Supervised Learning - Unit 3

Decision Trees

S.V.Jansi Rani

Associate Professor / CSE

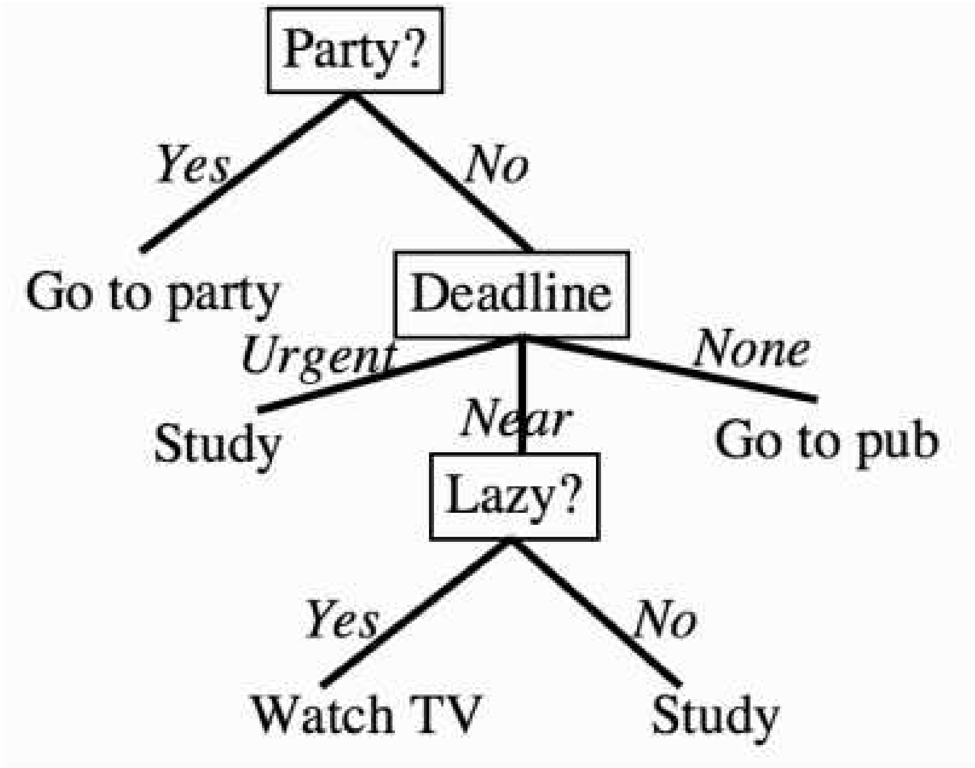
Agenda

- Introduction
- Constructing decision trees
- Information Gain using ID3
- Gini Impurity
- Classification Problem

S.V.JANSI RANI/Assoc Prof /CSE/SSNCE

Using Decision Trees

- Powerful data - binary tree
- Cost of using it is even lower: O(log N), where N is the number of datapoints.
- ML -querying the trained algorithm -fast
- DT - breaks classification down into a set of choices about
- each feature - starting at the root (base) of the tree and progressing down to the leaves- classification decision.
- Turned into a set of if-then rules, suitable for use in a rule induction system.



A simple decision tree to decide how you will spend the evening

- DT-popular is that we can turn them into a set of
- logical disjunctions (if ... then rules) that then go into program code very simply
- —the first part of the tree above can be turned into:

- `if` *there is a party* `then` *go to it*
- `if` *there is not a party* `and` *you have an urgent deadline* `then` *study*
- etc.

Constructing Decision Trees

- Three features - state of your energy level, the date of your nearest deadline, and whether or not there is a party tonight.
- How based on those features, we can construct the tree.
- Algorithms build the tree in a greedy manner starting at the root, choosing the most informative feature at each step.
- Choose a question that gives you the most information given what you know already. Thus, you would ask 'Is it
- an animal?' before you ask 'Is it a cat?'.
- How much information is provided to you by knowing certain facts - encoding this mathematically-information theory.
- **Information theory, a mathematical representation of the conditions and parameters affecting the transmission and processing of information.**

Entropy

- Measure of information entropy, which describes the amount of impurity in a set of features. The entropy H of a set of probabilities p_i is

$$\text{Entropy}(p) = -\sum_i p_i \log_2 p_i,$$

- log - binary digits (bits), 0log 0 =0
- H - high, information is high, select that feature

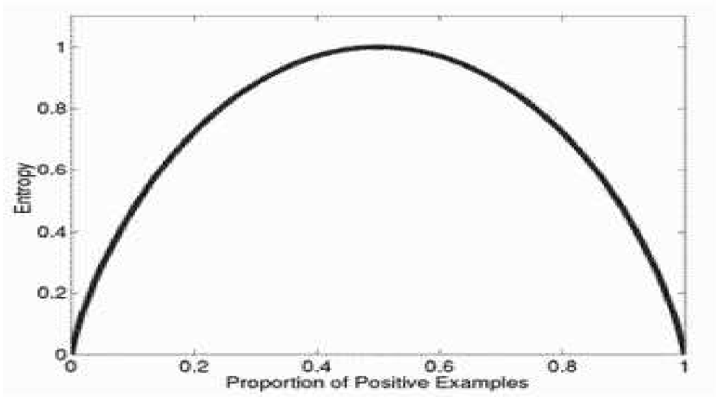


FIGURE 12.2 A graph of entropy, detailing how much information is available from finding out another piece of information given what you already know.

