UNIVERSITY OF LEEDS

Class: Machine Learning

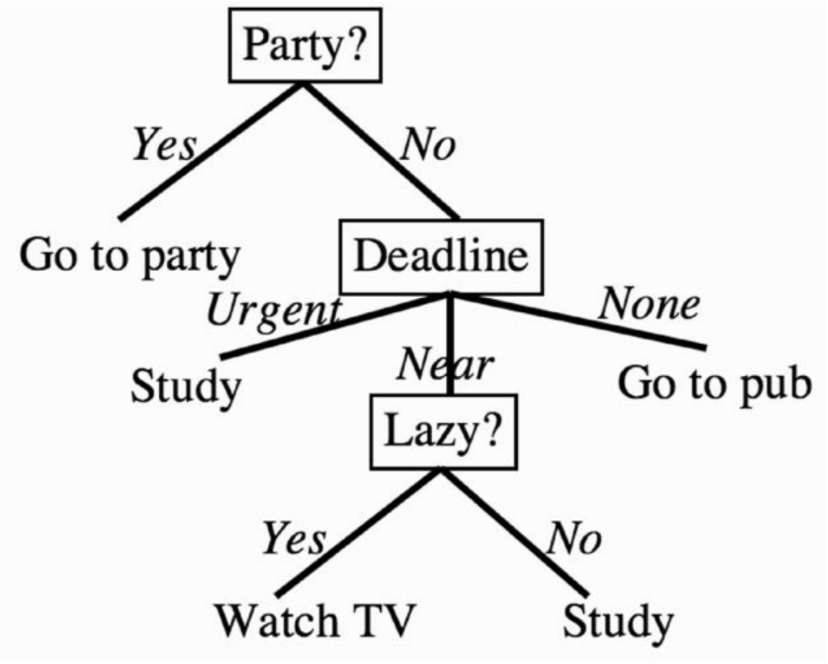
Decision Trees

Instructor: Matteo Leonetti

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



Nonmetric data

How to choose the variable for each split?

History

1983 - Ross Quinlan (U. of Sidney)

Learning efficient classification procedures and their application to chess end games.

Entropy and information

How much information do I receive, with a message X?

