CLASSIFICATION BY DECISION TREE INDUCTION

Decision tree

– A flow-chart-like tree structure

– Internal node denotes a test on an attribute

– Branch represents an outcome of the test

– Leaf nodes represent class labels or class distribution

• ***Decision tree generation consists of two phases***

- Tree construction
    - At start, all the training examples are at the root
    - Partition examples recursively based on selected attributes
- Tree pruning

• Identify and remove branches that reflect noise or outliers

• ***Use of decision tree: Classifying an unknown sample***

– Test the attribute values of the sample against the decision tree

Training Dataset

This follows an example from Quinlan's ID3

Algorithm for decision tree induction

• Basic algorithm (a greedy algorithm)

– Tree is constructed in a top-down recursive divide-and-conquer manner

– At start, all the training examples are at the root

– Attributes are categorical (if continuous-valued, they are discretized in advance)

– Examples are partitioned recursively based on selected attributes

– Test attributes are selected on the basis of a heuristic or statistical measure (e.g., information gain)

• ***Conditions for stopping partitioning***

– All samples for a given node belong to the same class

– There are no remaining attributes for further partitioning – majority voting is employed for classifying the leaf

– There are no samples left

Extracting Classification Rules from Trees

• Represent the knowledge in the form of IF-THEN rules

• One rule is created for each path from the root to a leaf

Each attribute-value pair along a path forms a conjunction

• The leaf node holds the class prediction

• Rules are easier for humans to understand

Example

IF *age* = "<=30" AND *student* = "*no*" THEN *buys_computer* = "*no*"

IF *age* = "<=30" AND *student* = "*yes*" THEN *buys_computer* = "*yes*"

IF *age* = "31…40" THEN *buys_computer* = "*yes*"

IF *age* = ">40" AND *credit_rating* = "*excellent*" THEN *buys_computer* = "*yes*"

IF *age* = ">40" AND *credit_rating* = "*fair*" THEN *buys_computer* = "*no*"

Avoid Overfitting in Classification

The generated tree may overfit the training data

– Too many branches, some may reflect anomalies due to noise or outliers

– Result is in poor accuracy for unseen samples

• Two approaches to avoid over fitting

Prepruning:

Halt tree construction early—do not split a node if this would result in the goodness measure falling below a threshold

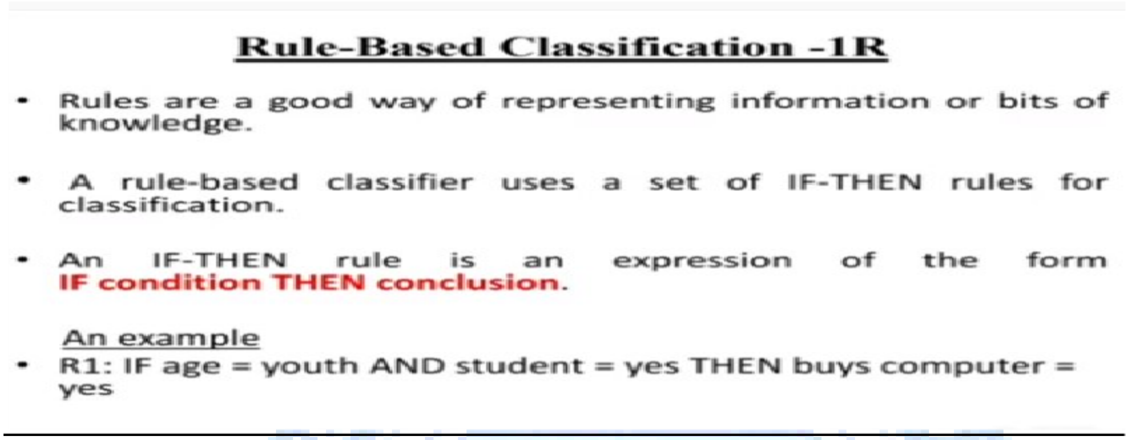
□ Difficult to choose an appropriate threshold

Post pruning:

□ Remove branches from a "fully grown" tree—get a sequence of progressively pruned trees

□ Use a set of data different from the training data to decide which the "best pruned tree"

