Part C: Classification Homework

Authors: Penny Herrera, Robby Dosanjh

1. For the following data set, apply ID3 separately, and show all steps of derivation (computation, reasoning, developing / final decision trees, and rules).

	color	shape	size	class
1	red	square	big	+
2	blue	square	big	+
3	red	round	small	-
4	green	square	small	-
5	red	round	big	+
6	green	round	big	-

$$Entropy(t) = -\sum_j \mathrm{p}(j|t)\log_2 \mathrm{p}(j|t)$$

Here class is the target attribute and has two values (+ and -). So it is a binary classification problem.

For a binary classification problem

- If all examples are positive or all are negative then entropy will be ***zero*** i.e. low.
- If half of the examples are of positive class and half are of negative class then entropy is **one** i.e. high.

1. Calculating Initial Entropy

   Out of 6 instances, 3 are + and 3 are –

$$P(+) = -\left(\frac{3}{6}\right) * \log_2\left(\frac{3}{6}\right) = 0.5$$

$$P(-) = -\left(\frac{3}{6}\right) * \log_2\left(\frac{3}{6}\right) = 0.5$$

$$Entropy(t) = E(t) = 0.5 + 0.5 = 1$$

   Note: 1 indicates that the classes are highly impure. It is true in our case as there are equal number of observations with target class + and –

2. For every feature we will calculate entropy and information gain

   For attribute color

$$E(Color = red) = -\frac{2}{3} * \log_2\frac{2}{3} - \frac{1}{3} * \log_2\frac{1}{3} \approx 0.92$$

$$E(Color = blue) = -\frac{1}{1} * \log_2 \frac{1}{1} - 0 = 0$$

$$E(Color = green) = -0 - \frac{2}{2} * \log_2 \frac{2}{2} = 0$$

Average Entropy = $\frac{3}{6}(0.92) + \frac{1}{6}(0) + \frac{2}{6}(0) = 0.46$

Gain(Outlook) = $1 - 0.46 = \mathbf{0.54}$

For attribute shape

$$E(shape = square) = -\frac{2}{3} * \log_2 \frac{2}{3} - \frac{1}{3} * \log_2 \frac{1}{3} \approx 0.92$$

$$E(shape = round) = -\frac{1}{3} * \log_2 \frac{1}{3} - \frac{2}{3} * \log_2 \frac{2}{3} = 0.92$$

Average Entropy = $\frac{3}{6}(0.92) + \frac{3}{6}(0.92) = 0.92$

Gain(Outlook) = $1 - 0.92 = \mathbf{0.08}$

For attribute size

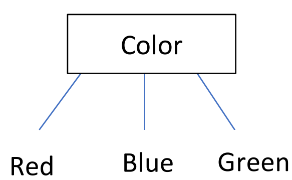
$$E(size = big) = -\frac{3}{4} * \log_2 \frac{3}{4} - \frac{1}{4} * \log_2 \frac{1}{4} \approx 0.81$$

$$E(size = small) = 0 - \frac{2}{2} * \log_2 \frac{2}{2} = 0$$

Average Entropy = $\frac{4}{6}(0.81) + \frac{2}{6}(0) = 0.54$

Gain(Outlook) = $1 - 0.54 = \mathbf{0.46}$

Feature ‘color’ provides more information on the ‘class’ as it has the highest information gain and hence will be chosen as the first splitting attribute



Likewise, we create the entire tree by selecting the splitting attribute as the attribute that gives the most information.

p= positive class

n= negative class

1. For the following data set, apply ID3 separately, and show all steps of derivation (computation, reasoning, developing / final decision trees, and rules).

	color	shape	size	class
1	red	square	big	+
2	blue	square	big	+
3	red	round	small	-
4	green	square	small	-
5	red	round	big	+
6	green	round	big	-

$$\text{Entropy} = \frac{-p}{p+n}\log_2\left(\frac{p}{p+n}\right) + \frac{-n}{p+n}\log_2\left(\frac{n}{p+n}\right)$$

avg info $\longrightarrow$ $I(\text{attribute}) = \sum \frac{p_i + n_i}{p+n} \text{ Entropy}(A)$

Gain = Entropy(s) − I (Attribute)

