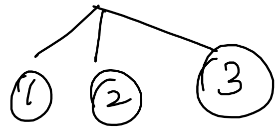
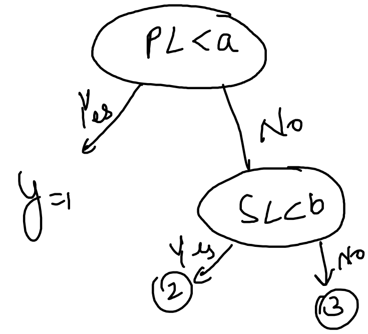
Sunday, May 26, 2024 9:36 AM

Decision Trees

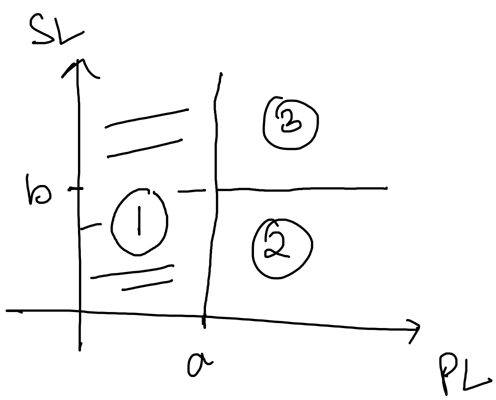
IRIS Dataset = [SL, SW, PL, PW]






```
if PL<a :
    y1 = 1
else :
    if SL<b :
        y1 = 2
    else :
        y1 = 3
```

Recursive Partitioning :→ (Axis Parallel hyperplanes)



(a, 0)

```
if PL<a :
    yi=1
else :
    if SL<b :
        yi=2
    else :
        yi=3
```

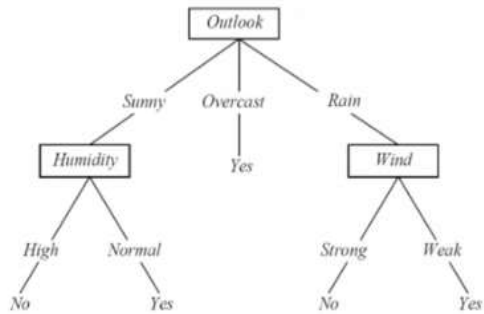
Decision Trees

→ Entropy
→ Gini Impurity
→ Information Gain

Entropy → Randomness in dataset

$$H_D(Y) = -\sum_{i=1}^{n} p_i \lg(p_i) \qquad \lg \Rightarrow \log_2$$

Outlook	Temperature	Humidity	Windy	PlayTennis
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Rainy	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Sunny	Mild	High	False	No
Sunny	Cool	Normal	False	Yes
Rainy	Mild	Normal	False	Yes
Sunny	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Rainy	Mild	High	True	No



Parent's Entropy

$$H_D(Y) = -\sum_{i=1}^{n} p_i \lg(p_i)$$

$$= -P(Yes)\lg P(Yes) - P(No)\lg P(No)$$

$$P(Yes) = \frac{9}{14} \qquad P(No) = \frac{5}{14}$$

$$H_D(Y) = -\frac{9}{14} \times \lg\left(\frac{9}{14}\right) - \frac{5}{14}\lg\left(\frac{5}{14}\right)$$

$$= 0.94$$

entropy of each column:

0.94

Outlook
- Sunny (5) $(2Y, 3N) \rightarrow -\frac{2}{5}\lg\frac{2}{5} - \frac{3}{5}\lg\frac{3}{5} = 0.97$
- Overcast $(4Y, 0N) \rightarrow -\frac{4}{4}\lg\frac{4}{4} - 0\lg 0 = 0$
- Rainy $(3Y, 2N) \rightarrow -\frac{3}{5}\lg\frac{3}{5} - \frac{2}{5}\lg\frac{2}{5} = 0.97$

weighted entropy $= \frac{5}{14}\times 0.97 + \frac{4}{14}\times 0 + \frac{5}{14}\times 0.97$

$\Rightarrow \quad 0.97 \times \frac{5}{7} = 0.69$

Temperature
- Hot $\frac{}{4}$ $(2Y, 2N) = -\frac{2}{4}\lg\frac{2}{4} - \frac{2}{4}\lg\frac{2}{4} = 1$
- Mild $\frac{}{6}$ $(4Y, 2N) = -\frac{4}{6}\lg\frac{4}{6} - \frac{2}{6}\lg\frac{2}{6} = 0.91$
- Cool $\frac{}{4}$ $(3Y, 1N) = -\frac{3}{4}\lg\frac{3}{4} - \frac{1}{4}\lg\frac{1}{4} = 0.81$

weighted entropy $= \frac{4}{14}\times 1 + \frac{6}{14}\times 0.91 + \frac{4}{14}\times 0.81$