ID3(Iterative Dichotomiser 3)

- It generates decision tree from a dataset.
- The important idea is to work out how much the entropy of the whole training set would decrease if we choose each particular feature for the next classification step. This is known as the information gain
- Example: S = {s1 = true, s2 = false, s3 = false, s4 = false} and one feature F that can have values {f1, f2, f3}. The feature value for s1 could be f2, for s2 it could be f2, for s3, f3 and for s4, f1

S_1 T f_2

S_2 F f_2

S_3 F f_3

S_4 F f_1

$$\text{Gain}(S, F) = \text{Entropy}(S) - \sum_{f \in \text{values}(F)} \frac{|S_f|}{|S|} \text{Entropy}(S_f).$$

$-\sum_i p_i \log_2 p_i$ $f \in$ values(F)

$\log_2 p = \frac{\ln p}{\ln 2}$

$$\begin{aligned} \text{Entropy}(S) &= -p_{\oplus} \log_2 p_{\oplus} - p_{\ominus} \log_2 p_{\ominus} \\ &= -\frac{1}{4} \underbrace{\log_2 \frac{1}{4}}_{(-2)} - \frac{3}{4} \underbrace{\log_2 \frac{3}{4}}_{-0.415} \\ &= 0.5 + 0.311 = 0.811. \end{aligned}$$

S_1 + f_2
S_2 f f_2
S_3 f f_3
S_4 f f_1

$+ve$ $-ve$

$$\frac{|S_{f_1}|}{|S|}\text{Entropy}(S_{f_1}) = \frac{1}{4}\times\left(-\frac{0}{1}\log_2\frac{0}{1}-\frac{1}{1}\log_2\frac{1}{1}\right) = 0$$

$\ln 0.5/\ln 2$ $\ln 0.5/\ln 2$

$$\frac{|S_{f_2}|}{S}E(S_{f_2}) = \frac{2}{4}\times\left(-\frac{1}{2}\underbrace{\log_2\frac{1}{2}}_{(-1)} - \frac{1}{2}\underbrace{\log_2\frac{1}{2}}_{(-1)}\right) = 1/2$$

$$\frac{|S_{f_3}|}{|S|}E(S_{f_3}) = \frac{1}{4}\times\left[-\frac{0}{1}(\,) - \frac{1}{1}\log_2(1/1)\right] = 0.$$

$$\text{Gain}(S,F) = 0.811 - (0 + 0.5 + 0) = 0.311$$

Deadline?	Is there a party?	Lazy?	Activity
Urgent	Yes	Yes	Party
Urgent	No	Yes	Study
Near	Yes	Yes	Party
None	Yes	No	Party
None	No	Yes	Pub
None	Yes	No	Party
Near	No	No	Study
Near	No	Yes	TV
Near	Yes	Yes	Party
Urgent	No	No	Study

$E(S) = -P_P \log_2 P_P - P_S \log_2 P_S - P_{pub} \log_2 P_{pub} - P_{TV} \log_2 P_{TV}$

$= -\frac{5}{10}(\log_2 5/10) - 3/10 \log_2 3/10 - 1/10 \log_2 1/10 - 1/10 \log_2 1/10$

$= 0.5 + 0.5211 + 0.3322 \times 2 = 1.6855$

$G(S,F) = E(S) - \sum \frac{|S_f|}{|S|} E(S_f)$

f ∈ value of (F)

G(S,D), G(S,P), G(S,L)

$G(S, \text{Deadline}) = E(S) - \frac{|S_{urgent}|}{10} E(S_{urg}) - \frac{|S_{near}|}{10} E(S_{near}) - \frac{|S_{none}|}{10} E(S_{none})$

$= 1.6855 - \frac{3}{10}\left[-\frac{2}{3}\log_2 \frac{2}{3} - \frac{1}{3}\log_2 \frac{1}{3}\right] - \frac{4}{10}\left[-\frac{2}{4}\log_2 2/4 - 2\left(\frac{1}{4}\log_2 \frac{1}{4}\right)\right] - \frac{3}{10}\left[-\frac{2}{3}\log_2 2/3 - \frac{1}{3}\log_2 1/3\right]$

$= 1.6855 - (0.2755)2 - 0.6 = 0.5345$

F	Y	P
U	N	Y S
U	N	N S
Ne	Y	Y P
Ne	N	N S
Ne	N	Y TV
Ne	Y	Y P
None	Y	N P
N	N	Y Pub
No	Y	N P

$G(S, \text{Party}) = E(S) -$

S_1	t	f_2
S_2	f	f_2
S_3	f	f_3
S_4	f	f_1

$$Gain(S,F) = E(S) - \sum_{f \in V[F]} \frac{|S_f|}{|S|} E(S_f)$$

$$= E(S) - \frac{S_{f_1}}{S} E(S_{f_1}) - \frac{S_{f_2}}{S} E(S_{f_2}) - \frac{S_{f_3}}{S} E(S_{f_3})$$

$$\frac{|S_{f_1}|}{|S|}\text{Entropy}(S_{f_1}) = \frac{1}{4} \times \left(-\frac{0}{1}\log_2\frac{0}{1} - \frac{1}{1}\log_2\frac{1}{1}\right) = 0$$

$$\frac{2}{4}\left[-\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2}\right] = 0.5$$

0

0

$$= 0.811 - [0 + 0.5 + 0] = 0.311$$

S.V.JANSI RANI/Assoc Prof /CSE/SSNCE

$$
\begin{aligned}
\frac{|S_{f_1}|}{|S|}\text{Entropy}(S_{f_1}) &= \frac{1}{4} \times \left(-\frac{0}{1}\log_2\frac{0}{1} - \frac{1}{1}\log_2\frac{1}{1}\right) \\
&= 0 \\
\frac{|S_{f_2}|}{|S|}\text{Entropy}(S_{f_2}) &= \frac{2}{4} \times \left(-\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2}\right) \\
&= \frac{1}{2} \\
\frac{|S_{f_3}|}{|S|}\text{Entropy}(S_{f_3}) &= \frac{1}{4} \times \left(-\frac{0}{1}\log_2\frac{0}{1} - \frac{1}{1}\log_2\frac{1}{1}\right) \\
&= 0
\end{aligned}
$$

$$
\text{Gain}(S, F) = 0.811 - (0 + 0.5 + 0) = 0.311.
$$

Deadline?	Is there a party?	Lazy?	Activity
Urgent	Yes	Yes	Party
Urgent	No	Yes	Study
Near	Yes	Yes	Party
None	Yes	No	Party
None	No	Yes	Pub
None	Yes	No	Party
Near	No	No	Study
Near	No	Yes	TV
Near	Yes	Yes	Party
Urgent	No	No	Study

