Module-2

3 a. Discuss the two approaches to prevent over fitting the data. **(08 Marks)**

b. Consider the following set of training examples:

Instance	Classification	a_1	a_2
1	1	1	1
2	1	1	1
3	0	1	0
4	1	0	0
5	0	0	1
6	0	0	1

(i) What is the entropy of this collection of training examples with respect to the target function classification?

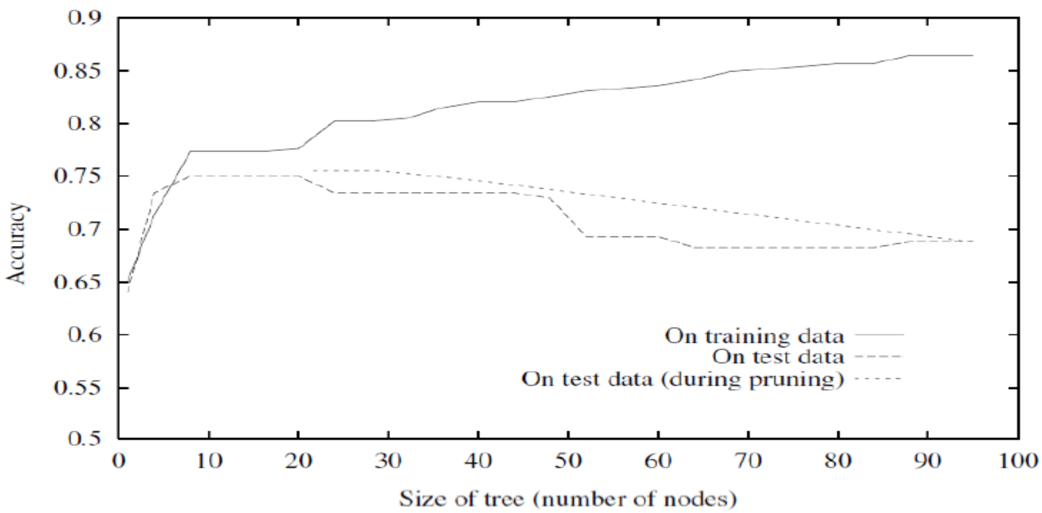
(ii) What is the information gain of a_2 relative to these training examples? **(08 Marks)**

3a

Reduced-Error Pruning

- Reduced-error pruning, is to consider each of the decision nodes in the tree to be candidates for pruning
- ***Pruning*** a decision node consists of removing the subtree rooted at that node, making it a leaf node, and assigning it the most common classification of the training examples affiliated with that node
- Nodes are removed only if the resulting pruned tree performs no worse than-the original over the validation set.
- Reduced error pruning has the effect that any leaf node added due to coincidental regularities in the training set is likely to be pruned because these same coincidences are unlikely to occur in the validation set

The impact of reduced-error pruning on the accuracy of the decision tree is illustrated in below figure



- The additional line in figure shows accuracy over the test examples as the tree is pruned. When pruning begins, the tree is at its maximum size and lowest accuracy over the test set. As pruning proceeds, the number of nodes is reduced and accuracy over the test set increases.
- The available data has been split into three subsets: the training examples, the validation examples used for pruning the tree, and a set of test examples used to provide an unbiased estimate of accuracy over future unseen examples. The plot shows accuracy over the training and test sets.

Pros and Cons

Pro: Produces smallest version of most accurate *T* (subtree of *T*)

Con: Uses less data to construct *T*

Can afford to hold out $D_{validation}$?. If not (data is too limited), may make error worse (insufficient D_{train})

Rule Post-Pruning

Rule post-pruning is successful method for finding high accuracy hypotheses

- Rule post-pruning involves the following steps:
- Infer the decision tree from the training set, growing the tree until the training data is fit as well as possible and allowing overfitting to occur.
- Convert the learned tree into an equivalent set of rules by creating one rule for each path from the root node to a leaf node.
- Prune (generalize) each rule by removing any preconditions that result in improving its estimated accuracy.
- Sort the pruned rules by their estimated accuracy, and consider them in this sequence when classifying subsequent instances.

