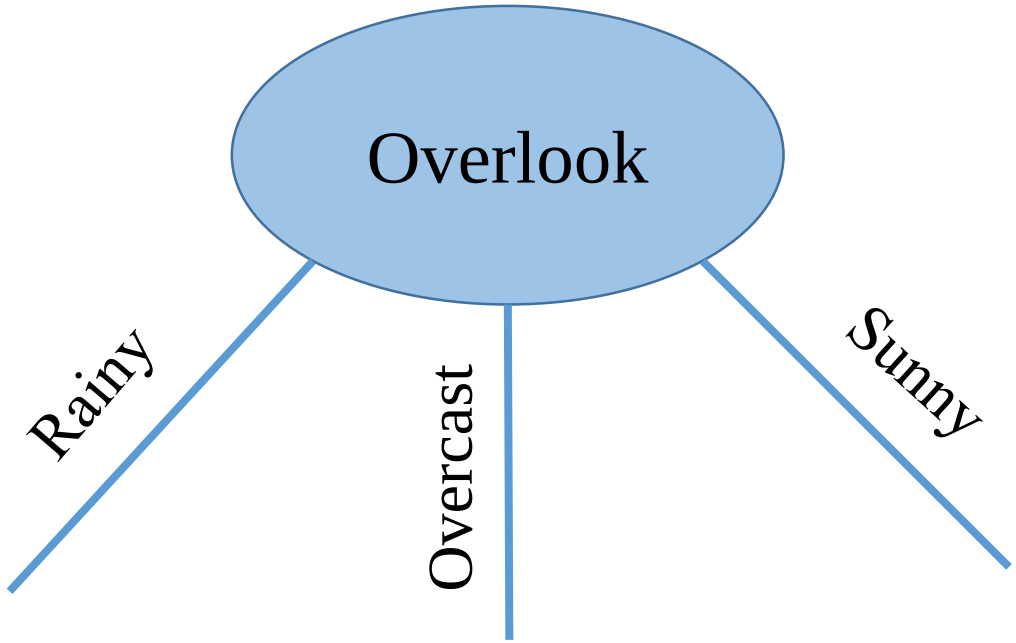
The maximum gain is 0.06455 so the chosen threshold should be 70.5

Gain(S,Oulook)=
Entropy(S)-(5/14)Entropy(sunny)-(4/14)Entropy(overlast)-(5/14)Entropy(rainy)= **0.2467498**

Gain(S,Windy)=
= Entropy(S)-(8/14)Entropy(FALSE)-(6/14)Entropy(TRUE)= **0.04812703**

Gain(S,temperature)=
= Entropy(S)-(5/14)Entropy(≤70.5)-(9/14)Entropy(>70.5)= **0.06455**

Gain(S,humidity)=
= Entropy(S)-(7/14)Entropy(≤82.5)-(7/14)Entropy(>82.5)= **0.1518**


Overlook
Rainy
Overcast
Sunny

Oulook=sunny

| No. | outlook | temperature | humidity | windy | play |
|---|---|---|---|---|---|
| 1 | sunny | 85 | 85 | FALSE | no |
| 2 | sunny | 80 | 90 | TRUE | no |
| 3 | overcast | 83 | 86 | FALSE | yes |
| 4 | rainy | 70 | 96 | FALSE | yes |
| 5 | rainy | 68 | 80 | FALSE | yes |
| 6 | rainy | 65 | 70 | TRUE | no |
| 7 | overcast | 64 | 65 | TRUE | yes |
| 8 | sunny | 72 | 95 | FALSE | no |
| 9 | sunny | 69 | 70 | FALSE | yes |
| 10 | rainy | 75 | 80 | FALSE | yes |
| 11 | sunny | 75 | 70 | TRUE | yes |
| 12 | overcast | 72 | 90 | TRUE | yes |
| 13 | overcast | 81 | 75 | FALSE | yes |
| 14 | rainy | 71 | 91 | TRUE | no |

| No. | outlook | temperature | humidity | windy | play |
|---|---|---|---|---|---|
| 1 | sunny | 85 | 85 | FALSE | no |
| 2 | sunny | 80 | 90 | TRUE | no |
| 8 | sunny | 72 | 95 | FALSE | no |
| 9 | sunny | 69 | 70 | FALSE | yes |
| 11 | sunny | 75 | 70 | TRUE | yes |

$$Gain(S) = -(2/5)\log_2(2/5) - (3/5)\log_2(3/5) = 0.9709$$

| No. | outlook | temperature | humidity | windy | play |
|---|---|---|---|---|---|
| 9 | sunny | 69 | 70 | FALSE | yes |
| 11 | sunny | 75 | 70 | TRUE | yes |
| 1 | sunny | 85 | 85 | FALSE | no |
| 2 | sunny | 80 | 90 | TRUE | no |
| 8 | sunny | 72 | 95 | FALSE | no |