X a random variable over possible messages

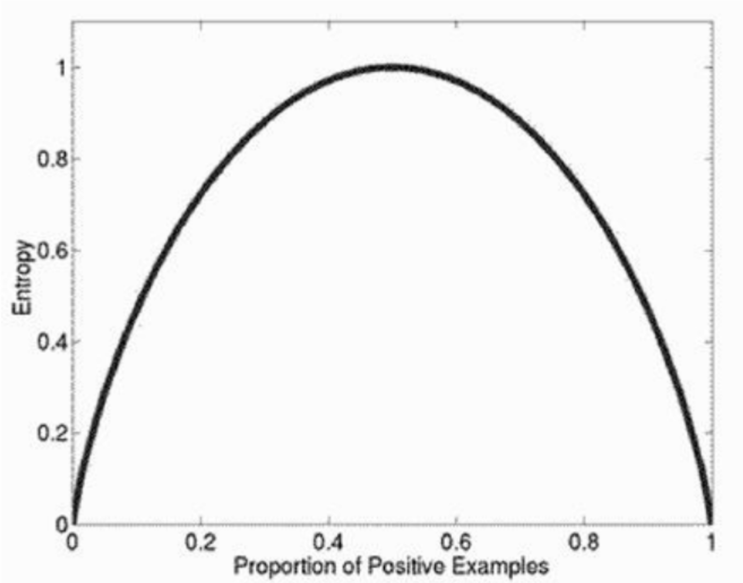
Information

$$I(x) = -\log_2 P(x)$$

Entropy

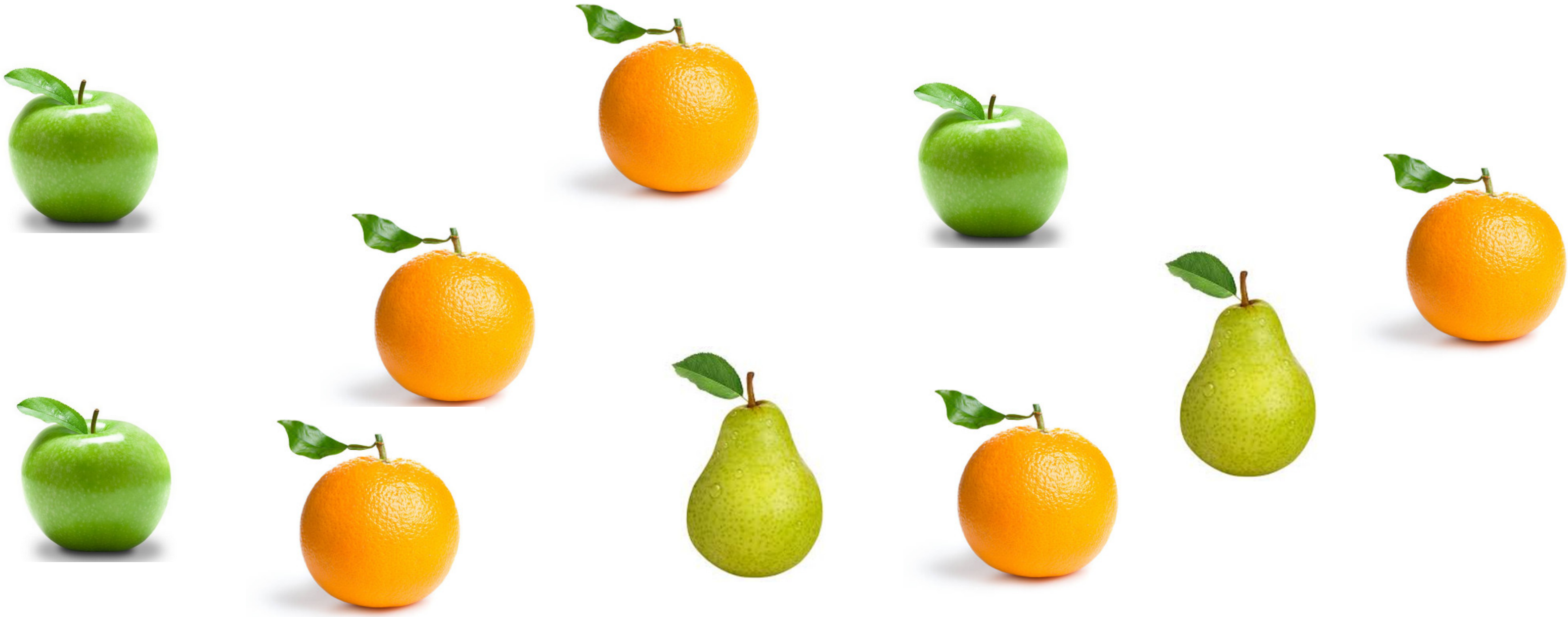
$$\mathrm{H} = E[I] = \sum_i -p_i \log_2 p_i$$

