Decision Trees

Learning outcomes

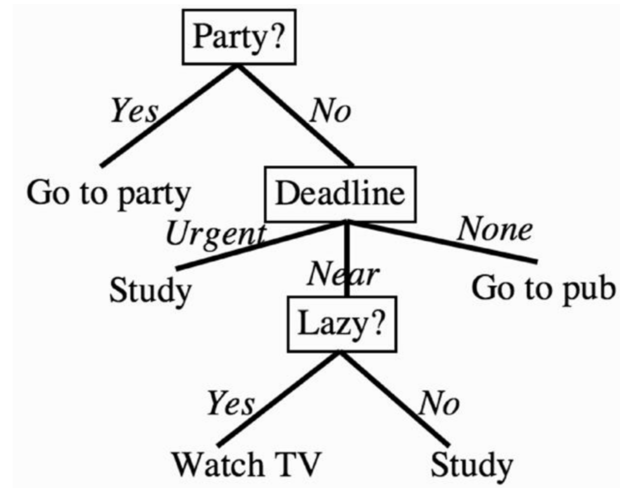
- Define the entropy of a set
- Compute the entropy of a given set
- Define the information gain for a given feature
- Define the Gini Impurity of a set
- Implement the ID3 and CART algorithms

Making Decisions

UNIVERSITY OF LEEDS

Differently from NNs, and most other ML methods, decision trees can work with non-metric data, that is, features that are not numbers.

Building a decision tree amounts to deciding the order in which the splits should be made.



How to choose the variable for each split?

Entropy and Information

UNIVERSITY OF LEEDS

Information: amount of a 'surprise' in an outcome with probability of p_i :

$$I_i = \log_2(\frac{1}{p_i}) = -\log_2(p_i)$$

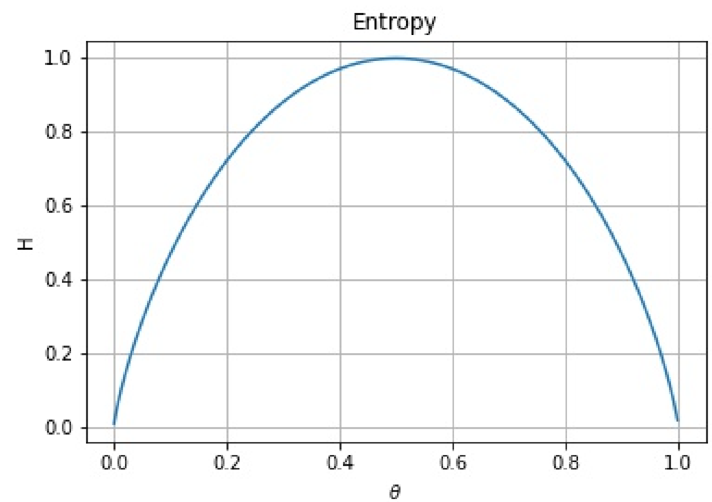
Entropy: expected amount of 'surprise' in a probability distribution:

