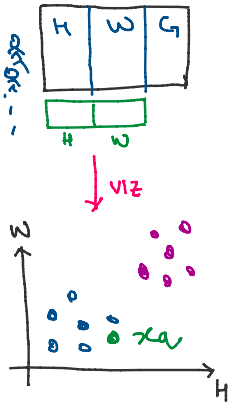
* Supervised Learning
* Rule Based Approach
* Tasks → Classification (How?)
  ↳ Entropy, Information Gain
  Regression (How?)
  ↳ MSE

- Given Data containing Height, Weights & Gender of some individuals.

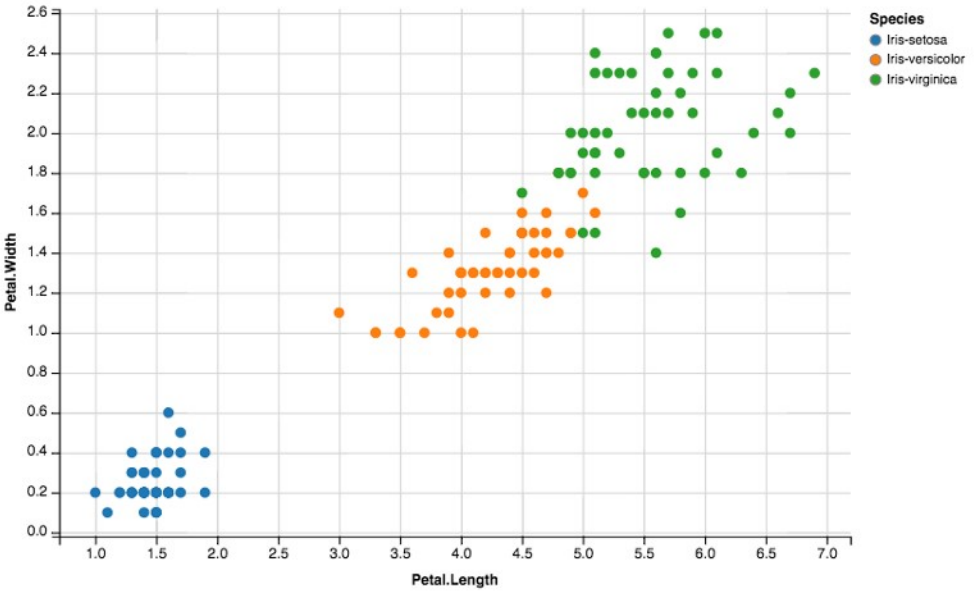


* Predict the Gender for x_a

Type → Task
? ?

Evaluation Metric → ?

1. Target Variable → Gender (y)
   Gender ∈ {M, F}
2. Task → Classification
3. Evaluation Metrics → Accuracy
4. (How?) Algorithms →
   Make the splits in such a way that impurity of new partitions is reduced.
   (i.e Try to get homogeneous nodes)

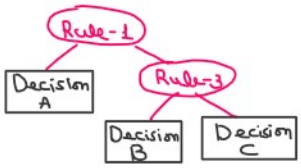


Rules ?

1. For setosa
   - PL < 2.5
   - PW < 0.8
2. For versicolor
   - PL < 5.0
   - PW < 1.6
3. For virginica
   - PL ≥ 5.0
   - PW ≥ 1.6

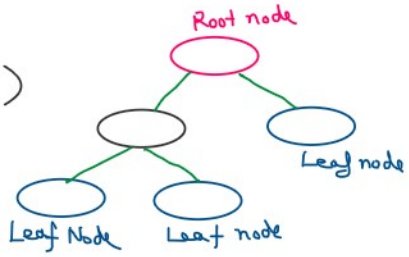
Learning Phase

X_train , y_train → Decision Tree Algo → Model (Set of rules arranged in tree structure)



Understanding the Tree Structure

* Nodes (Root, Intermediate & Leaf nodes)
* Branches (No cycles)

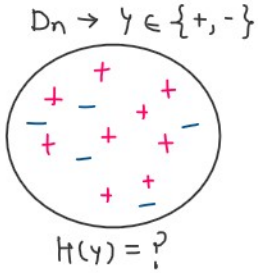


→ Represent rules in 'ROOT' & 'INTERMEDIATE' Nodes

→ Represent decisions in the 'LEAF' Nodes

Consider this binary classification data set:

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes

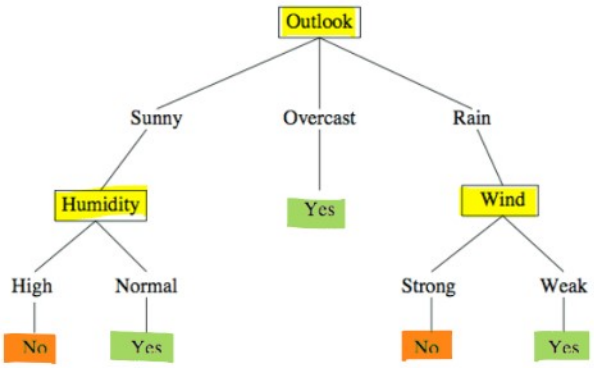


Rules ?