$$0 \log_2 0 = 0$$



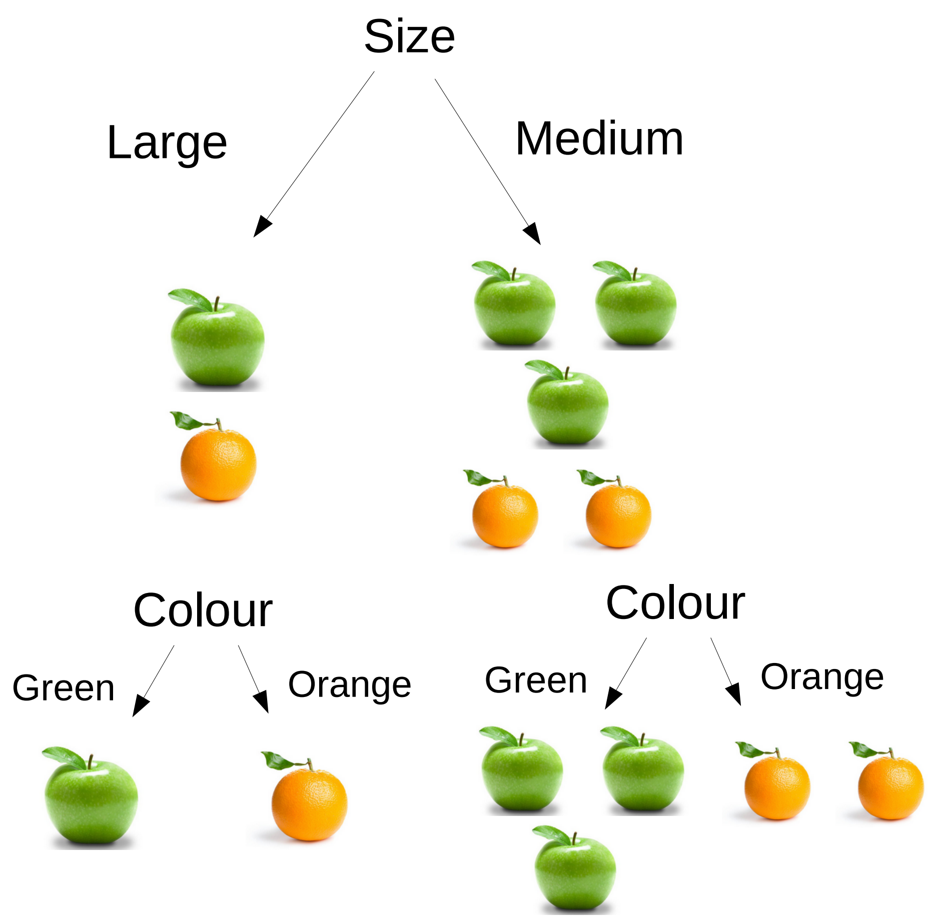
Fruits

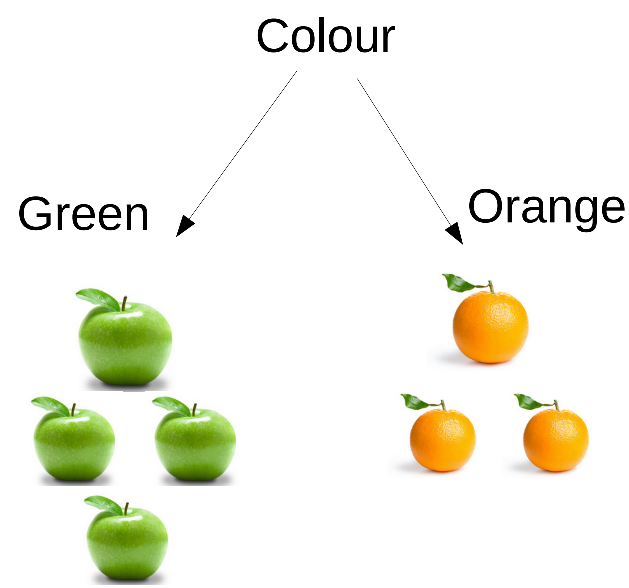
$$H = E[I] = \sum_{i \in classes} -p_i \log_2 p_i$$



