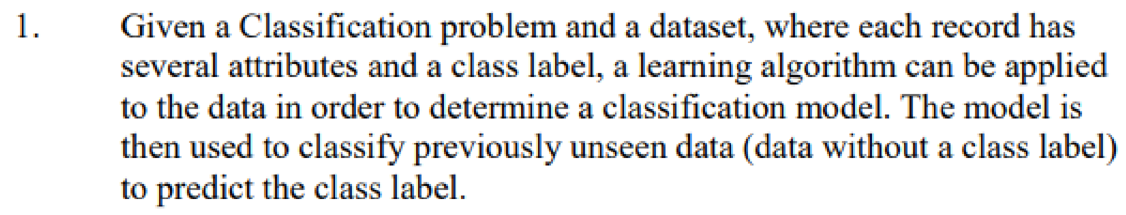
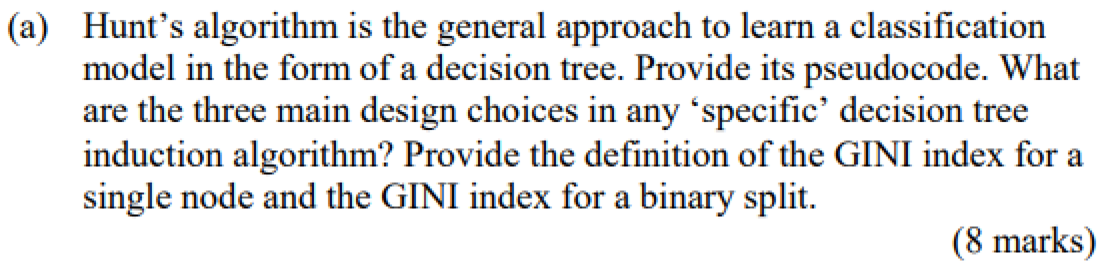
Question 1

Candidate Number: 59069

Data Science Algorithms & Tools - CS3DS19



1. Given a Classification problem and a dataset, where each record has several attributes and a class label, a learning algorithm can be applied to the data in order to determine a classification model. The model is then used to classify previously unseen data (data without a class label) to predict the class label.



(a) Hunt's algorithm is the general approach to learn a classification model in the form of a decision tree. Provide its pseudocode. What are the three main design choices in any 'specific' decision tree induction algorithm? Provide the definition of the GINI index for a single node and the GINI index for a binary split.

(8 marks)

PSEUDOCODE:

AS A RECURSIVE FUNCTION:

```
Hunts(Dataset, Attributes, testcase)
    Dataset
    IF Dataset contains only 1 Class: THEN
        Create Leaf Node labelled as Class
    ELSE IF Dataset is empty
        Create Leaf Node Labelled as Default class
    ELSE IF Attributes  is empty
        Create Leaf Node Labelled as Majority Class
    ELSE
        SPLIT data using attribute list and respective testcase
        SUBSET = Split()
        Remove Chosen Attribute and test case from list
        Hunts(SUBSET, Attributes, testcase)
# CALL INITIAL FUNCTION:
 Hunts(trainingData, attributeList, ConditionList )
```

Design Choices:

1. **Entropy**

   Measure for the degree of uncertainty, impurity, randomness, or disorder of the inputted data.

2. **Classification Error**

   Measure of misclassified labels.

3. **Gini Index**

Gini index is the measure for the degree of probability that a certain instance could be classified into the wrong class if it were to be chosen at random. Gini index varies between 0 and 1, where 0 is the most pure classification, implying that all records belong to a single class, and 1 is the most impure, demonstrating an equal distribution of records amongst classes.

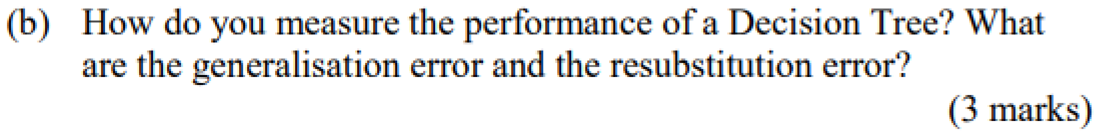
Gini index is defined as $1 - \sum_{i=1}^{n} (P_i)^2$.