$H_D(Y, \text{temp})$

$= 0.91$

Humidity
- 7 High $(3Y, 4N) \rightarrow -\frac{3}{7}\lg\frac{3}{7} - \frac{4}{7}\lg\frac{4}{7} = 0.98$
- 7 Normal $(6Y, 1N) \rightarrow -\frac{6}{7}\lg\frac{6}{7} - \frac{1}{7}\lg\frac{1}{7} = 0.59$

└── Normal $\rightarrow -\frac{6}{7}\lg\frac{6}{7} - \frac{1}{7}\lg\frac{1}{7} = 0.59$

$H_D(Y, \text{humidity})$

weighted entropy $= \frac{7}{14} \times 0.98 + \frac{7}{14} \times 0.59$

$= \frac{7}{14} \times 1.57 = 0.78$

Windy
- 6 (3Y, 3N) True $\rightarrow -\frac{3}{6}\lg\frac{3}{6} - \frac{3}{6}\lg\frac{3}{6} = 1$
- 8 (6Y, 2N) False $\rightarrow -\frac{6}{8}\lg\frac{6}{8} - \frac{2}{8}\lg\frac{2}{8} = 0.81$

weighted entropy

$H_D(Y, \text{Windy}) = \frac{6}{14} \times 1 + \frac{8}{14} \times 0.81$

$= 0.89$

Choosing the column for split:

a) Compare weighted entropies of column & choose column with least entropy

outlook	temp	humidity	windy
0.69	0.91	0.78	0.89

0.69   0.91

(b) Information Gain (IG) = Parent's entropy − weighted entropy of each column

highest IG $\Rightarrow$ $IG(Y, \text{outlook}) = 0.94 - 0.69 = 0.25$

$IG(Y, \text{temp}) = 0.94 - 0.91 = 0.03$

$IG(Y, \text{humidity}) = 0.94 - 0.78 = 0.16$

$IG(Y, \text{windy}) = 0.94 - 0.89 = 0.05$

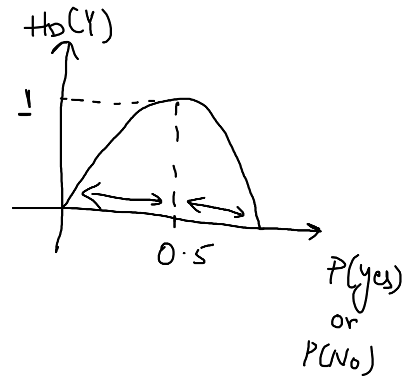
outlook is chosen because it has highest information gain.

**

Properties

(1) $P(\text{yes}) = \frac{1}{2}$ $P(\text{No}) = \frac{1}{2}$

$$H_D(y) = -\frac{1}{2}\lg\frac{1}{2} - \frac{1}{2}\lg\frac{1}{2} = 1$$



(2) $P(\text{yes}) = 1$ $P(\text{No}) = 0$

$$H_D(y) = -\frac{1}{1}\lg\frac{1}{1} - 0\lg 0 = 0$$

Gini Impurity (I_G)

$IG \neq I_G$

$$I_G(Y) = 1 - \sum_{i=1}^{n} p_i^2$$

→ binary classification ⇒

$$I_G(Y) = 1 - [P(+)^2 + P(-)^2]$$

→ multiclass classification ⇒

$$I_G(Y) = 1 - [p(y_1)^2 + p(y_2)^2 + p(y_3)^2 + \dots + p(y_n)^2]$$

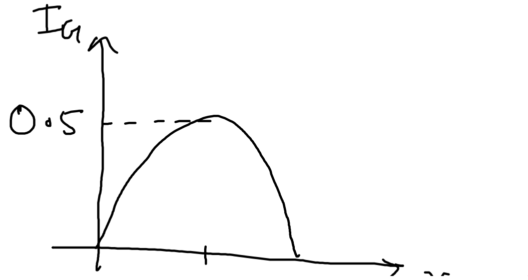
Properties:

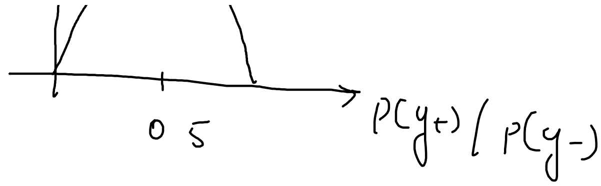
(1) $P(y+) = 0.5 \qquad p(y-) = 0.5$

$$I_G = 1 - [0.5^2 + 0.5^2] = 1 - [0.5] = 0.5$$

(2) $P(y+) = 1 \qquad p(y-) = 0$

$$I_G = 1 - [1^2 + 0^2] = 0$$





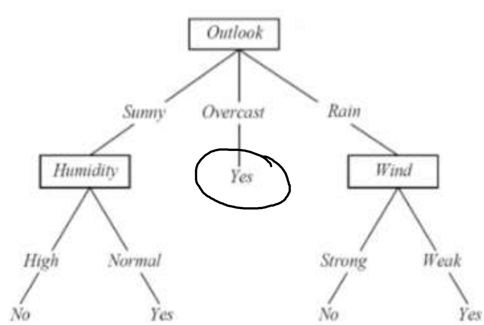
Comparison of Gini Impurity & Entropy:

1) $H_D(Y) = -\sum_{i=1}^{n} p_i \lg p_i$ $\qquad I_G = 1 - \sum_{i=1}^{n} p_i^2$

Since, I_G is computationally efficient, it can be used for larger datasets

Entropy shouldn't be used for large dataset.

Outlook	Temperature	Humidity	Windy	PlayTennis
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Rainy	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Sunny	Mild	High	False	No
Sunny	Cool	Normal	False	Yes
Rainy	Mild	Normal	False	Yes
Sunny	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Rainy	Mild	High	True	No



Parent's $I_G = 1 - [p(yes)^2 + p(No)^2]$

$= 1 - \left[\left(\frac{9}{14}\right)^2 + \left(\frac{5}{14}\right)^2\right]$

$= 0.45$

I_G for each column:

0.45

Outlook
- 5 (2Y, 3N) Sunny $\Rightarrow 1-\left[\left(\frac{2}{5}\right)^2+\left(\frac{3}{5}\right)^2\right]=0.48$
- 4 (4Y, 0N) overcast $\Rightarrow 1-\left[\left(\frac{4}{4}\right)^2+0^2\right]\Rightarrow 0$
- 5 (3Y, 2N) rainy $\Rightarrow 1-\left[\left(\frac{3}{5}\right)^2+\left(\frac{2}{5}\right)^2\right]=0.48$

weighted $I_G \Rightarrow \frac{5}{14}\times 0.48+\frac{4}{14}\times 0+\frac{5}{14}\times 0.48$

$\Rightarrow \frac{5}{14}\times 0.96 \Rightarrow 0.342$

Temperature
- 4 (2Y, 2N) Hot $\Rightarrow 1-\left[\left(\frac{2}{4}\right)^2+\left(\frac{2}{4}\right)^2\right]=0.5$
- 6 (4Y, 2N) Mild $\Rightarrow 1-\left[\left(\frac{4}{6}\right)^2+\left(\frac{2}{6}\right)^2\right]=0.444$
- 4 (3Y, 1N) Cool $\Rightarrow 1-\left[\left(\frac{3}{4}\right)^2+\left(\frac{1}{4}\right)^2\right]=0.375$

Weighted $I_G = \frac{4}{14}\times 0.5+\frac{6}{14}\times 0.444+\frac{4}{14}\times 0.375 = 0.44$

Humidity
- 7 High (3Y, 4N) $\Rightarrow 1-\left[\left(\frac{3}{7}\right)^2+\left(\frac{4}{7}\right)^2\right]=0.48$
- 7 Normal (6Y, 1N) $\Rightarrow 1-\left[\left(\frac{6}{7}\right)^2+\left(\frac{1}{7}\right)^2\right]=0.24$

Weighted $I_G = \frac{7}{14}\times 0.48+\frac{7}{14}\times 0.24 = 0.36$

6 (3Y, 3N) True $\Rightarrow 1-\left[\left(\frac{3}{6}\right)^2+\left(\frac{3}{6}\right)^2\right]=0.5$

Windy
- True: 6 (3Y, 3N) $\Rightarrow 1 - \left[\left(\frac{3}{6}\right)^2 + \left(\frac{3}{6}\right)^2\right] = 0.5$
- False: 8 (6Y, 2N) $\Rightarrow 1 - \left[\left(\frac{6}{8}\right)^2 + \left(\frac{2}{8}\right)^2\right] = 0.375$

weighted $I_G \Rightarrow \frac{6}{14} \times 0.5 + \frac{8}{14} \times 0.375$

$\Rightarrow 0.42$

Choose column for split.

a) choose column with least I_G

outlook	temp	humidity	windy
0.342	0.44	0.36	0.42

b) Info Gain: Parent's I_G − weighted I_G for each column

highest IG chosen for split

$I_G(Y, \text{outlook}) = 0.45 - 0.342 = 0.108$

$I_G(Y, \text{temp}) = 0.45 - 0.44 = 0.01$

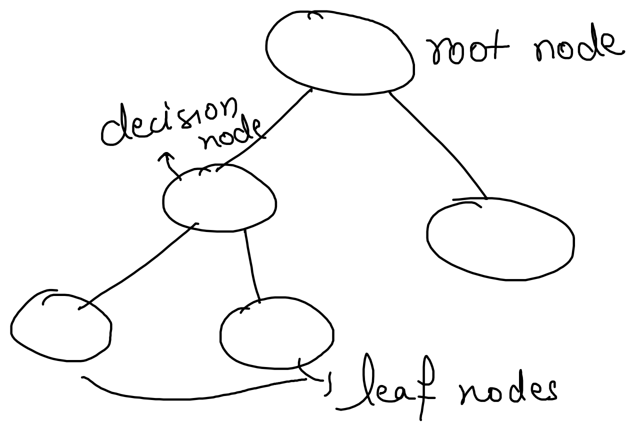
$I_G(Y, \text{humidity}) = 0.45 - 0.36 = 0.09$