$$H = E[I] = \sum_i -p_i \log_2(p_i)$$

Example: For a binary random variable

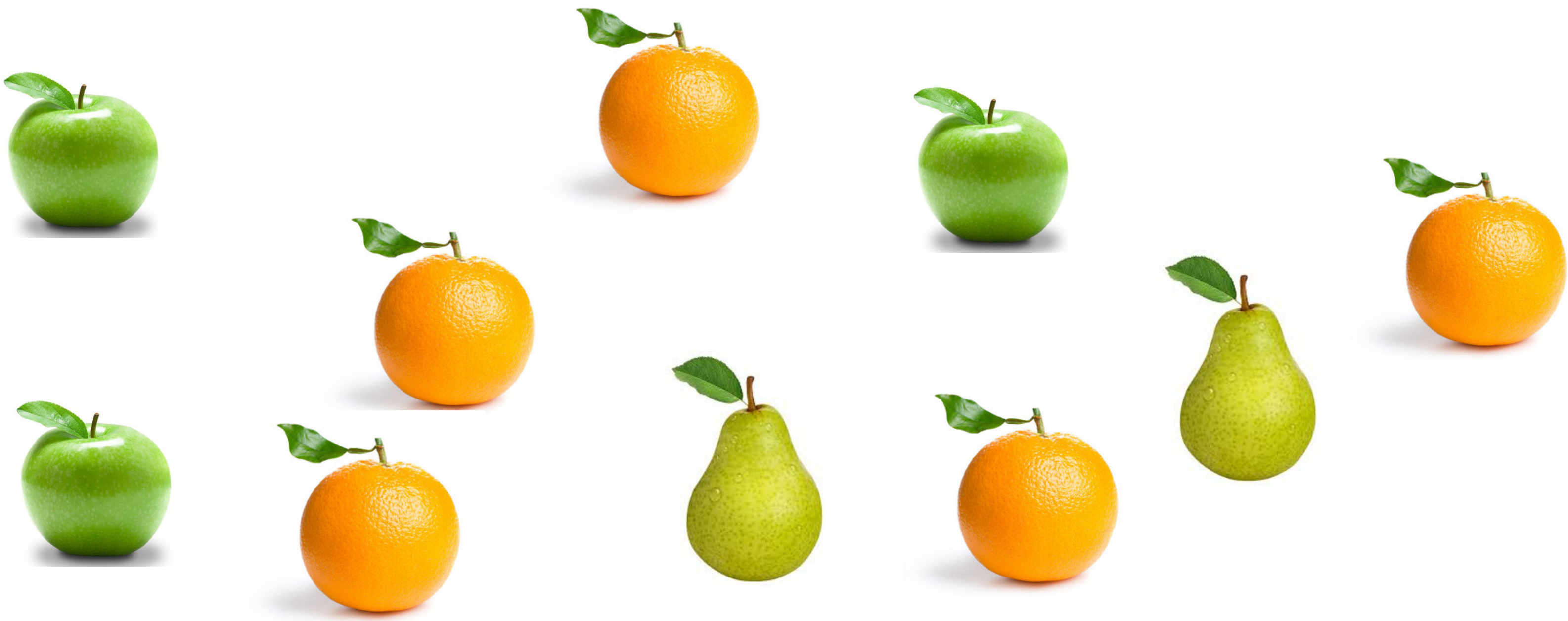
$$p(x = Head) = \theta$$

$$p(x = Tail) = 1 - \theta$$



Oranges & apples & pears

$$H = E[I] = \sum_{i \in classes} -p_i \log_2 p_i$$



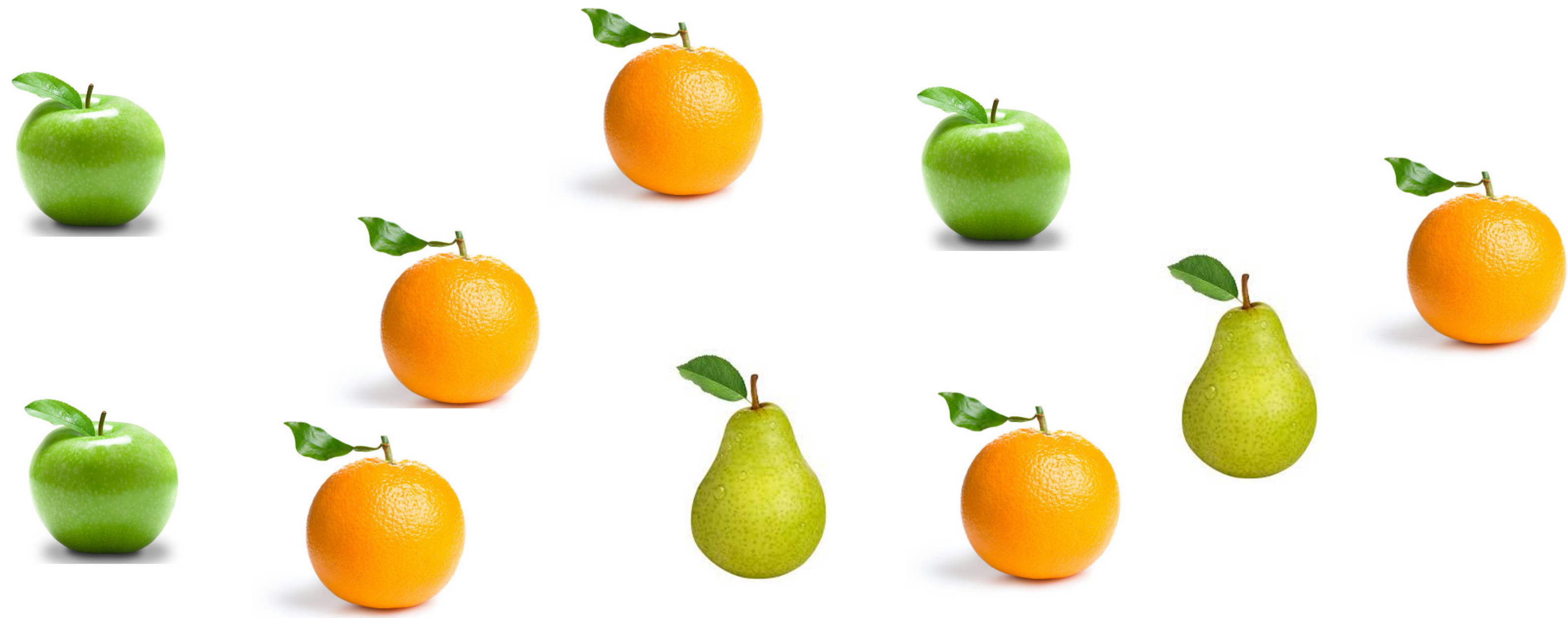
$H = ?$

$$\log_a(x) = \frac{\log_b(x)}{\log_b(a)}$$