D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No
X_q	Rain	Hot	Normal	Weak	?

Given X_q → Outlook = 'Rain' & Temp = 'Hot' & Humidity = 'Normal' & Wind = 'Weak'

We can describe this data set with the following decision tree:



Ways to create Decision Trees?

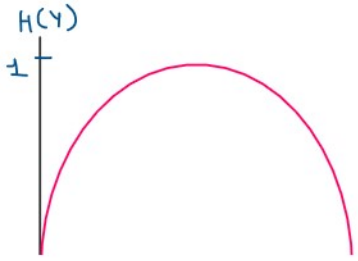
1. ID3 (Iterative Dichotomiser 3 (i.e Division)) → Only for classification task on categorical data.
   ↳ For classification – (Entropy, Information Gain)
2. C 4.5 → (Handles categorical as well as numerical data)
   ↳ For Classification – (Entropy , Information Gain Ratio)
   ↳ For Regression – (MSE, Minimize the sum of Weighted MSE)
3. C5.0 (Commercial)
4. SLIQ
5. SPRINT } (If the data is not on a single disk)
6. CART (Classification & Regression Trees) → sklearn implements
   ↳ For Classification – (Gini Index , Minimize the Sum of weighted Gini Index)
   ↳ For Regression – (MSE , Minimize the sum of weighted MSE)

Concepts

1. Entropy (Measure of impurity/randomness)

→ $$H(\text{Target}) = -\sum_{i=1}^{c} P(y_i) \log_2 (P(y_i))$$

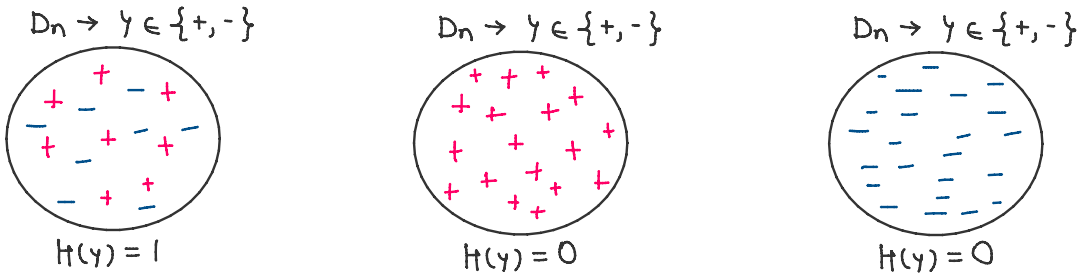
here, c → Number of classes



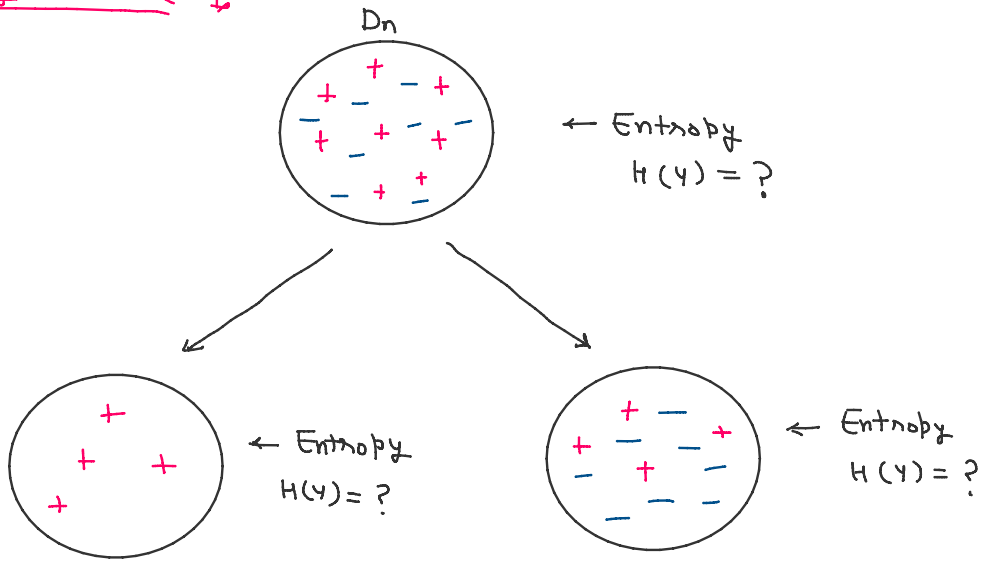
here, c → Number of classes

$$0 \leq H(Y) \leq 1$$

* More the Entropy, harder it is to draw the conclusions.