$E(S) = -P_P \log_2 P_P - P_S \log_2 P_S - P_{pub} \log_2 P_{pub} - P_{TV} \log_2 P_{TV}$

$= -\frac{5}{10} \log_2 \frac{5}{10} - \frac{3}{10} \log_2 \frac{3}{10} - \frac{1}{10} \log_2 \frac{1}{10} - \frac{1}{10} \log_2 \frac{1}{10}$

$= +0.5 + 0.5211 + 0.3322 + 0.3322 = 1.6855$

$\frac{\ln y}{\ln 2}$

$G(S, Deadline) = E(S) - \frac{|S_{urg}|}{|S|} E(S_{urg}) - \frac{|S_{near}|}{|S|} E(S_{near}) - \frac{|S_{none}|}{|S|} E(S_{none})$

$= 1.6855 - \frac{3}{10}\left[-\frac{1}{3} \log_2 \frac{1}{3} - \frac{2}{3} \log_2 \frac{2}{3}\right] - \frac{4}{10}\left[-\frac{2}{4} \log \frac{2}{4} - \frac{1}{4} \log \frac{1}{4} - \frac{1}{4} \log \frac{1}{4}\right] - \frac{3}{10}\left[\quad\right]$

$= 1.6855 - 0.$

$- 0.6 -$

$-\sum p_i \log_2 p_i$

$-P_+ \log_2 P_+ - P_- \log_2 P_-$

$Gain(S, Party) = 1.6855 - \frac{5}{10}\left[-\frac{3}{5} \log \frac{3}{5} - \frac{1}{5} \log \frac{1}{5} - \frac{1}{5} \log \frac{1}{5}\right] - \frac{5}{10}\left[-\frac{5}{5} \log_2 \frac{5}{5}\right]$

$= 1.6855 - 0.6855 - 0 = 1.0$

Deadline?	Is there a party?	Lazy?	Activity
Urgent	Yes	Yes	Party
Urgent	No	Yes	Study
Near	Yes	Yes	Party
None	Yes	No	Party
None	No	Yes	Pub
None	Yes	No	Party
Near	No	No	Study
Near	No	Yes	TV
Near	Yes	Yes	Party
Urgent	No	No	Study

$$G(S, Party) = E(S) - \frac{|S_y|}{|S|} E(S_y) - \frac{|S_n|}{|S|} E(S_n)$$

$$= 1.6855 - \frac{5}{10}\left[-\frac{5}{5}\log \frac{5}{5}\right] - \frac{5}{10}\left[-\frac{3}{5}\log \frac{3}{5} - 2\left(\frac{1}{5}\log \frac{1}{5}\right)\right]$$

$$= 1.6855 - 0 - 0.6855 = 1.0$$

Gain(S, Party)

Y P | N S
Y P | N Pb
Y P | N S
Y P | N TV
Y P | N S

$$G(S, Lazy) = E(S) - \frac{6}{10}\left[-\frac{3}{6}\log_2 \frac{3}{6} - 3\left(\frac{1}{6}\log \frac{1}{6}\right)\right] - \frac{4}{10}\left[-2\left(\frac{2}{4}\log \frac{2}{4}\right)\right]$$

$$= 1.6855 - 1.0755 - 0.4 = 0.21$$

$G(S, D) \rightarrow 0.5345$

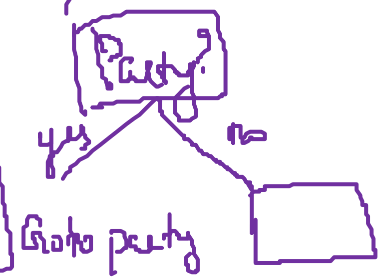
$G(S, P) - 1.0$

$G(S, L) - 0.21$

Lazy Ac

Y P N P
Y PS N P
Y P N S
Y Pb N S
Y TV
Y ? 3 P

1 S
1 Pb 1 TV



Deadline?	Is there a party?	Lazy?	Activity
Urgent	Yes	Yes	Party
Urgent	No	Yes	Study
Near	Yes	Yes	Party
None	Yes	No	Party
None	No	Yes	Pub
None	Yes	No	Party
Near	No	No	Study
Near	No	Yes	TV
Near	Yes	Yes	Party
Urgent	No	No	Study

Deadline?	Is there a party?	Lazy?	Activity
Urgent	Yes	Yes	Party
Urgent	No	Yes	Study
Near	Yes	Yes	Party
None	Yes	No	Party
Nonc	No	Ycs	Pub
None	Yes	No	Party
Near	No	No	Study
Near	No	Yes	TV
Near	Yes	Yes	Party
Urgent	No	No	Study

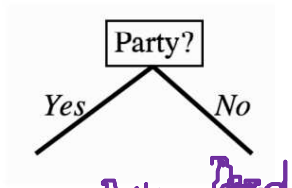


FIGURE 12.6 The decision tree after one step of the algorithm.