Oranges & apples & pears

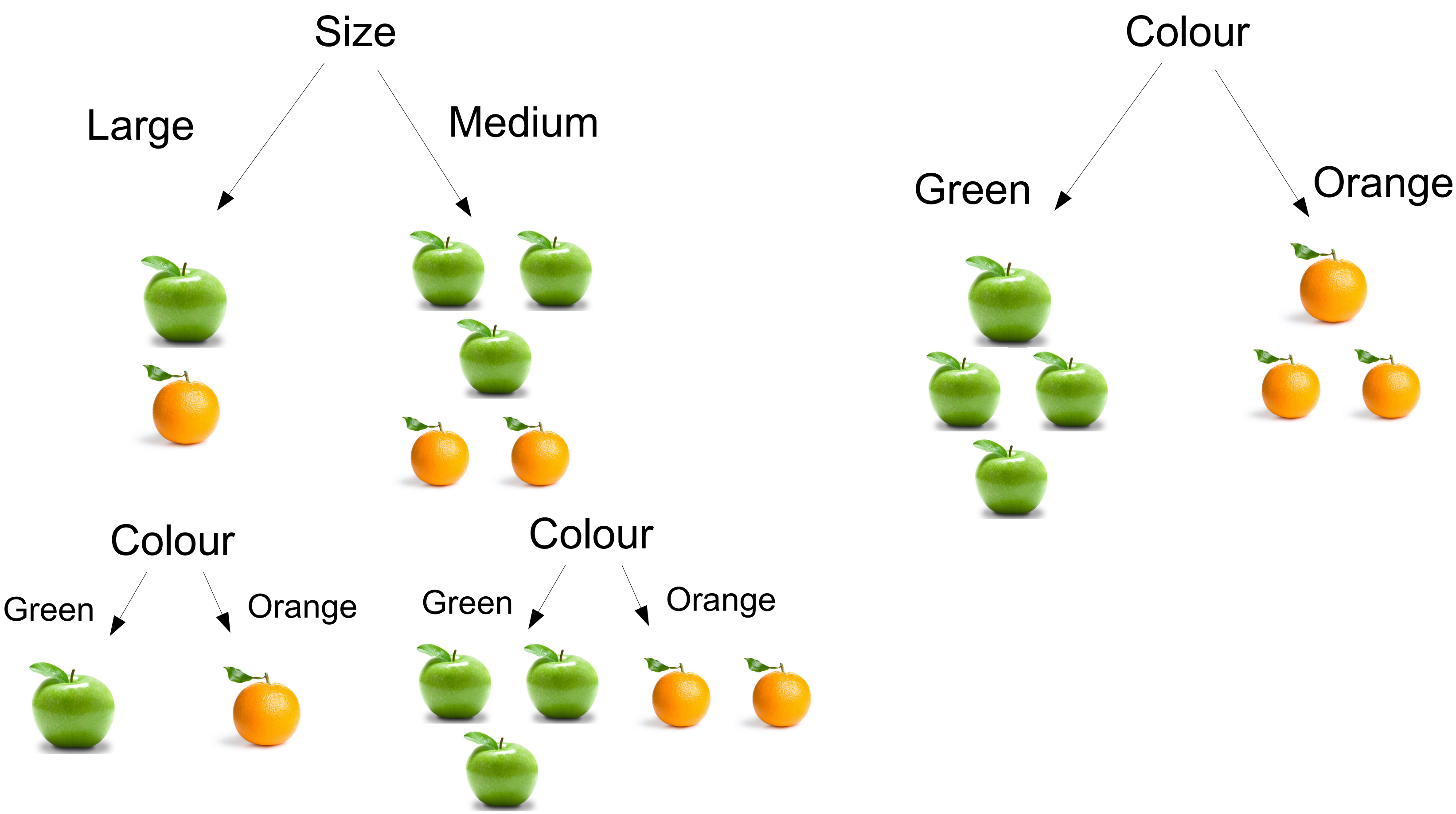
UNIVERSITY OF LEEDS

$$\mathrm{H} = E[I] = \sum_{i \in classes} -p_i \log_2 p_i$$



$$\mathrm{H} = \overbrace{-\frac{3}{10}\log_2\frac{3}{10}}^{\text{Apples}} \overbrace{-\frac{5}{10}\log_2\frac{5}{10}}^{\text{Oranges}} \overbrace{-\frac{2}{10}\log_2\frac{2}{10}}^{\text{Pears}} = 1.485$$