At a single node ($GINI(t) = 1 - \sum_{j}^{n} (p\ (j \mid t\ )\ )^2$) where $p\ (j \mid t\ )$

Binary Split:

For a binary split, it is defined as the weighted sum of the gini index divided by the child nodes.

$GINI_{split} = \sum_{i=1}^{k} (\frac{n_i}{n} GINI(i))$ where n_i is the number of records at the current child node, and n is the total number of records at the parent node.



(b) How do you measure the performance of a Decision Tree? What are the generalisation error and the resubstitution error?

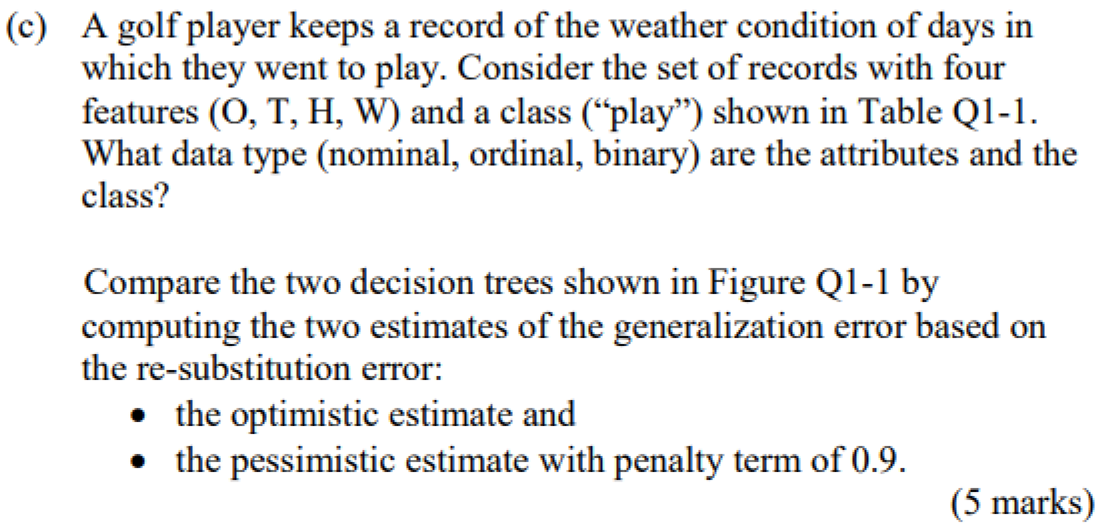
(3 marks)

The performance of a decision tree is assessed through the accuracy of the results it produces.

There are two main methods for assessing the performance of a decision tree

Generalisation error: The generalization is the estimation of the error made on the future data. This can be done via an optimistic approach, or a pessimistic approach.

Another option is the value representing the Resubstitution error, which is the calculated error value on the training dataset. If this value is too low, it is likely that the decision tree is extremely overfitted to the data.



(c) A golf player keeps a record of the weather condition of days in which they went to play. Consider the set of records with four features (O, T, H, W) and a class (“play”) shown in Table Q1-1. What data type (nominal, ordinal, binary) are the attributes and the class?

Compare the two decision trees shown in Figure Q1-1 by computing the two estimates of the generalization error based on the re-substitution error:

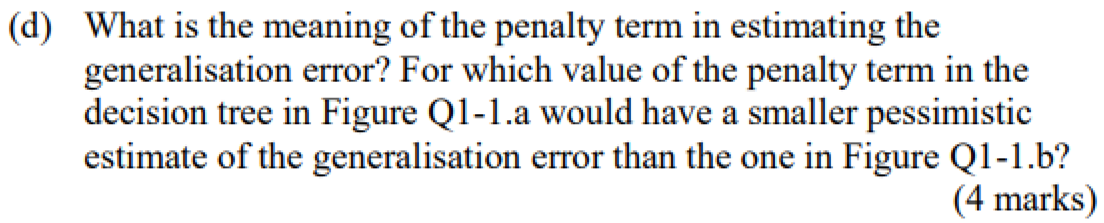
- the optimistic estimate and
- the pessimistic estimate with penalty term of 0.9.