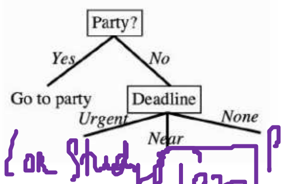


FIGURE 12.7 The tree after another step

Deadline?	Is there a party?	Lazy?	Activity
Urgent	No	Yes	Study
None	No	Yes	Pub
Near	No	No	Study
Near	No	Yes	TV
Urgent	No	~~Yes~~ No	Study

$$E(S) = \sum_i P_i \log_2 P_i = -\frac{3}{5}\log_2\frac{3}{5} - 2\left[\frac{1}{5}\log_2\frac{1}{5}\right] = 1.371$$

$$G(S, \text{Deadline}) = E(S) - \underset{\text{Urg}}{\_\_} \; \underset{\text{Near}}{\_\_} \; \underset{\text{None}}{\_\_}$$

$$= 1.371 - \frac{2}{5}\left[-\frac{2}{2}\log_2\frac{2}{2}\right] - \frac{2}{5}\left[2\left(-\frac{1}{2}\log_2 \tfrac{1}{2}\right)\right] - \frac{1}{5}\left(-\frac{1}{1}\log_2\frac{1}{1}\right)$$

$$= 1.371 - 0.4 = 0.971$$

$$G(S, \text{Lazy}) = 1.371 - \frac{3}{5}\left(\left(-\frac{1}{3}\log_2 \tfrac{1}{3}\right)3\right) - \frac{2}{5}\left(-\frac{2}{2}\log_2 \tfrac{2}{2}\right)$$

$$= 1.371 - 0.9509 = 0.42$$

Classification Example

Deadline?	Is there a party?	Lazy?	Activity
Urgent	Yes	Yes	Party
Urgent	No	Yes	Study
Near	Yes	Yes	Party
None	Yes	No	Party
None	No	Yes	Pub
None	Yes	No	Party
Near	No	No	Study
Near	No	Yes	TV
Near	Yes	Yes	Party
Urgent	No	No	Study

$$
\begin{aligned}
\text{Entropy}(S) &= -p_{\text{party}} \log_2 p_{\text{party}} - p_{\text{study}} \log_2 p_{\text{study}} \\
&\quad - p_{\text{pub}} \log_2 p_{\text{pub}} - p_{\text{TV}} \log_2 p_{\text{TV}} \\
&= -\frac{5}{10}\log_2\frac{5}{10} - \frac{3}{10}\log_2\frac{3}{10} - \frac{1}{10}\log_2\frac{1}{10} - \frac{1}{10}\log_2\frac{1}{10} \\
&= 0.5 + 0.5211 + 0.3322 + 0.3322 = 1.6855
\end{aligned}
$$

$$
\begin{aligned}
\text{Gain}(S, \text{Deadline}) &= 1.6855 - \frac{|S_{\text{urgent}}|}{10}\text{Entropy}(S_{\text{urgent}}) \\
&- \frac{|S_{\text{near}}|}{10}\text{Entropy}(S_{\text{near}}) - \frac{|S_{\text{none}}|}{10}\text{Entropy}(S_{\text{none}}) \\
&= 1.6855 - \frac{3}{10}\left(-\frac{2}{3}\log_2\frac{2}{3} - \frac{1}{3}\log_2\frac{1}{3}\right) \\
&- \frac{4}{10}\left(-\frac{2}{4}\log_2\frac{2}{4} - \frac{1}{4}\log_2\frac{1}{4} - \frac{1}{4}\log_2\frac{1}{4}\right) \\
&- \frac{3}{10}\left(-\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3}\right) \\
&= 1.6855 - 0.2755 - 0.6 - 0.2755 \\
&= 0.5345
\end{aligned}
$$

$$
\begin{aligned}
\text{Gain}(S, \text{Party}) &= 1.6855 - \frac{5}{10}\left(-\frac{5}{5}\log_2\frac{5}{5}\right) \\
&- \frac{5}{10}\left(-\frac{3}{5}\log_2\frac{3}{5} - \frac{1}{5}\log_2\frac{1}{5} - \frac{1}{5}\log_2\frac{1}{5}\right) \\
&= 1.6855 - 0 - 0.6855 \\
&= 1.0 \qquad (
\end{aligned}
$$

$$
\begin{aligned}
\text{Gain}(S, \text{Lazy}) &= 1.6855 - \frac{6}{10}\left(-\frac{3}{6}\log_2\frac{3}{6} - \frac{1}{6}\log_2\frac{1}{6} - \frac{1}{6}\log_2\frac{1}{6} - \frac{1}{6}\log_2\frac{1}{6}\right) \\
&- \frac{4}{10}\left(-\frac{2}{4}\log_2\frac{2}{4} - \frac{2}{4}\log_2\frac{2}{4}\right) \\
&= 1.6855 - 1.0755 - 0.4 \\
&= 0.21 \qquad (
\end{aligned}
$$

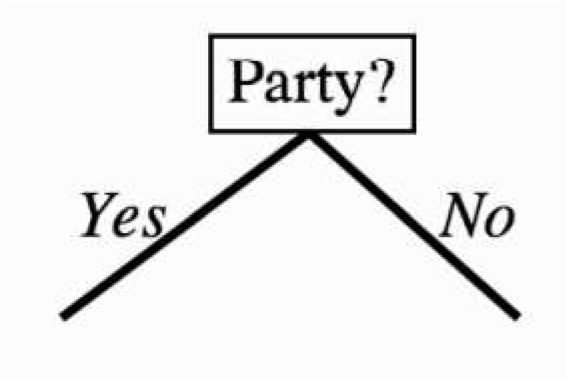


FIGURE 12.6 The decision tree after one step of the algorithm.

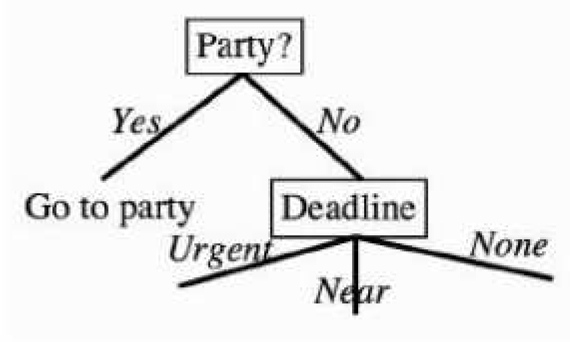


FIGURE 12.7 The tree after another step.