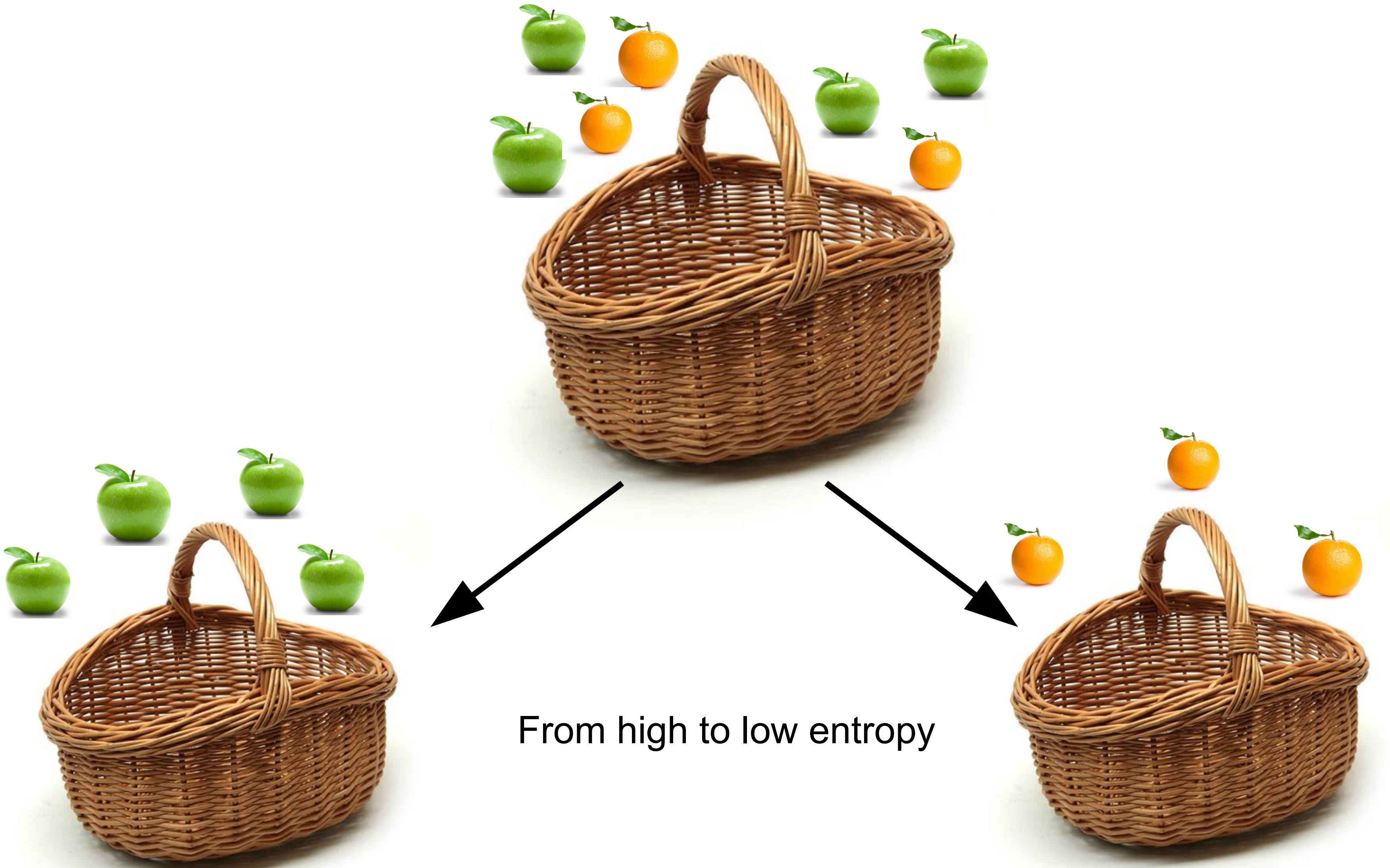
Splitting oranges and apples



Entropy of the set

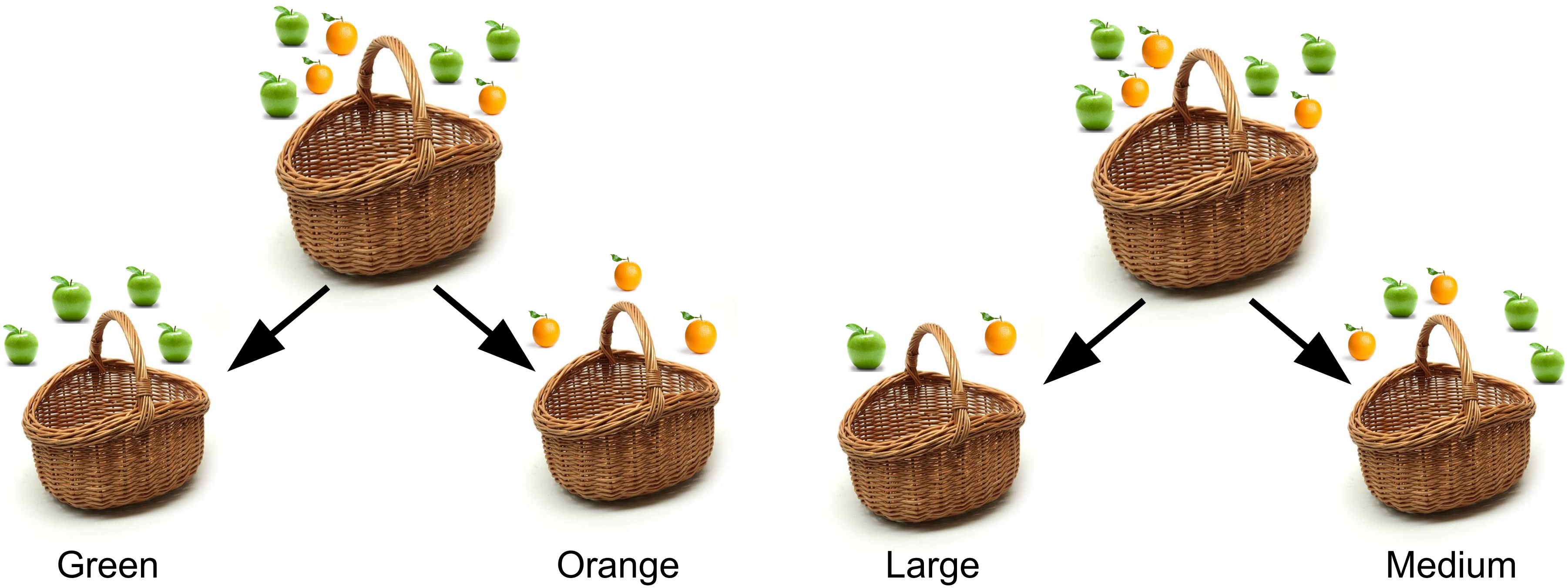
$$H = -p_O \log_2(p_O) - p_A \log_2(p_A) = -\frac{3}{7}\log_2\left(\frac{3}{7}\right) - \frac{4}{7}\log_2\left(\frac{4}{7}\right) = 0.985$$