3 positive & 3 Negative values:

$$\text{Entropy}(s) = \frac{-3}{3+3}\log_2\left(\frac{3}{3+3}\right) + \frac{-3}{3+3}\log_2\left(\frac{3}{3+3}\right)$$

$$\text{Entropy}(s) = \frac{-1}{2}\log_2\left(\frac{1}{2}\right) + -\frac{1}{2}\log_2\left(\frac{1}{2}\right)$$

$$\text{Entropy}(s) = 0.5 + 0.5 = 1$$

Color Red:

$$\text{Entropy} = \frac{-p}{p+n}\log_2\left(\frac{p}{p+n}\right) + \frac{-n}{p+n}\log_2\left(\frac{n}{p+n}\right)$$

$$= -\frac{2}{3}\cdot\log_2\left(\frac{2}{3}\right) - \frac{1}{3}\cdot\log_2\left(\frac{1}{3}\right)$$

$$= 0.918 \approx 0.92$$

Color Blue:

$$\text{Entropy} = \frac{-p}{p+n}\log_2\left(\frac{p}{p+n}\right) + \frac{-n}{p+n}\log_2\left(\frac{n}{p+n}\right)$$

$$\text{Entropy} = \frac{-1}{1}\cdot\log_2\left(\frac{1}{1}\right) - 0 = 0$$

Color green:

$$\text{Entropy} = \frac{-p}{p+n}\log_2\left(\frac{p}{p+n}\right) + \frac{-n}{p+n}\log_2\left(\frac{n}{p+n}\right)$$

$$\text{Entropy} = 0 - \frac{2}{2}\cdot\log\left(\frac{2}{2}\right) = 0$$

avg info $\longrightarrow$ $I(\text{attribute}) = \sum \frac{p_i + n_i}{p+n}$ Entropy (A)

avg entropy

$$\text{avg entropy} = \frac{P_{red} + N_{red}}{P+n} \cdot \text{Entropy(red)} + \frac{P_{green} + N_{green}}{P+n} \cdot \text{Entropy(green)} + \frac{P_{blue} + N_{blue}}{P+n} \cdot \text{Entropy(blue)}$$

$$\text{avg entropy} = \frac{2+1}{3+3} \cdot 0.918 + \frac{0+2}{3+3} \cdot 0 + \frac{1+0}{3+3} \cdot 0$$

avg entropy $= 0.459$

gain (outlook) = Entropy (S) − avg entropy $= 1 - 0.459 = 0.541$

For Shapes :

square :

$$\text{Entropy} = \frac{-p}{p+n} \log_2\left(\frac{p}{p+n}\right) + \frac{-n}{p+n} \log_2\left(\frac{n}{p+n}\right)$$

$$= -\frac{2}{3} \cdot \log_2\left(\frac{2}{3}\right) - \frac{1}{3} \cdot \log_2\left(\frac{1}{3}\right)$$

Entropy $= 0.918 \approx 0.92$

round :

$$\text{Entropy} = \frac{-p}{p+n} \log_2\left(\frac{p}{p+n}\right) + \frac{-n}{p+n} \log_2\left(\frac{n}{p+n}\right)$$

$$= \frac{-1}{3} \cdot \log_2\left(\frac{1}{3}\right) - \frac{2}{3} \cdot \log_2\left(\frac{2}{3}\right)$$

Entropy $= 0.918 \approx 0.92$

avg info $\longrightarrow$ $I(\text{attribute}) = \sum \frac{p_i + n_i}{p+n}$ Entropy (A)

avg entropy

$$\text{avg entropy} = \frac{P_{square} + n_{square}}{p+n} \cdot \text{Entropy(square)} + \frac{P_{round} + n_{round}}{p+n} * \text{Entropy(round)}$$

$$\text{avg entropy} = \frac{2+1}{3+3} * 0.918 + \frac{1+2}{3+3} \cdot 0.918$$

avg entropy $= 0.918 \approx 0.92$

gain (outlook) = Entropy (S) − avg entropy (shape) $= 1 - 0.918 = 0.082$

size:

big:

$$\text{Entropy} = \frac{-p}{p+n}\log_2\left(\frac{p}{p+n}\right) + \frac{-n}{p+n}\log_2\left(\frac{n}{p+n}\right)$$

$$\text{Entropy} = \frac{-3}{4}\cdot\log_2\left(\frac{3}{4}\right) - \frac{1}{4}\cdot\log_2\left(\frac{1}{4}\right) \approx 0.81$$

small:

$$\text{Entropy} = \frac{-p}{p+n}\log_2\left(\frac{p}{p+n}\right) + \frac{-n}{p+n}\log_2\left(\frac{n}{p+n}\right)$$