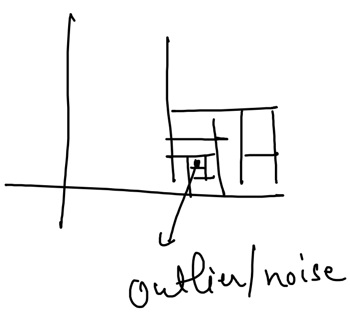
$I_G(Y, \text{windy}) = 0.45 - 0.42 = 0.03$

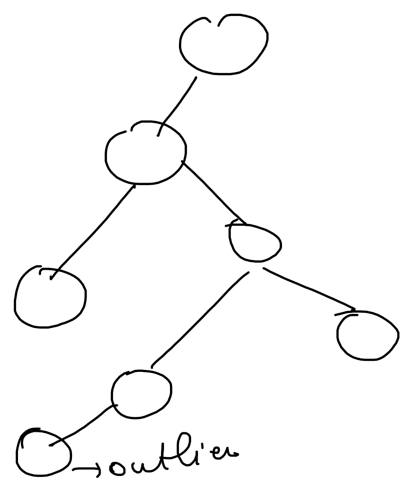
When to stop a tree:

→ pure node

→ If you have very few samples in terminal nodes

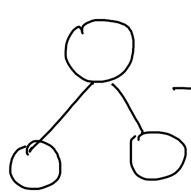






Hyperparameter: max_depth↑ ⇒ height (depth)↑ ⇒ Overfitting↑

↳ (max value = 32)



→ decision Stump

(tree of depth = 1)

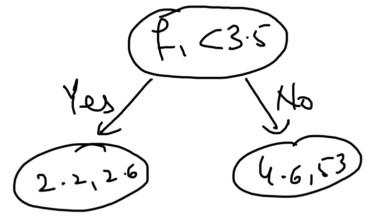
Splitting Numerical features

f_1	y
2.2	1
2.6	1
3.5	0
4.6	1
5.3	0

1. sort the f_1
2. $f_1 < 2.6$ $\quad f_1 < 5.3$

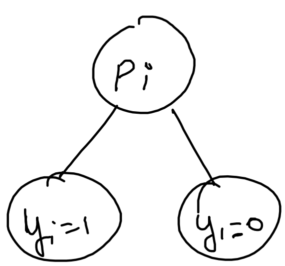
$f_1 < 3.5 \to$ Lowest IG/H

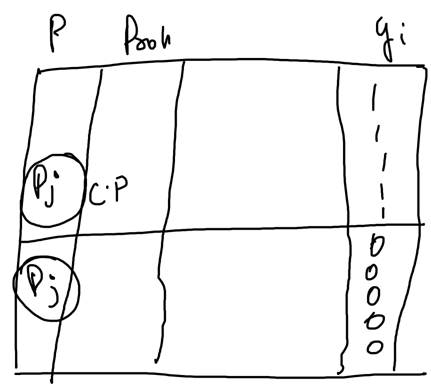
$f_1 < 4.6$



Feature Engineering: column with lot of categories

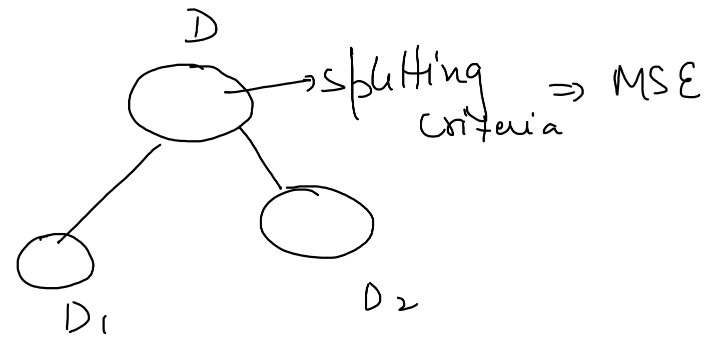
PINCODE





$$P(y=0/P_j) = \frac{P_j \cap y=0}{P_j} \qquad P(y=1|P_j) = \frac{P_j \cap y}{P_j}$$

Regression in DT



Advantages:

- easy to intrepret
- important features can be extracted
- No need to standardize