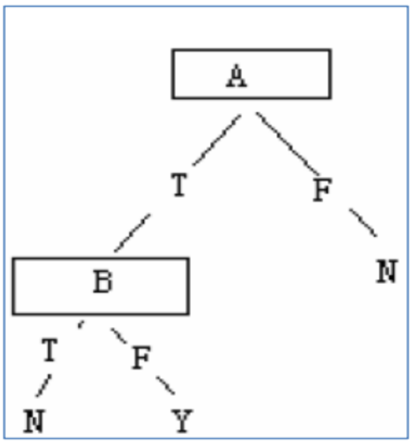
3b. < refer https://www.youtube.com/watch?v=JO2wiZif2OM >

4 a. Define decision tree. Construct the decision tree to represent the following Boolean functions:
i) $A \wedge \neg B$ ii) $A \vee [B \wedge C]$ iii) A XOR B **(06 Marks)**
b. Write the ID3 algorithm. **(06 Marks)**
c. What do you mean by gain and entropy? How it is used to build the decision tree. **(04 Marks)**

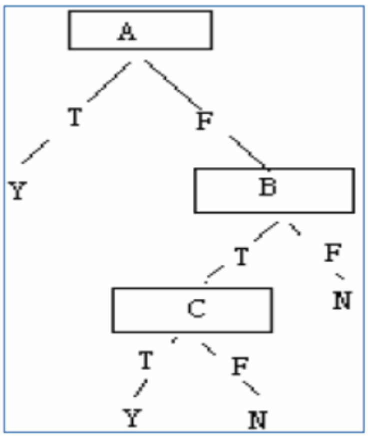
4a.

A decision tree is a graphical representation of all possible solutions to decisions based on certain conditions.

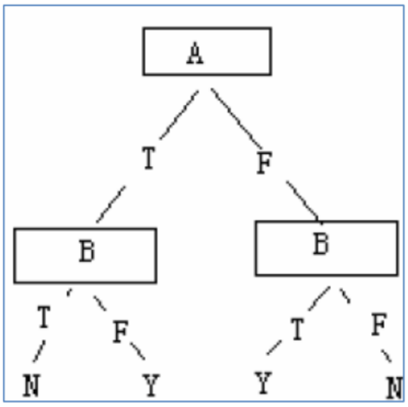
1) $A \wedge \neg B$



2) $A \vee [B \wedge C]$



3) A XOR B = $(A \wedge \neg B) \vee (\neg A \wedge B)$



4b.

The basic algorithm is ID3 which learns decision trees by constructing them top-down

---

ID3(Examples, Target_attribute, Attributes)

Examples are the training examples. Target_attribute is the attribute whose value is to be predicted by the tree. Attributes is a list of other attributes that may be tested by the learned decision tree. Returns a decision tree that correctly classifies the given Examples.

- Create a Root node for the tree
- If all Examples are positive, Return the single-node tree Root, with label = +
- If all Examples are negative, Return the single-node tree Root, with label = -
- If Attributes is empty, Return the single-node tree Root, with label = most common value of Target_attribute in Examples

- Otherwise Begin
  - A ← the attribute from Attributes that best* classifies Examples
  - The decision attribute for Root ← A
  - For each possible value, v_i, of A,
    - Add a new tree branch below *Root*, corresponding to the test A = v_i
    - Let *Examples* $_{vi}$, be the subset of Examples that have value v_i for *A*
    - If *Examples* $_{vi}$, is empty
      - Then below this new branch add a leaf node with label = most common value of Target_attribute in Examples
      - Else below this new branch add the subtree
        ID3(*Examples* $_{vi}$, Targe_tattribute, Attributes – {A}))
- End
- Return Root

4c.

GAIN

The statistical property called i*nformation gain* is the good quantitative measure of the worth of an attribute **.**

-> Information Gain measures how well a given attribute separates the training examples according to their target classification.

ENTROPY

Define and measures Randomness in the Data

-> Entropy is just a metric which measures the impurity

-> If the target attribute can take on c possible values the entropy can be as large as $\boldsymbol{log2c}$

•Information gain uses the notion of entropy , commonly used in information Theory.

3 a. What is a decision tree & discuss the use of decision tree for classification purpose with an example. (08 Marks)

b. Write and explain decision tree for the following transactions: (08 Marks)

Tid	Refund	Martial status	Taxable Income	Cheat
1	Yes	Single	125 K	No
2	No	Married	100 K	No
3	No	Single	70 K	No
4	Yes	Married	120 K	No
5	No	Divorced	95 K	Yes
6	No	Married	60 K	No
7	Yes	Divorced	220 K	No
8	No	Single	85 K	Yes
9	No	Married	75 K	No
10	No	Single	90 K	Yes

OR

3a.