Deadline?	Is there a party?	Lazy?	Activity
Urgent	No	Yes	Study
None	No	Yes	Pub
Near	No	No	Study
Near	No	Yes	TV
Urgent	No	~~Yes~~ No	Study

$$
\begin{aligned}
\text{Gain}(S, \text{Deadline}) &= 1.371 - \frac{2}{5}\left(-\frac{2}{2}\log_2\frac{2}{2}\right) \\
&- \frac{2}{5}\left(-\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2}\right) - \frac{1}{5}\left(-\frac{1}{1}\log_2\frac{1}{1}\right) \\
&= 1.371 - 0 - 0.4 - 0 \\
&= 0.971 \\
\text{Gain}(S, \text{Lazy}) &= 1.371 - \frac{4}{5}\left(-\frac{2}{4}\log_2\frac{2}{4} - \frac{1}{4}\log_2\frac{1}{4} - \frac{1}{4}\log_2\frac{1}{4}\right) \\
&- \frac{1}{5}\left(-\frac{1}{1}\log_2\frac{1}{1}\right) \\
&= 1.371 - 1.2 - 0 \\
&= 0.1710
\end{aligned}
$$

Wrong

book

→ Calculating info. gain:

$$G(s, \text{deadline}) = 1.371 - \overset{\swarrow \text{urgent}}{\frac{2}{5}}\left(-\frac{2}{2}\log_2\frac{2}{2}\right)$$

$$\text{near} \longrightarrow -\frac{2}{5}\left(-\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2}\right)$$

$$\text{none} \longrightarrow -\frac{1}{5}\left(-\frac{1}{1}\log_2\frac{1}{1}\right)$$

$$= 1.371 - 0 - 0.4 - 0$$

$$\therefore G(s, \text{deadline}) = 0.971$$

$$G(s, \text{lazy}) = 1.371 - \overset{\swarrow \text{yes}}{\frac{3}{5}}\left(-\frac{1}{3}\log_2\frac{1}{3} - \frac{1}{3}\log_2\frac{1}{3} - \frac{1}{3}\log_2\frac{1}{3}\right)$$

$$\text{no} \longrightarrow -\frac{2}{5}\left(\frac{2}{2}\log_2\frac{2}{2}\right)$$

$$= 1.371 - 0.9509 - 0$$

$$\therefore G(s, \text{lazy}) = 0.42$$

The ID3 Algorithm

- If all examples have the same label:
  - return a leaf with that label
- Else if there are no features left to test:
  - return a leaf with the most common label
- Else:
  - choose the feature $\hat{F}$ that maximises the information gain of S to be the next node using Equation (12.2)
  - add a branch from the node for each possible value f in $\hat{F}$
  - for each branch:
    - calculate S_f by removing $\hat{F}$ from the set of features
    - recursively call the algorithm with S_f, to compute the gain relative to the current set of examples

Dealing with Continuous variables

- So far dealt with discrete or categorical values
- For a continuous variable there is not just one place to split it: the variable can be broken between any pair of datapoints

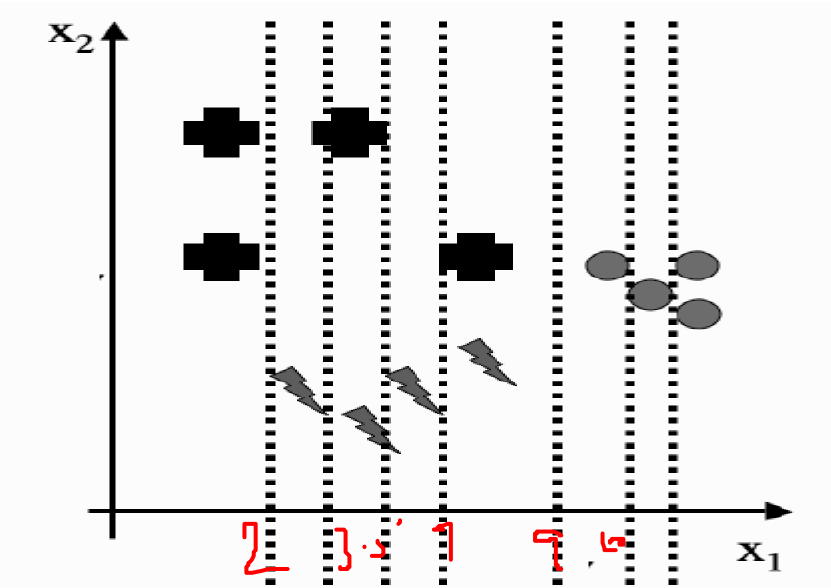


FIGURE 12.3 Possible places to split the variable x_1, between each of the datapoints as the feature value increases.

- The trees -univariate trees, because they pick one feature (dimension) at a time and split according to that one.
- There are also algorithms that make multivariate trees by picking combinations of features.
- This can make for considerably smaller trees if it is possible to find straight lines that separate the data well, but are not parallel to any axis.

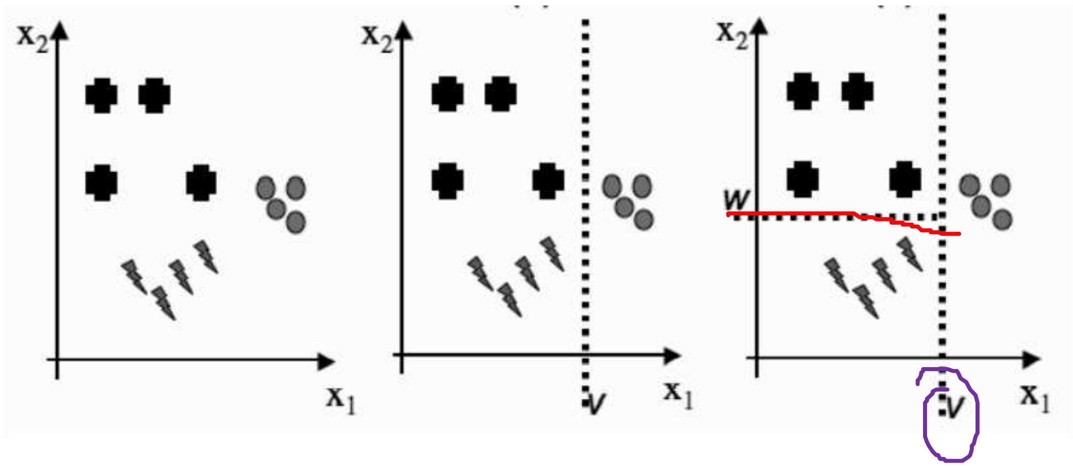


FIGURE 12.4 The effect of decision tree choices. The two-dimensional dataset shown in (a) is split first by choosing feature x_1 (b) and then x_2, (c) which separates out the three classes. The final tree is shown in Figure 12.5.