77.5

*Entropy(humidity≤80)=0*

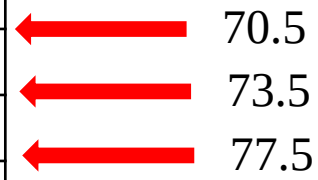
*Entropy(humidity>80)=0*

*Gain*(Outlook=sunny, humidity)= 0.9709

| No. | outlook | temperature | humidity | windy | play |
|---|---|---|---|---|---|
| 1 | sunny | 85 | 85 | FALSE | no |
| 2 | sunny | 80 | 90 | TRUE | no |
| 8 | sunny | 72 | 95 | FALSE | no |
| 9 | sunny | 69 | 70 | FALSE | yes |
| 11 | sunny | 75 | 70 | TRUE | yes |

$$Gain(S) = -(2/5)\log_2(2/5) - (3/5)\log_2(3/5) = 0.9709$$

| No. | outlook | temperature | humidity | windy | play |
|---|---|---|---|---|---|
| 9 | sunny | 69 | 70 | FALSE | yes |
| 8 | sunny | 72 | 95 | FALSE | no |
| 11 | sunny | 75 | 70 | TRUE | yes |
| 2 | sunny | 80 | 90 | TRUE | no |
| 1 | sunny | 85 | 85 | FALSE | no |




*Gain*(Outlook=sunny, temperatura(70.5))= 0.3218

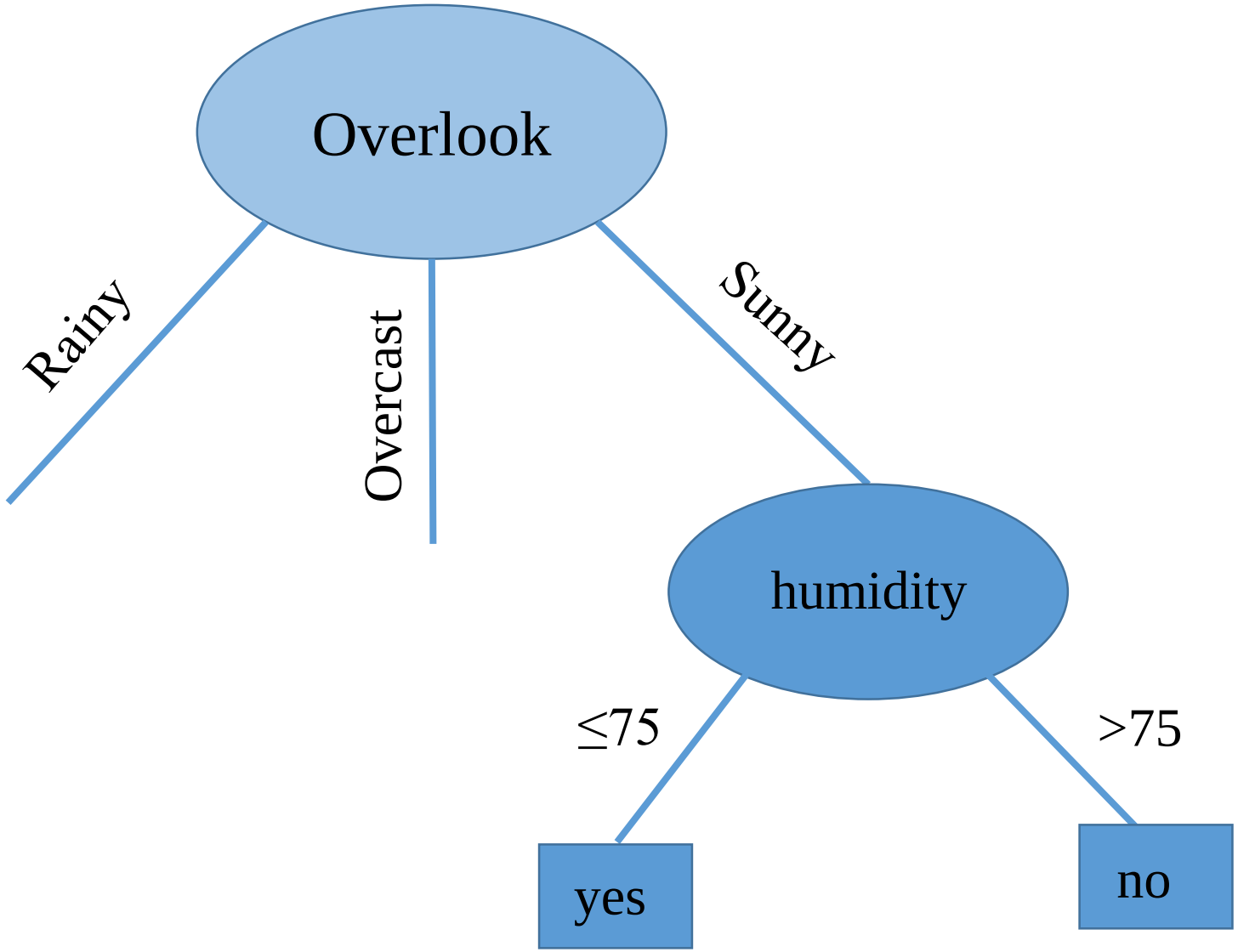
*Gain*(Outlook=sunny, temperatura(73.5))= 0.0199

*Gain*(Outlook=sunny, temperatura(77.5))= 0.4199

| No. | outlook | temperature | humidity | windy | play |
|---|---|---|---|---|---|
| 1 | sunny | 85 | 85 | FALSE | no |
| 2 | sunny | 80 | 90 | TRUE | no |
| 8 | sunny | 72 | 95 | FALSE | no |
| 9 | sunny | 69 | 70 | FALSE | yes |
| 11 | sunny | 75 | 70 | TRUE | yes |

$$\textit{Gain}(S) = -(2/5)\log_2(2/5) - (3/5)\log_2(3/5) = 0.9709$$

$$\textit{Gain}(\text{Outlook=sunny, Windy}) = 0.0199$$


Overlook
Rainy
Overcast
Sunny
humidity
≤75
>75
yes
no