RULE BASED CLASSIFICATION

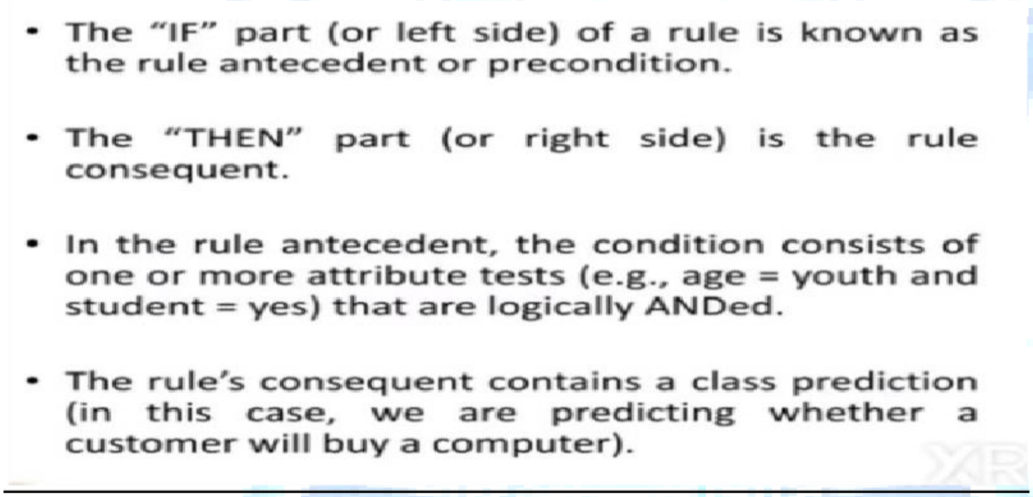


Rule-Based Classification -1R

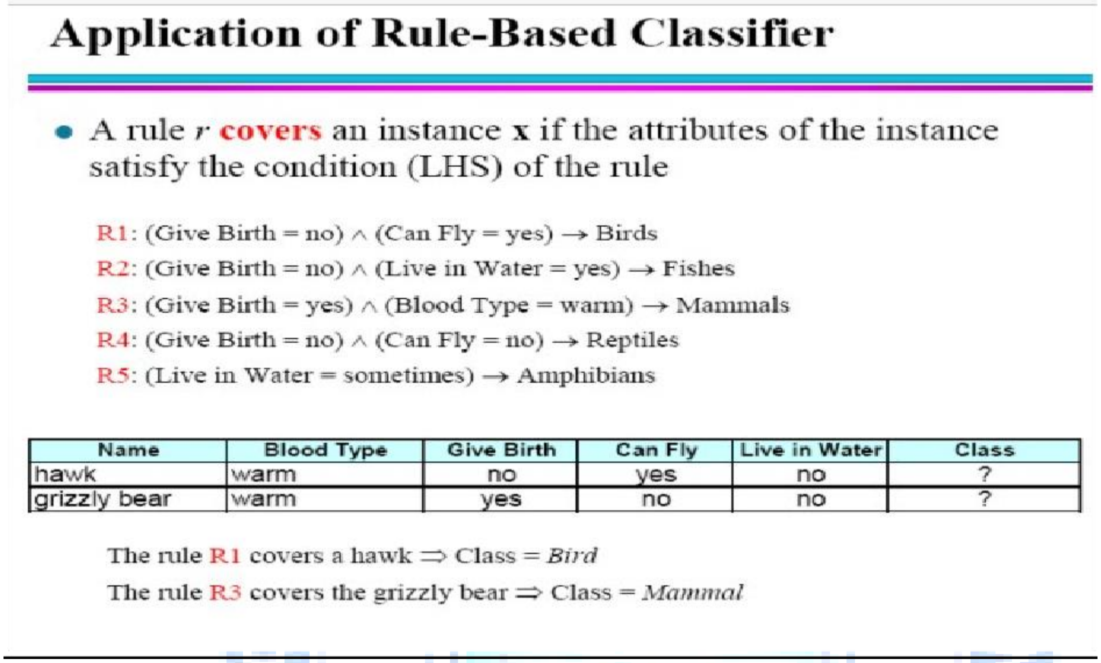
- Rules are a good way of representing information or bits of knowledge.
- A rule-based classifier uses a set of IF-THEN rules for classification.
- An IF-THEN rule is an expression of the form **IF condition THEN conclusion**.