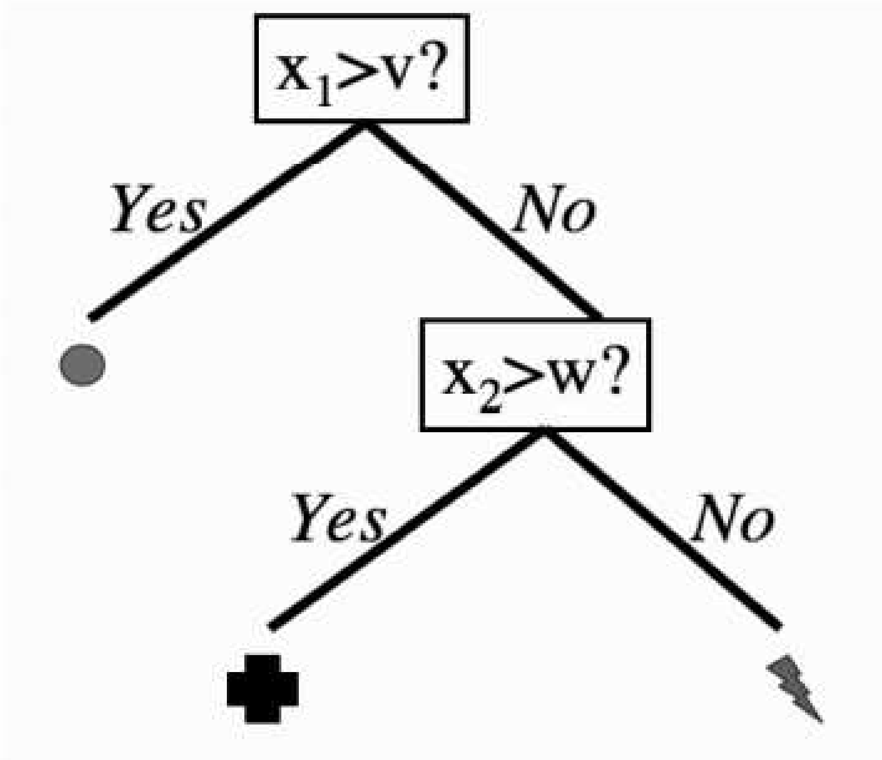
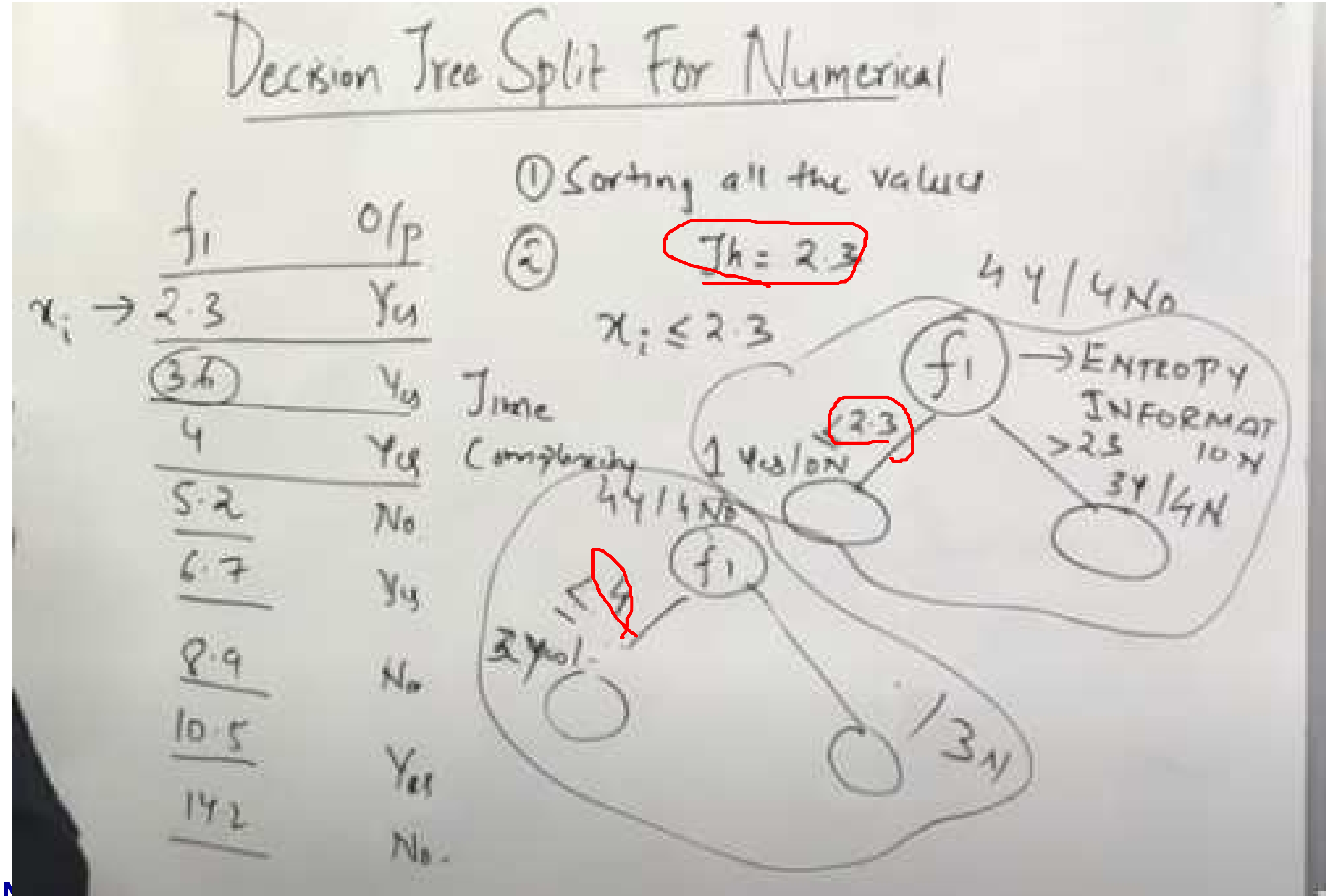