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

Apples and Oranges

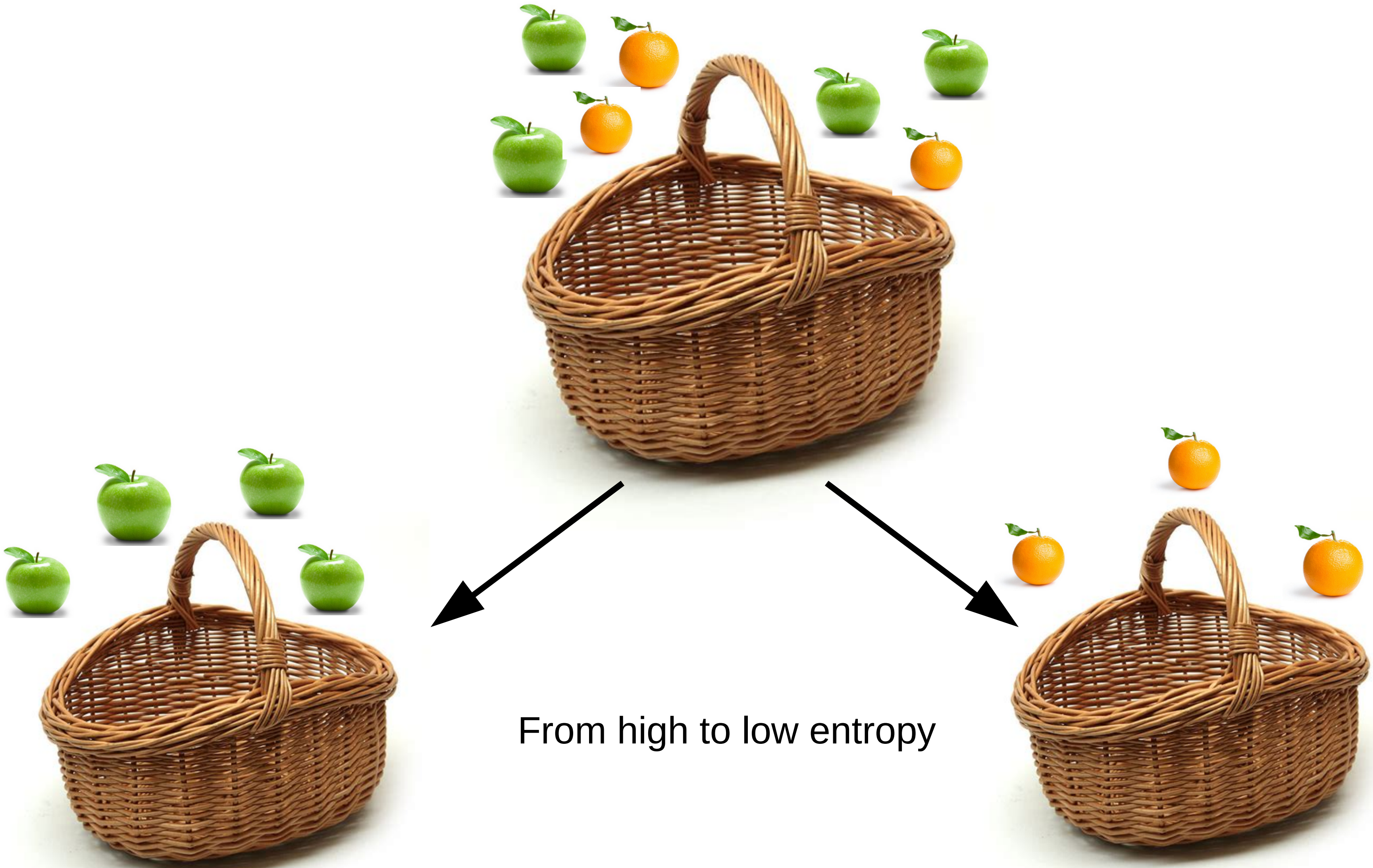




Entropy of the set

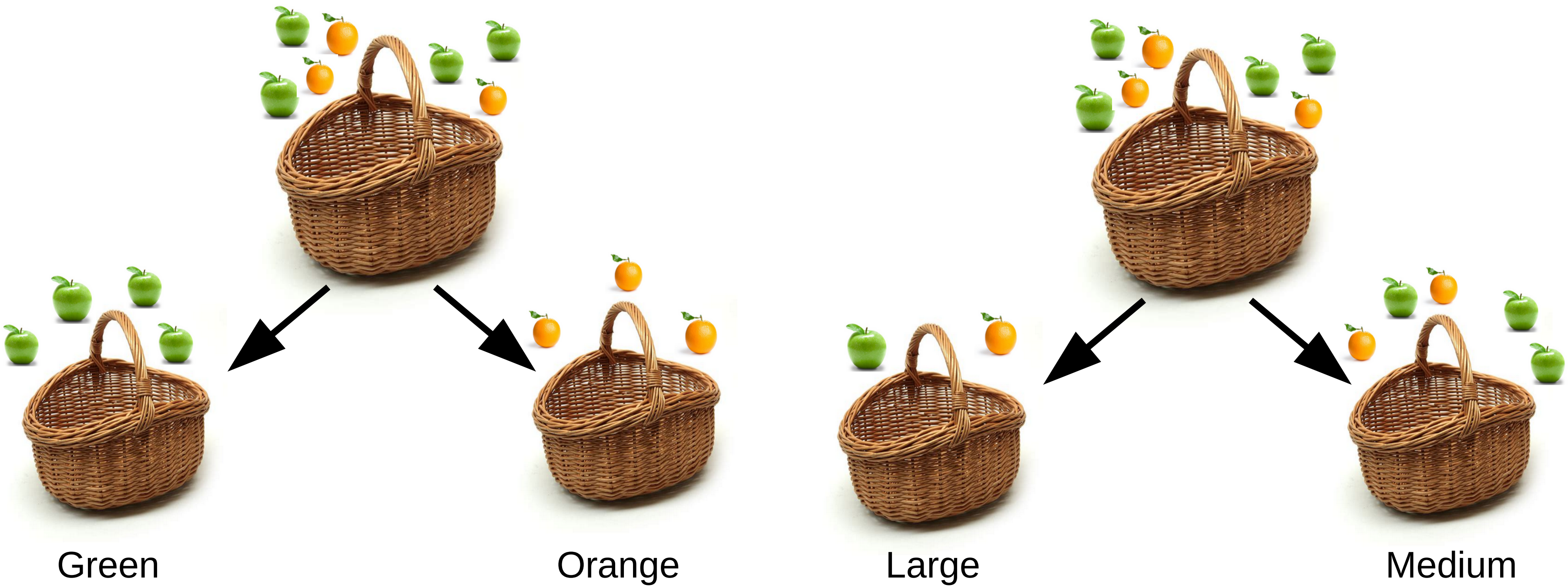
$$H = -p_O \log_2(p_O) - p_A \log_2(p_A) = -\frac{3}{7}\log_2\left(\frac{3}{7}\right) - \frac{4}{7}\log_2\left(\frac{4}{7}\right) = 0.985$$

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

$$H = 0.985$$

$$H_{colour} = 0$$

$$H_{size} = 0.98$$

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

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

Information gain

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:

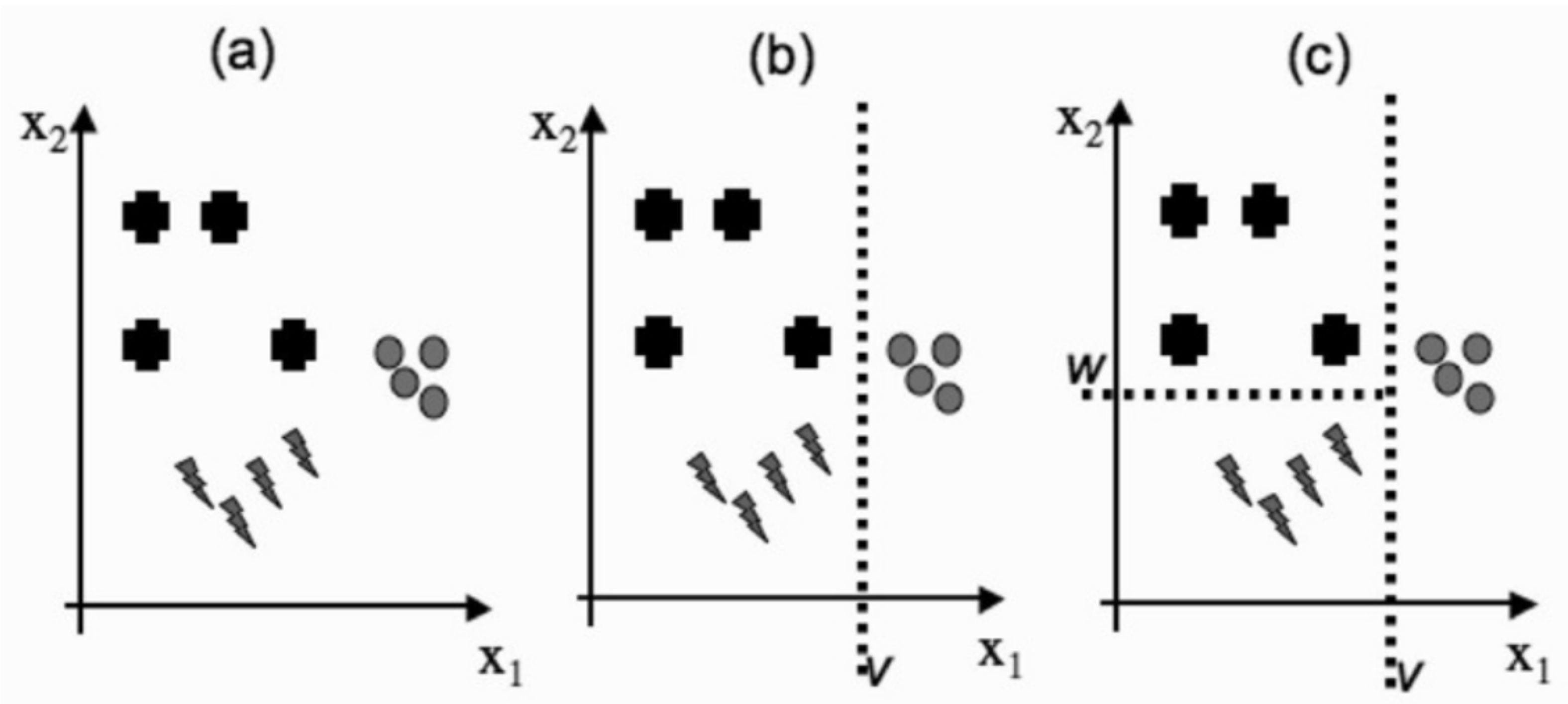
$$\mathrm{H}_{\mathrm{size}}=\overbrace{\frac{2}{7}}^{\text{fraction in Large}}\overbrace{\left(-\frac{1}{2}\log_2\frac{1}{2}-\frac{1}{2}\log_2\frac{1}{2}\right)}^{\text{entropy of Large}}+\overbrace{\frac{5}{7}}^{\text{fraction in Medium}}\overbrace{\left(-\frac{2}{5}\log_2\frac{2}{5}-\frac{3}{5}\log_2\frac{3}{5}\right)}^{\text{entropy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

The ID3 algorithm

UNIVERSITY OF LEEDS

- 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

Visualizing splits



Characteristics

Greedy with respect to G → potential local minimum

Deals with noisy data (by assigning the label to most common class)

