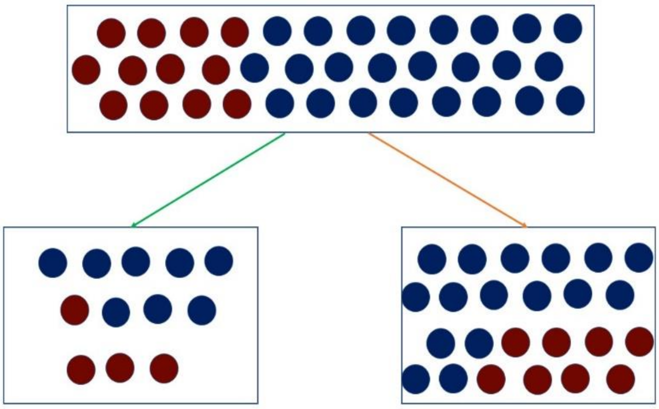
MLNN Laboratory Session

8 March 2023

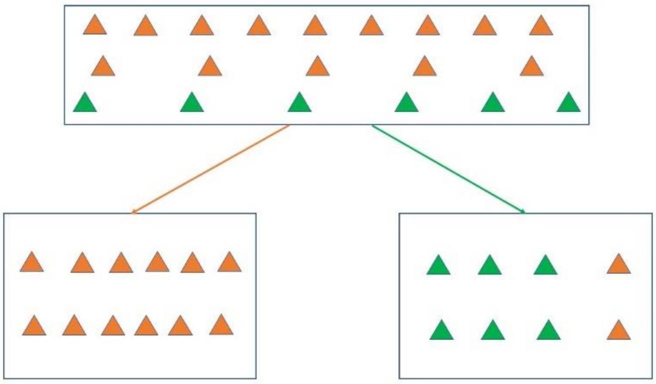
Problem 1

Calculate the information gain and GINI indices for the following splits. *(**The formulae for the information gain and GINI index are given at the end of this problem sheet.)***

(a)



(b)



Problem 2

You are given an anonymised list of customers of an insurance company, containing their age group and a satisfaction marker. The list also shows if the ustomer has renewed their policy or not. Write Python code that creates an optimal decision tree model for predicting whether a customer will renew..

The data is stored in the `insurance.csv` file. The first column is the age group ("Y" and "O" for the age ranges 25–45 and 46–65, respectively); the second column shows the satisfaction marker ("S" and "U" for generally satisfied and dissatisfied, respectively. The third column indicates whether the customer renewed ("R") or left ("L").

INFORMATION GAIN (see also lecture notes)

To calculate the information gain in a split, first, we need to calculate entropies for each node using the following formula

$$E = -\sum_i p_i log_2 p_i$$

where p_i is the fraction of elements with a certain label.

The information gain is then calculated as a difference between the information entropy in the parent node and average information entropy in the child nodes,

$$IG = E_p - E_{CA}.$$

The latter is calculated as

$$E_{CA} = f_{C1}E_{C1} + f_{C2}E_{C2},$$

where f_{C1} and f_{C2} are fractions of the elements going from the parent node to the first and second child nodes, respectively, and E_{C1} and E_{C2} are the information entropies in the first and second child nodes, respectively.

GINI index

Similar to the entropy, GINI index is a proxy of a “purity” of information in the node. For each node it can be calculated as

$$GINI = 1 - \sum_i p_i^2$$