Entropy of the set



Entropy of the set

UNIVERSITY OF LEEDS



$$H_{\text{colour}} = \overbrace{\frac{4}{7}}^{\text{fraction in Green}} \overbrace{(0)}^{\text{entropy of Green}} + \overbrace{\frac{3}{7}}^{\text{fraction in Orange}} \overbrace{(0)}^{\text{entropy of Orange}} = 0$$

$$H_{\text{size}} = \overbrace{\frac{2}{7}}^{\text{fraction in Large}} \overbrace{\left(-\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2}\right)}^{\text{entropy of Large}} + \overbrace{\frac{5}{7}}^{\text{fraction in Medium}} \overbrace{\left(-\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5}\right)}^{\text{entropy of Medium}} = 0.98$$

Entropy of the set

$$H = 0.985$$

$$H_{colour} = 0$$

$$G(\text{Colour}) = H - H_{colour} = 0.985$$

$$H_{size} = 0.98$$

$$G(\text{Size}) = H - H_{size} = 0.005$$

Information gain

UNIVERSITY OF LEEDS

Set of elements

elements in S with feature F = f

$$G(S,F)=\mathrm{H}(S)-\sum_{f\in values(F)}\frac{|S_f|}{|S|}\mathrm{H}(S_f)$$

Feature

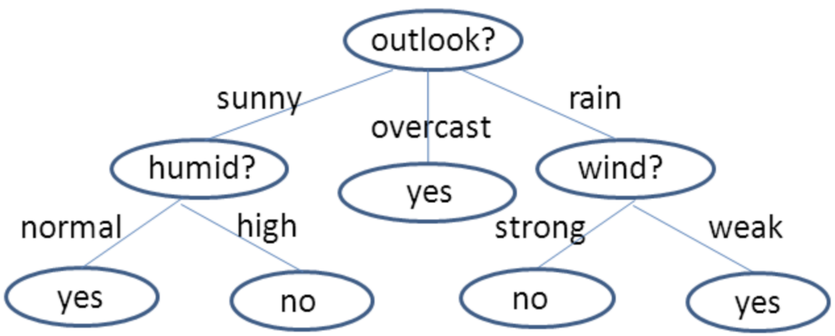
compare with:

$$\mathrm{H}_{\mathrm{size}}=\overbrace{\frac{2}{7}}^{\text{fraction in Large}}\overbrace{\left(-\frac{1}{2}\log_2\frac{1}{2}-\frac{1}{2}\log_2\frac{1}{2}\right)}^{\text{entropy of Large}}+\overbrace{\frac{5}{7}}^{\text{fraction in Medium}}\overbrace{\left(-\frac{2}{5}\log_2\frac{2}{5}-\frac{3}{5}\log_2\frac{3}{5}\right)}^{\text{entropy of Medium}}=0.98$$