FIGURE 12.5 The final tree created by the splits in Figure 12.4.

Numerical - time complexity is more



Computational Complexity

- Tree is approximately balanced, then the cost at each node consists of searching through the d possible features (although this decreases by 1 at each level, that doesn't affect the complexity in the O(·) notation) and then computing the information gain for the dataset for each split.
- This has cost O(dn log n), where n is the size of the dataset at that node.
- For the root, n = N, and if the tree is balanced, then n is divided by 2 at each stage down the tree. Summing this over the approximately logN levels in the tree gives computational cost **O(dN2 logN)**.

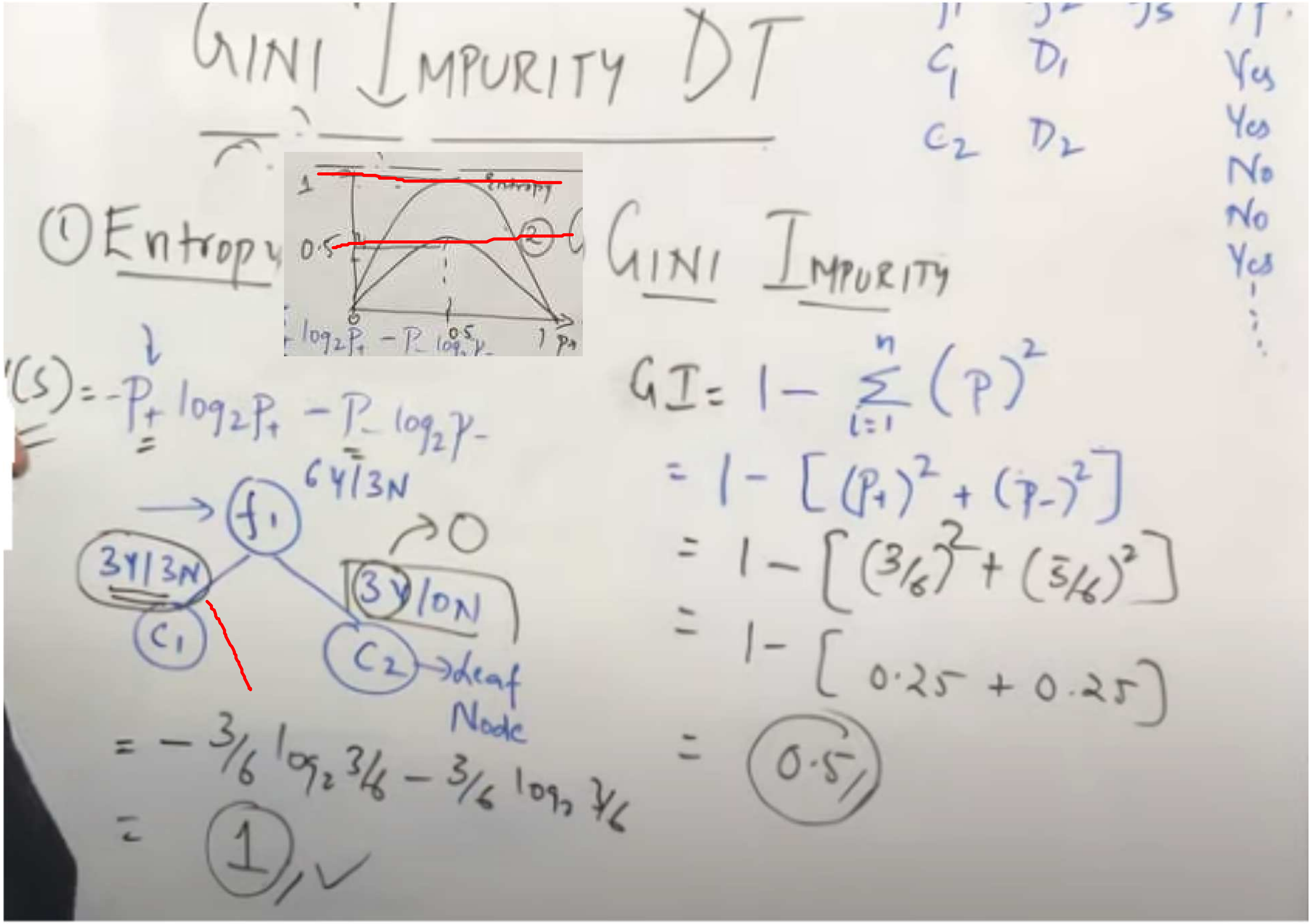
Gini Impurity

$$G_k = \sum_{i=1}^{c} \sum_{j \neq i} N(i)N(j), \tag{12.8}$$

where c is the number of classes. In fact, you can reduce the algorithmic effort required by noticing that $\sum_i N(i) = 1$ (since there has to be some output class) and so $\sum_{j \neq i} N(j) = 1 - N(i)$. Then Equation (12.8) is equivalent to:

$$G_k = 1 - \sum_{i=1}^{c} N(i)^2. \tag{12.9}$$

- Computationaly efficient


GINI IMPURITY DT
C1 D1 Yes
C2 D2 Yes
No
No
Yes
① Entropy
② GINI IMPURITY
Entropy
1
0.5
0
0.5
1
H(S) = -P+ log2 P+ - P- log2 P-
GI = 1 - Σ (p)^2
= 1 - [(P+)^2 + (P-)^2]
= 1 - [(3/6)^2 + (3/6)^2]
= 1 - [0.25 + 0.25]
= 0.5
6Y/3N
f1
3Y/3N
3Y/0N
C1
C2
Leaf Node
= -3/6 log2 3/6 - 3/6 log2 3/6
= 1