A decision tree is a graphical representation of all possible solutions to decisions based on certain conditions.

<Refer others>

3b.<refer others>

4 a. For the transactions shown in the table compute the following :

(i) Entropy of the collection of transaction records of the table with respect to classification.

(ii) What are the information gain of a_1 and a_2 relative to the transactions of the table? (08 Marks)

Instance	1	2	3	4	5	6	7	8	9
a_1	T	T	T	F	F	F	F	T	F
a_2	T	T	F	F	T	T	F	F	T
Target class	+	+	–	+	–	–	–	+	–

b. Discuss the decision learning algorithm. (04 Marks)

c. List the issues of decision tree learning. (04 Marks)

1 of 2

4a.<refer others>

4b.< Id3 algorithm>

……

4c.

Practical issues in learning DT include determining how deeply to grow the DT ,

handling continuous attributes , choosing and appropriate attribute selection measure , handling training data with missing attribute values , handling attributes with different costs and improving computational efficiency.

1.Avoiding overfitting the data

2.Incorporating Continuous Valued Attributes

3.Alternative Measures for Selecting Attributes

4.Handling Training Examples with Missing Attribute Values

5.Handling Attributes with Differing Costs.

1. Avoiding Overfitting the Data :

- Building trees that "adapt too much" to the training examples may lead to "overfitting".
- Consider error of hypothesis h over
  - training data: $error_D(h)$ empirical error
  - entire distribution X of data: $error_X(h)$ expected error
- Hypothesis *h* *overfits* training data if there is an alternative hypothesis $h' \in H$ such that

$$error_D(h) < error_D(h') \quad \text{and}$$
$$error_X(h') < error_X(h)$$

i.e. h' behaves better over unseen data

2.Incorporating continuous valued attributes:

- So far discrete values for attributes and for outcome.
- Given a continuous-valued attribute A, dynamically create a new attribute A_c

$$A_c = \text{True } if\ A < c, \text{ False } otherwise$$

- How to determine threshold value *c* ?
- Example. *Temperature* in the *PlayTennis* example
  - Sort the examples according to *Temperature*

Temperature	40	48	| 60	72	80	| 90
PlayTennis	No	No	54 Yes	Yes	Yes	85 No

  - Determine candidate thresholds by averaging consecutive values where there is a change in classification: (48+60)/2=54 and (80+90)/2=85
  - Evaluate candidate thresholds (attributes) according to information gain. The best is $Temperature_{>54}$.The new attribute competes with the other ones

4. Handling Training Examples with Missing complete training data

->How to cope with the problem that the value of some attribute may be missing?

->Example : Blood Test Result in a medical diagnosis problem

-> The strategy: use other examples to guess attribute

1.Assign the value that is most common among the training examples at the node

2.Assign a probability to each value, based on frequencies, and assign valuesto missing attribute, according to this probability distribution

-> Missing values in new instances to be classified are treated accordingly, and the most probable classification is chosen (C 4.5)

5. Handling attributes with different costs

- Instance attributes may have an associated cost: we would prefer decision trees that use low-cost attributes
- ID3 can be modified to take into account costs:

1. Tan and Schlimmer (1990)

$$\frac{Gain^2(S, A)}{Cost(S, A)}$$

2. Nunez (1988)

$$\frac{2^{Gain(S, A)} - 1}{(Cost(A) + 1)^w} \quad w \in [0,1]$$

Module-2

3 Construct decision tree for the following data using ID3 algorithm.

Day	A1	A2	A3	Classification
1	True	Hot	High	No
2	True	Hot	High	No
3	False	Hot	High	Yes
4	False	Cool	Normal	Yes
5	False	Cool	Normal	Yes
6	True	Cool	High	No
7	True	Hot	High	No
8	True	Hot	Normal	Yes
9	False	Cool	Normal	Yes
10	False	Cool	High	No

(16 Marks)

OR

4 a. Explain the concept of decision tree learning. Discuss the necessary measure required to select the attributes for building a decision tree using ID3 algorithm. (08 Marks)

b. Discuss the issues of avoiding over fitting the data, handling continuous data and missing values in decision trees. (08 Marks)

3.<refer https://youtu.be/2A8AGfxs0D8 >

4.<refer others>