The ID3 algorithm

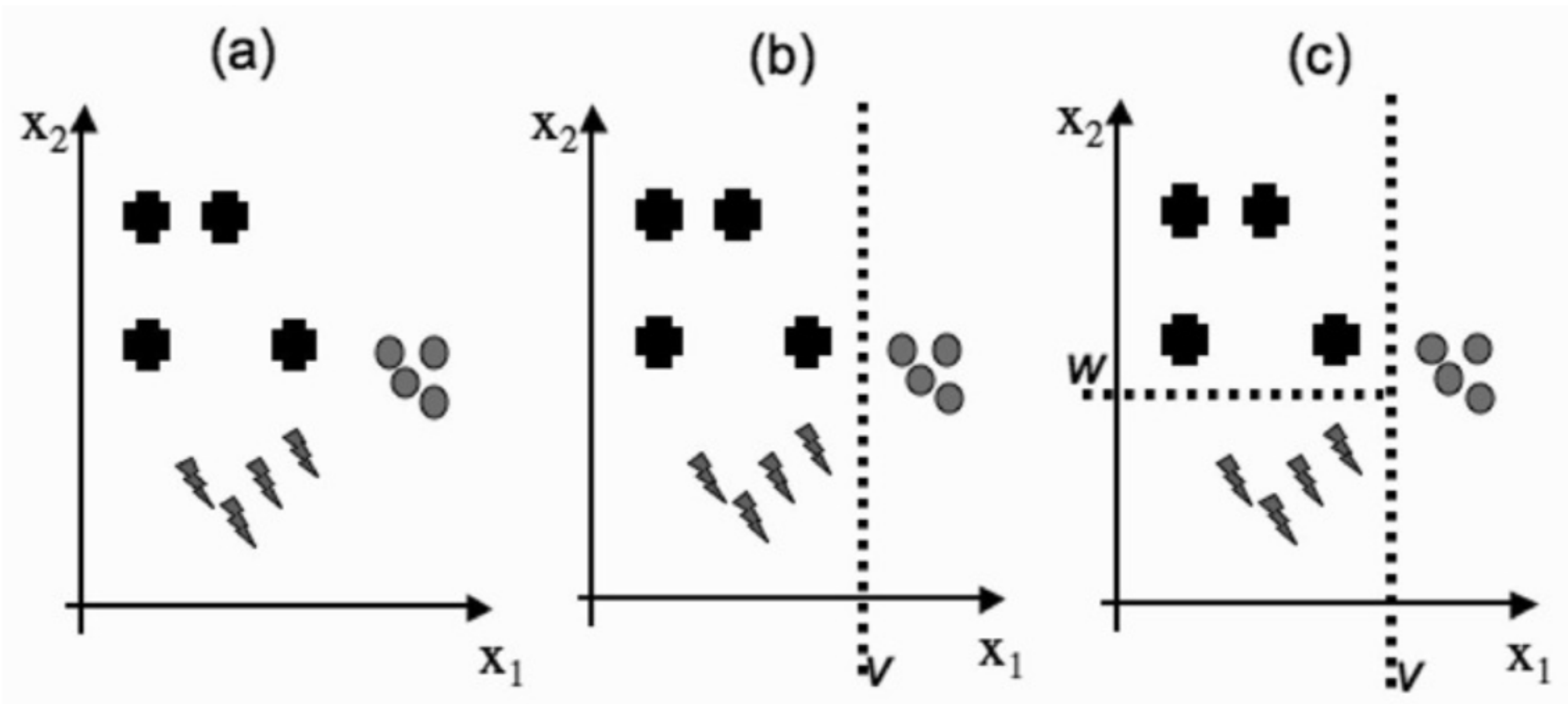
- It computes the information gain for each feature and chooses the one that produces the highest value.

- The output is the decision tree, with the decisions given in the leaves.



play	outlook	temp	humid	wind	day	moon
no	sunny	hot	high	weak	tuesday	full
no	sunny	hot	high	strong	tuesday	small
yes	overcast	hot	high	weak	tuesday	full
yes	rainy	mild	high	strong	tuesday	small
yes	rainy	cool	normal	weak	monday	full
no	rainy	cool	normal	strong	monday	small
yes	overcast	cool	normal	strong	monday	full
no	sunny	mild	high	weak	monday	small
yes	sunny	cool	normal	weak	monday	full
yes	rainy	mild	normal	weak	monday	small
yes	sunny	mild	normal	strong	friday	full
no	rainy	cool	normal	strong	friday	small
yes	overcast	cool	normal	strong	friday	full
no	sunny	mild	high	weak	friday	small
yes	sunny	cool	normal	weak	friday	full
yes	overcast	hot	high	weak	friday	small
yes	rainy	mild	high	strong	friday	full
yes	rainy	mild	normal	weak	friday	small
yes	sunny	mild	normal	strong	tuesday	full

Visualizing splits



Characteristics of ID3:

UNIVERSITY OF LEEDS

- Does not capture the curvature in the decision boundries
- Greedy with respect to G (chooses max gain)→ potential local minimum
- Deals with noisy data (by assigning the label to most common class)
- Always uses all the features: Prone to overfitting (pruning is applied).
- Does not work with continuous variables (C4.5 algorithm does)