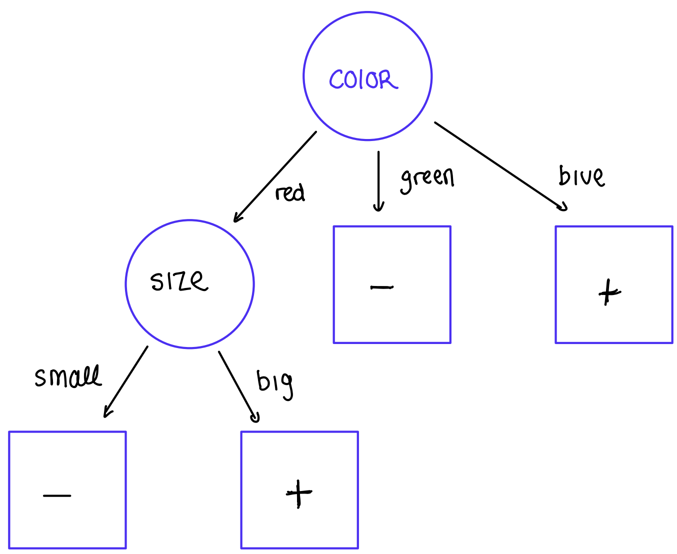
$$\text{Entropy} = 0 - \frac{2}{2}\cdot\log_2\left(\frac{2}{2}\right) \approx 0$$

$$\text{avg entropy} = \frac{p_{big}+n_{big}}{p+n} * \text{Entropy}(big) + \frac{p_{small}+n_{small}}{p+n}\cdot\text{Entropy}(small)$$

$$\text{avg entropy} = \frac{3+1}{3+3} * 0.81 + \frac{0+2}{3+3} * 0 = 0.54$$

$$\text{gain}(\text{outlook}) = \text{Entropy}(S) - \text{avg entropy}(\text{size}) = 1 - 0.54 = 0.46$$

Final descion tree:



(ii) understand what impact may happen to your created tree, if you later add a new missing attribute after creating the tree? You may add any new attribute, for example, "pattern of shirt". The values may be "checked", "striped", and one or two more. Be creative, add your own new favorite attribute or keep "pattern of shirt". The class column remains the same. What are some of the different possible changes you may expect to see on the classification decision tree you just created? Add this analysis to your solution document and submit. What if a data scientist provided his or her results with high confidence, by missing this attribute altogether? What if his or her results are used for decision making on how many million more shirts to produce for the next year? Do you think the data scientist surprises and makes an impact on the manager and CEO in case he or she discovers the new attribute and it's influence in getting more reliable results valuable to the company?

If a new attribute, such as "pattern of shirt" with values like "checked", "striped", "plain", etc., is introduced after the tree has been created, then we would have to reevaluate the Information Gain. The decision tree might change significantly if the new attribute provides a higher information gain compared to the existing attributes. This would cause the new attribute to become the root or a primary node depending on where it introduces the most significant reduction in entropy. This also introduces tree complexity. The tree might become more complex as this new attribute would add additional branches. This could potentially lead to better classification accuracy but also increase the risk of overfitting. If the "pattern of shirt" significantly influences the class outcome, previous analyses without this attribute might lead to incorrect decisions. For instance, if a particular pattern significantly correlates with a high sales class, missing this could result in underproduction or overproduction of certain shirts. A new attribute also leads to changes in decision-making. If a data scientist initially disregarded the "pattern of shirt" attribute and later found it to be highly influential, this could drastically impact business decisions. For instance, recognizing that a specific pattern drastically increases shirt sales could lead to a strategic shift in production, affecting how many million more shirts to produce for the next year. In terms of how Managerial and CEO Impact gets effected, discovering a previously overlooked attribute that significantly influences outcomes can surprise managers and CEOs. If the new attribute leads to more reliable results and better aligns production with market demand, this can enhance the data scientist's credibility and contribute positively to the company's strategy and financial success. This scenario underscores the importance of thorough data analysis and the potential consequences of overlooking significant attributes. It highlights the necessity for ongoing review and updating of analytical models to incorporate new insights and data.