An example

- R1: IF age = youth AND student = yes THEN buys computer = yes



- The "IF" part (or left side) of a rule is known as the rule antecedent or precondition.
- The "THEN" part (or right side) is the rule consequent.
- In the rule antecedent, the condition consists of one or more attribute tests (e.g., age = youth and student = yes) that are logically ANDed.
- The rule's consequent contains a class prediction (in this case, we are predicting whether a customer will buy a computer).



Application of Rule-Based Classifier

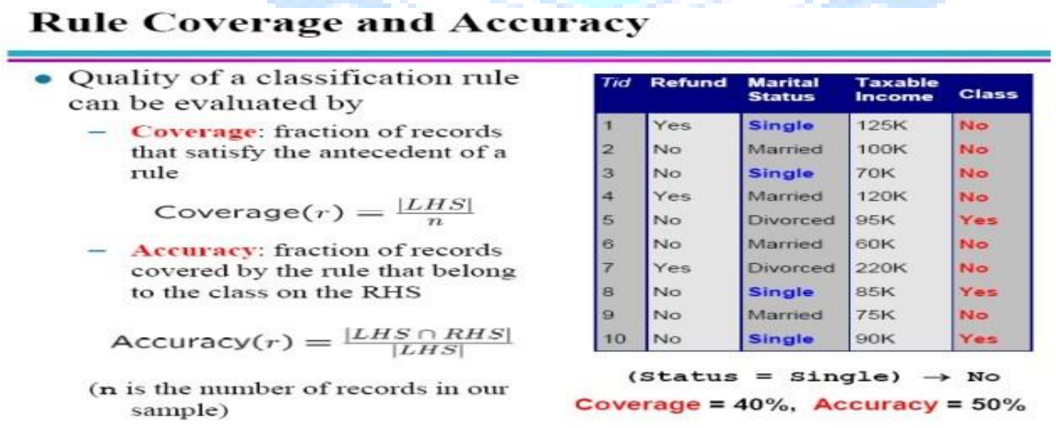
- A rule *r* **covers** an instance **x** if the attributes of the instance satisfy the condition (LHS) of the rule

R1: (Give Birth = no) ∧ (Can Fly = yes) → Birds
R2: (Give Birth = no) ∧ (Live in Water = yes) → Fishes
R3: (Give Birth = yes) ∧ (Blood Type = warm) → Mammals
R4: (Give Birth = no) ∧ (Can Fly = no) → Reptiles
R5: (Live in Water = sometimes) → Amphibians

Name	Blood Type	Give Birth	Can Fly	Live in Water	Class
hawk	warm	no	yes	no	?
grizzly bear	warm	yes	no	no	?

The rule R1 covers a hawk ⇒ Class = *Bird*

The rule R3 covers the grizzly bear ⇒ Class = *Mammal*



Rule Coverage and Accuracy

- Quality of a classification rule can be evaluated by
  - **Coverage**: fraction of records that satisfy the antecedent of a rule

$$\text{Coverage}(r) = \frac{|LHS|}{n}$$

  - **Accuracy**: fraction of records covered by the rule that belong to the class on the RHS

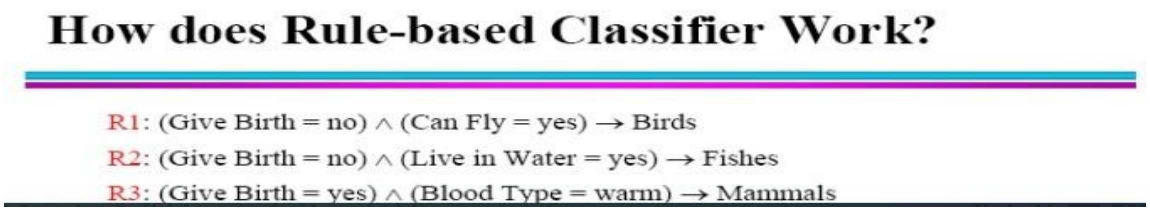
$$\text{Accuracy}(r) = \frac{|LHS \cap RHS|}{|LHS|}$$

(n is the number of records in our sample)

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

(Status = Single) → No

Coverage = 40%, Accuracy = 50%



How does Rule-based Classifier Work?

R1: (Give Birth = no) ∧ (Can Fly = yes) → Birds
R2: (Give Birth = no) ∧ (Live in Water = yes) → Fishes
R3: (Give Birth = yes) ∧ (Blood Type = warm) → Mammals