A Different Criterion: Gini Impurity

		Colour	
		Green	Orange
Size	Large	P A P P	O
	Medium	A A A	O O

$$G(S) = \overbrace{\frac{4}{10}}^{apples}\overbrace{\left(\frac{3}{10}+\frac{3}{10}\right)}^{not\,apples} + \overbrace{\frac{3}{10}}^{pears}\overbrace{\left(\frac{4}{10}+\frac{3}{10}\right)}^{not\,pears} + \overbrace{\frac{3}{10}}^{oranges}\overbrace{\left(\frac{4}{10}+\frac{3}{10}\right)}^{not\,oranges} = \sum_{i}^{C} p_i(1-p_i)$$

Gini split:

$$G(S)=\sum_{i}^{C} p_i(1-p_i)=\overbrace{\sum_{i}^{C}}^{\text{\# of classes}}(p_i-p_i^2)=\sum_{i}^{C} p_i-\sum_{i}^{C} p_i^2=1-\sum_{i}^{C} p_i^2$$

$$G(S,F)=G(S)-\sum_{f\in values(F)}\frac{|S_f|}{|S|}G(S_f)$$

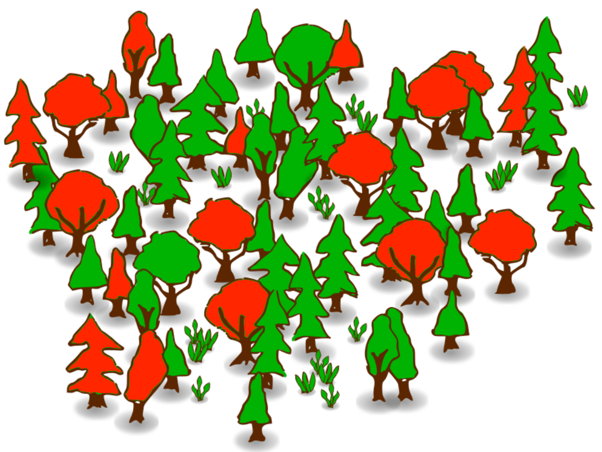
If we use the same definition of the information gain but with the Gini impurity instead of entropy we obtain the splitting principle of another algorithm: CART.
CART splits on the variable with the highest Gini gain, as opposed to information gain.

Random forests

UNIVERSITY OF LEEDS

Trees are often used in collections called (quite appropriately) forests. Each tree is obtained by introducing randomness in some aspects, the most commons are:

- bagging, a random subset of the data is used
- a random subset of the features is used.



Each tree “votes” for a classification, and the classification with most votes is returned by the forest.

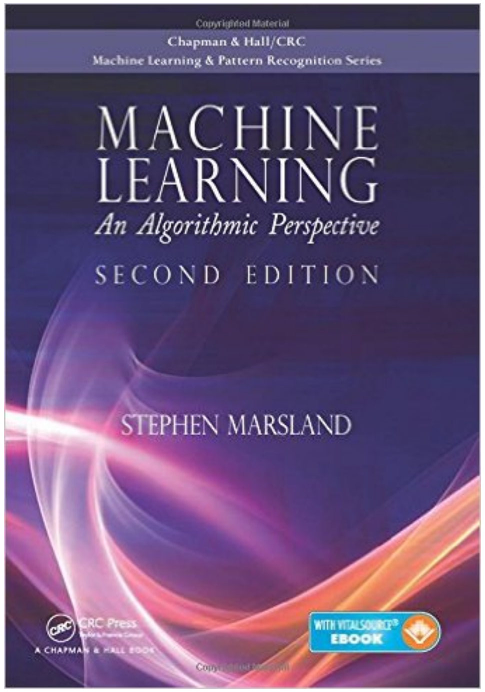
UNIVERSITY OF LEEDS

Conclusion

Learning outcomes

UNIVERSITY OF LEEDS

- Define the entropy of a set
- Compute the entropy of a given set
- Define the information gain for a given feature
- Define the Gini Impurity of a set
- Implement the ID3 and CART algorithms


Copyrighted Material
Chapman & Hall/CRC
Machine Learning & Pattern Recognition Series
MACHINE LEARNING
An Algorithmic Perspective
SECOND EDITION
STEPHEN MARSLAND
CRC Press
A CHAPMAN & HALL BOOK
WITH VITALSOURCE® EBOOK

Chapter 12