# Decision tree

*SONG JIANXING*

Data: weather\_norminal.CSV

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **num** | **outlook** | **temperature** | **humidity** | **windy** | **play** |
| 1 | sunny | hot | high | F | no |
| 2 | sunny | hot | high | T | no |
| 3 | Overcast | hot | high | F | yes |
| 4 | rainy | mild | high | F | yes |
| 5 | rainy | cool | normal | F | yes |
| 6 | rainy | cool | normal | T | no |
| 7 | Overcast | cool | normal | T | yes |
| 8 | sunny | mild | high | F | no |
| 9 | sunny | cool | normal | F | yes |
| 10 | rainy | mild | normal | F | yes |
| 11 | sunny | mild | normal | T | yes |
| 12 | Overcast | mild | high | T | yes |
| 13 | Overcast | hot | normal | F | yes |
| 14 | rainy | mild | high | T | no |

## Method I: ID3

response variable(play): [yes9,no5]

**Total:**

Entropy(s) = info([9,5]) = -9/14log(9/14)-5/14log5/14 (base number of log is 2)

=0.940

Use each of predictor variable (outlook, temperature, humindy, wind) as key to classification discuss.

**Outlook:**

***Sunny:***

Result of sunny: [yes2,no3]

Entropy(sunny) = Info([2,3]) = -2/5log(2/5)-3/5log(3/5) = 0.971

***Overcast:***

Result of overcast: [yes4,no0]

Entropy(overcast) = info([4,0]) = -4/4log(4/4) – 0log(0) = 0

***Rainy:***

Result of rainy: [yes3,no2]

Entropy(rainy) = info([3,2]) = -3/5log(3/5) – 2/5log(2/5) = 0.971

***Summary outlook***:

Entropy(outlook) = info([2,3],[4,0],[3,2]) = 5/14info[2,3] + 4/14info[4,0] + 5/14info[3,2]

= 5/14×0.971 + 4/14×0 + 5/14×0.971

= 0.694

***information gain:***

gain(outlook) = info([9,5]) - info([2,3],[4,0],[3,2])

= 0.940 – 0.694

= 0.246

**Temperature:**

***Hot:***

Result of hot: [yes2,no2]

Entropy(hot) = Info([2,2]) = -2/4log(2/4) – 2/4log(2/4) = 1

***Mild:***

Result of mild: [yes5,no1]

Entropy(mild) = info([5,1]) = -5/6log(5/6) – 1/6log(1/6) = 0.65

***Cool:***

Result of cool: [yes3,no1]

Entropy(cool) = info([3,1]) = -3/4log(3/4) – 1/4log(1/4) = 0.811

***Summary temperature:***

Entropy(temperature) = info([2,2],[5,1],[3,1]) = 4/14info([2,2]) + 6/14info([5,1]) + 4/14info([3,1])

= 4/14×1 + 6/14×0.65 + 4/14×0.811

= 0.796

***information gain:***

gain(temperature) = info([9,5]- info([2,2],[5,1],[3,1])

= 0.940 - 0.796

= 0.029

***Humidity:***

***High:***

Result of high: [yes3,no4]

Entropy(high) = info([3,4]) = -3/7log(3/7) – 4/7log(4/7) = 0.985

***Normal:***

Result of normal: [yes6,no1]

Entropy(normal) = info([6,1]) = -6/7log(6/7) – 1/7log(1/7) = 0.591

***Summary humidity:***

Entropy(humidity) = info([3,4],[6,1]) = 7/14info([3,4])+7/14info([6,1]) = 0.788

***information gain:***

gain(humidity) = info([9,5])- info([3,4],[6,1]) = 0.940-0.788 = 0.152

***Windy:***

***True:***

Result of true: [yes3,no3]

Entropy(ture) = Info([3,3]) = -3/6log3/6 – 3/6log3/6 = 1

***False:***

Result of false: [yes6,no2]

Entropy(false) = info([6,2]) = -6/8log(6/8) – 2/8log(2/8) = 0.811

***Summary windy:***

Entropy(windy) = info([3,3],[6,2]) = 0.892

***information gain:***

gain(windy) = info([9,5]) – info([3,3],[6,2]) = 0.05

gain(outlook) = 0.25

gain(tempearture) = 0.03

gain(humidity) = 0.15

gain(wind) = 0.05

compare them to find gain(outlook) is max

**A: Set ‘outlook’ as Root of the tree, the preliminary is as follows:**



Take ‘sunny’ as total information

Sunny: [yes2,no3]

Info([2,3]) = -2/5log(2/5)-3/5log(3/5) = 0.97`

**B: Sunny: All possible results are as follows**

***Result I***



Entropy(temperature) = Info([0,2],[1,1],[1,0]) = 0.40

Gain(temperature) = Info([2,3]) - Info([0,2],[1,1],[1,0]) = 0.57

***Result II***



Entropy(humidity) = info([0,3],[2,0]) = 0

Gain(humidity) = Info([2,3]) - info([0,3],[2,0]) = 0.97

***Result III***



Entropy(wind) = info([2,1],[1,1]) = 0.95

Gain(wind) = Info([2,3]) - info([2,1],[1,1]) = 0.02

Gain(temperature) = 0.57

Gain(humidity) = 0.97

Gain(wind) = 0.02

compare them to find gain(humidity) is max

**C: Set ‘humidity’ as Internal node of the ‘sunny’**

**D: For ‘overcast’,** **recursive termination has been formed, do not need to classficated**

**F: Rainy: All possible results are as follows**

Rainy:[yes3,no2]

Info([3,2]) = -3/5log(3/5) – 2/5log(2/5) = 0.971

***Result I***



Entropy(humidity) = info([1,1],[2,1]) = 0.951

Gain(humidity) = Info([3,2]) - info([1,1],[2,1]) = 0.02

***Result II***



Entropy(temperature) = info([2,1],[1,1]) = 0.951

Gain(temperature) = Info([3,2]) - info([1,1],[2,1]) = 0.02

***Result III***



Entropy(wind) = info([3,0],[0,2]) = 0

Gain(wind) = Info([3,2]) – info([3,0],[0,2]) = 0.971

Gain(humidity) = 0.02

Gain(temperature) = 0.02

Gain(wind) = 0.971

compare them to find gain(wind) is max

**G: Set ‘wind’ as Internal node of the ‘rainy’**

**H: To sum up, the decision tree is generated as follows**