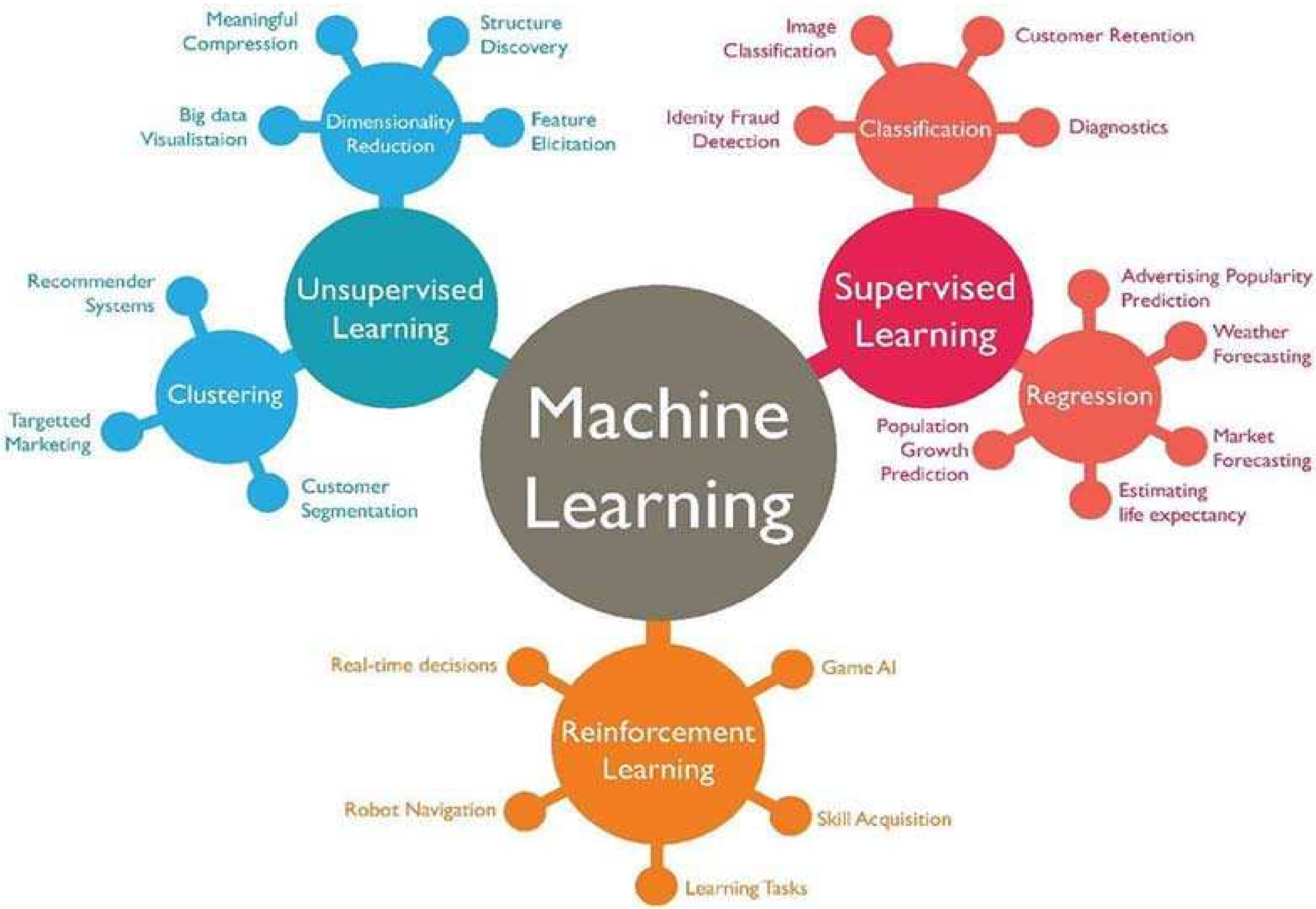
Exercise Problem

Problem 12.3 Turn this politically incorrect data from Quinlan into a decision tree to classify which attributes make a person attractive, and then extract the rules.

Height	Hair	Eyes	Attractive?
Small	Blonde	Brown	No
Tall	Dark	Brown	No
Tall	Blonde	Blue	Yes
Tall	Dark	Blue	No
Small	Dark	Blue	No
Tall	Red	Blue	Yes
Tall	Blonde	Brown	No
Small	Blonde	Blue	Yes

Check Your Understanding

- What is the disadvantage of classifying numerical values
- Why do we need gini impurity?


Machine Learning
Unsupervised Learning
Dimensionality Reduction
Meaningful Compression
Structure Discovery
Big data Visualistaion
Feature Elicitation
Clustering
Recommender Systems
Targetted Marketing
Customer Segmentation
Supervised Learning
Classification
Image Classification
Customer Retention
Idenity Fraud Detection
Diagnostics
Regression
Advertising Popularity Prediction
Weather Forecasting
Population Growth Prediction
Market Forecasting
Estimating life expectancy
Reinforcement Learning
Real-time decisions
Game AI
Robot Navigation
Skill Acquisition
Learning Tasks

Summary

- Constructing decision trees
- Information Gain using ID3
- Gini Impurity
- Classification Problem

techNews Time

Who is ready today?

THANK YOU

Courtsey : Stephen Marsland