Splitting a Node



2. Information Gain (Helps in determining the Best split)

→ $IG(\text{input feature}, \text{target}) = H(\text{Target}) - \sum_{i=1}^{p} \left\{ \frac{|D_i|}{|D|} * H_{D_i}(Y) \right\}$

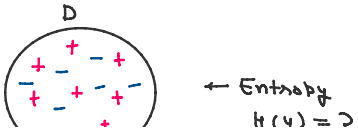
here, p → Total number of partitions

$|D_i|$ = no. of datapoints in each partition

$|D|$ = no. of datapoints before partitioning

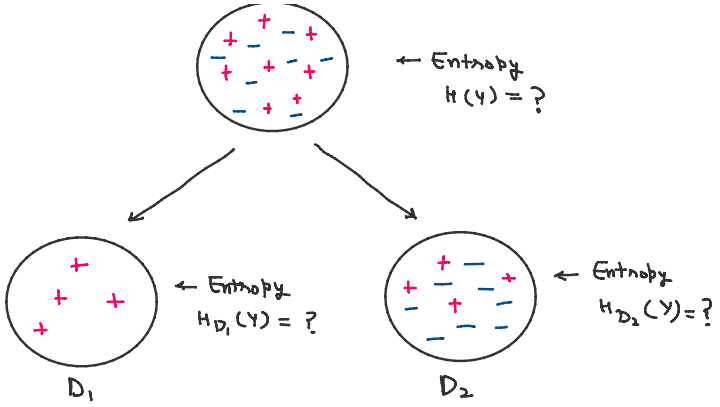
$H(Y)$ = Entropy before partition

$H_{D_i}(Y)$ = Entropy of D_i^{th} partition.



Information Gain (feature, Y)

$$= H(Y) - \sum \left\{ \frac{|D_i|}{|D|} * H_{D_i}(Y) \right\}$$



$$= H(Y) - \sum \left\{ \frac{|D_i|}{|D|} * H_{D_i}(Y) \right\}$$

$$= H(Y) - \left\{ \frac{|D_1|}{|D|} * H_{D_1}(Y) + \frac{|D_2|}{|D|} * H_{D_2}(Y) \right\}$$

$$= 1 - \left\{ \frac{4}{16} * 0 + \frac{12}{16} * \underline{?} \right\}$$

$$H_{D_2}(Y) = -\left(p(+ve) \log_2 p(+ve) + p(-ve) \log_2 p(-ve) \right)$$

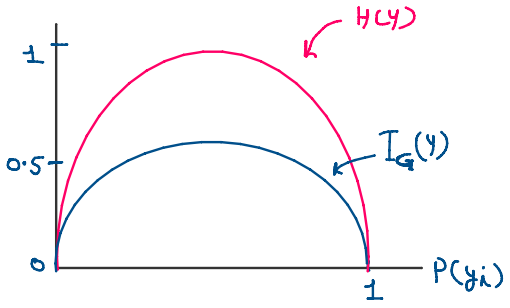
$$= -\left(\frac{4}{12} * \log_2 \frac{4}{12} + \frac{8}{12} \log_2 \frac{8}{12} \right)$$

$$= -\left(\frac{1}{3} * (\log 1 - \log 3) + \frac{2}{3} (\log 2 - \log 3) \right)$$

3. Gini Impurity

$$\rightarrow \quad I_G(\text{Target}) = 1 - \sum_{i=1}^{c} \left(P(y_i) \right)^2$$

$$\boxed{0 \leq I_G(Y) \leq 1 - \frac{1}{c}}$$



4. Information Gain Ratio

∵ Information Gain favours high branching features, let's look at Information Gain Ratio.

$$\text{Gain Ratio} = \frac{\text{Information Gain}}{\text{Intrinsic Information}}$$

(ie Penalize smaller partitions)

$$\text{Intrinsic Information} = -\sum_{i=1}^{b} \frac{|D_i|}{|D|} * \log_2 \frac{|D_i|}{|D|}$$

OR

Split Information

Decision Tree Algorithm	Task		Data	
	Classification	Regression	Categorical	Numerical
ID3	✓	✗	✓	✗
C4.5	✓	✓	✓	✓
CART (Sklearn)	✓	✓	✓	✓

Advantages of Decision Trees

a. Simple to interpret
b. Requires little data preparation
   (Numerical features – No rescaling required, Categorical feature – Label Encoding)
c. Very fast for prediction on query datapoints.
d. Can handle multi-class problems.
e. Provides feature importance.
f. Learns non-linear patterns.

Disadvantages of Decision Trees

a. Prone to overfitting & are unstable.
b. Prone to imbalanced data. Biased towards classes which dominates.
c. Predictions from decision tree are piecewise constant approximations.
d. Training time is more if lots of numerical features exist in input.