(5 marks)

Feature / class	Data Type
O - Outlook	Ordinal
T - Temperature	Nominal
H - Humidity	Nominal
W - Windy	Binary
play	Binary

Penalty Term = 0.9

TREE	OPTIMISTIC	PESSIMISTIC	Re
1.A	0.25	0.3625	
1.B	0.5	0.725	



(d) What is the meaning of the penalty term in estimating the generalisation error? For which value of the penalty term in the decision tree in Figure Q1-1.a would have a smaller pessimistic estimate of the generalisation error than the one in Figure Q1-1.b?
(4 marks)

A penalty Term is a factor used to account for the overfitting of the data on the training set, when being used to estimate the accuracy on a test data set. It is calculated by penalising the complexity of the model, resulting in a higher penalty for a model with more leaf nodes.

To work out the value of the penalty term, P, for which the pessimistic estimate of the decision tree in Figure Q1-1.a is smaller than the generalisation error of the tree in Figure Q1-1.b, first it must be determined at which value of P these two values are equal.

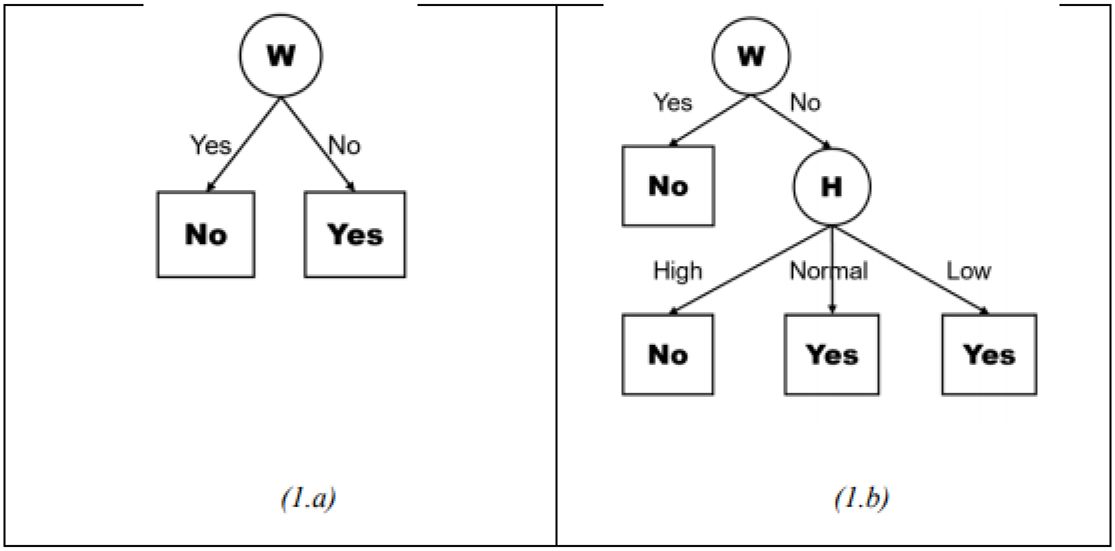
Tree	Total	MisClassifications	
1.A	16	4	
1.B	16	8	

This means it must be determined at which point $\frac{(4+(P\times 2))}{16}$ is equal to $\frac{8}{16}$therefore $(4 + (p \times 2)) = 8$ meaning that $p \times 2 = 4$ so $p = 2.$

Therefore for any value of Penalty Term, P, that is lower than 2, the pessimistic estimate of the decision tree in Figure Q1-1.a will be smaller than the generalisation of the error in Figure Q1-1.b.

ID	Outlook (O)	Temperature (T)	Humidity (H)	Windy (W)	play
1	Overcast	Cool	High	Yes	No
2	Overcast	Cool	Low	No	Yes
3	Overcast	Cool	Low	Yes	No
4	Overcast	Cool	Normal	Yes	No
5	Overcast	Hot	High	No	No
6	Overcast	Hot	Normal	No	Yes
7	Overcast	Mild	Low	Yes	Yes
8	Rainy	Cool	High	No	No
9	Rainy	Hot	High	No	No
10	Rainy	Hot	High	Yes	No
11	Rainy	Mild	Normal	No	Yes
12	Rainy	Mild	Normal	Yes	No
13	Sunny	Cool	Normal	No	Yes
14	Sunny	Cool	Normal	Yes	No
15	Sunny	Mild	High	No	Yes
16	Sunny	Mild	High	Yes	No

Table Q1-1. Golf data



(1.a)

(1.b)

Figure Q1-1. Decision Trees