Always uses all the features → prone to overfitting

↓

Pruning

Continuous variables → C4.5

Missing attributes

A Different Criterion: Gini Impurity

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

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

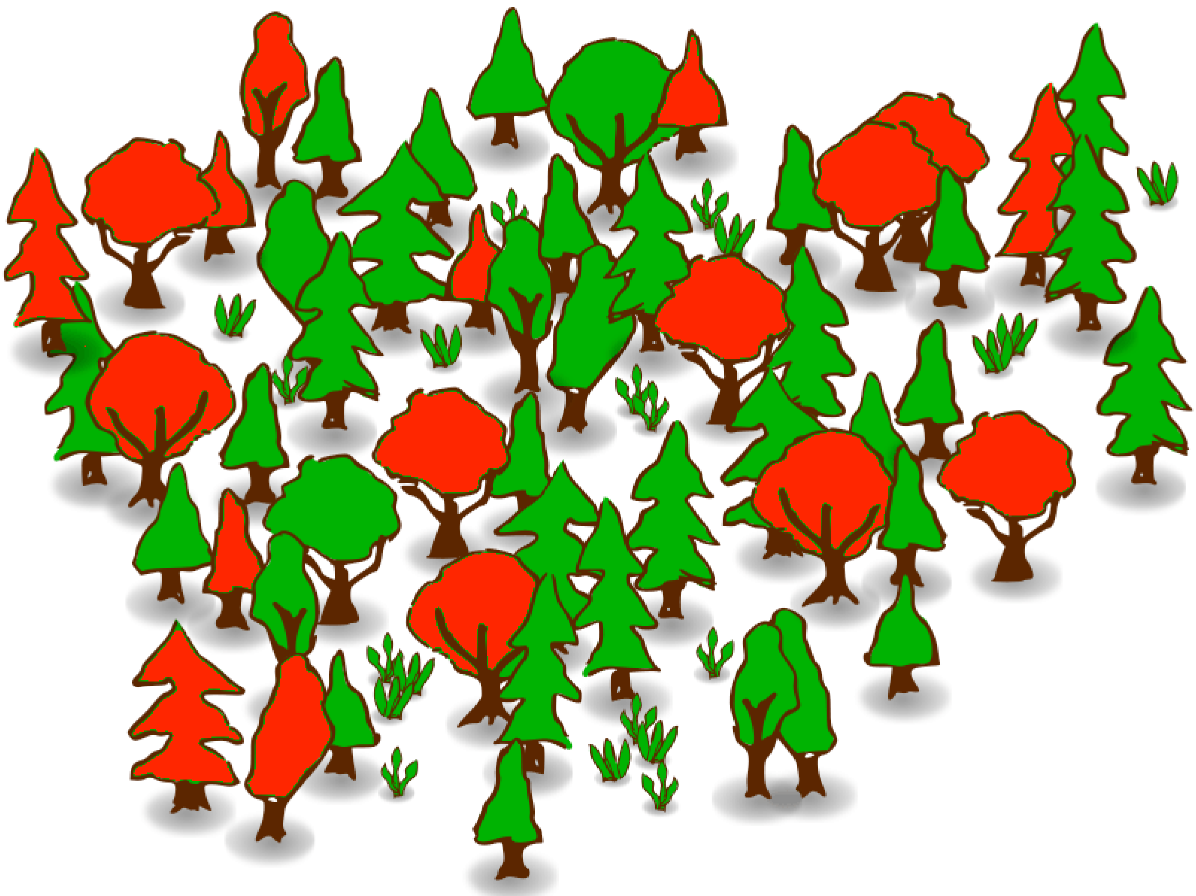
Gini split:

# of classes

$$G(S)=\sum_{i}^{C} p_i(1-p_i)=\sum_{i}^{C}(p_i-p_i^2)=\sum_{i}^{C} p_i-\sum_{i}^{C} p_i^2=1-\sum_{i}^{C} p_i^2$$

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

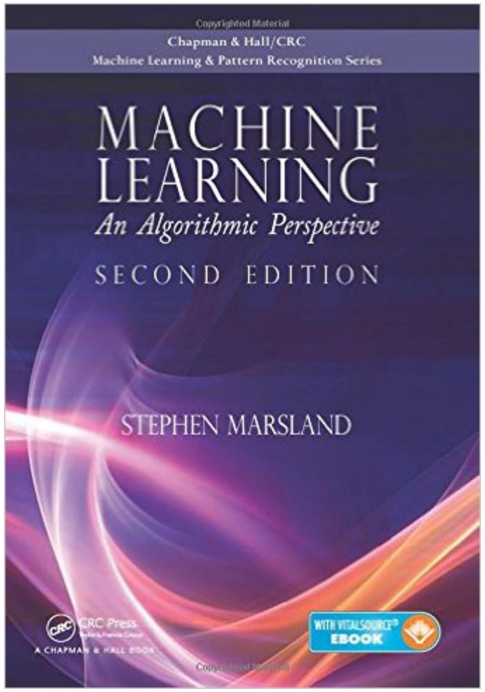
Random forests



Conclusion

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


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
Copyrighted Material

Chapter 12