DATA MINING CLASSIFICATION

PRESENTED BY

DR. NEHA SHARMA, INDIA

Secretary, Society for Data Science

OBJECTIVE

To understand-

- What is classification?
- What is prediction?
- Classification by decision tree induction

Classification vs. Prediction

- Classification
  - Predicts categorical class labels (discrete or nominal)
  - Classifies data (constructs a model) based on the training set and the values (class labels) in a classifying attribute and uses it in classifying new data
- Prediction (Regression)
  - Models continuous-valued functions, i.e., predicts unknown or missing values
- Typical applications
  - Credit approval
  - Target marketing
  - Medical diagnosis
  - Fraud detection

Classification Example

- Example training database
  - Two predictor attributes: Age and Car-type (**S**port, **M**inivan and **T**ruck)
  - Age is ordered, Car-type is categorical attribute
  - Class label indicates whether person bought product
  - Dependent attribute is *categorical*

Age	Car	Class
20	M	Yes
30	M	Yes
25	T	No
30	S	Yes
40	S	Yes
20	T	No
30	M	Yes
25	M	Yes
40	M	Yes
20	S	No

Prediction or Regression Example

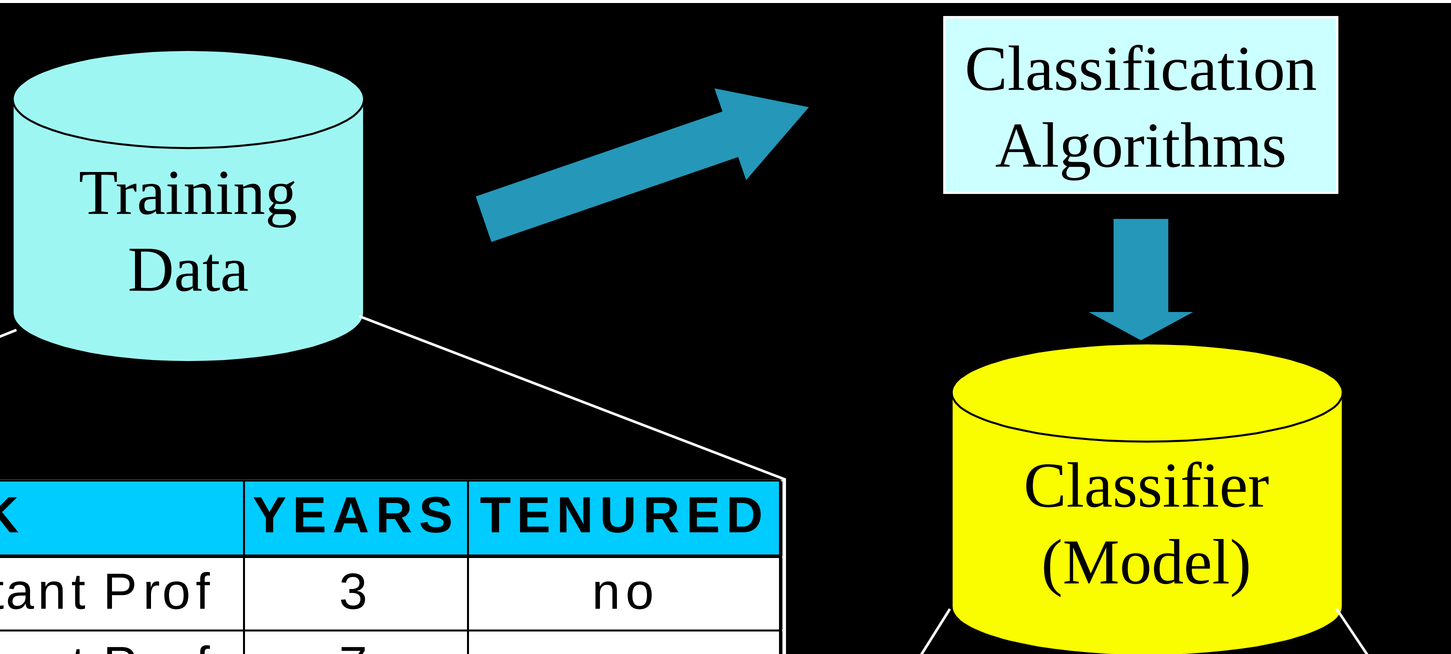
- Example training database
  - Two predictor attributes: Age and Car-type (**S**port, **M**inivan and **T**ruck)
  - Spent indicates how much person spent during a recent visit to the web site
  - Dependent attribute is *numerical*

Age	Car	Spent
20	M	$200
30	M	$150
25	T	$300
30	S	$220
40	S	$400
20	T	$80
30	M	$100
25	M	$125
40	M	$500
20	S	$420

Classification—A Two-Step Process

- Model construction: describing a set of predetermined classes
  - Each tuple/sample is assumed to belong to a predefined class, as determined by the class label attribute
  - The set of tuples used for model construction is training set
  - The model is represented as classification rules, decision trees, or mathematical formulae
- Model usage: for classifying future or unknown objects
  - Estimate accuracy of the model
    - The known label of test sample is compared with the classified result from the model
    - Accuracy rate is the percentage of test set samples that are correctly classified by the model
    - Test set is independent of training set, otherwise over-fitting will occur
  - If the accuracy is acceptable, use the model to classify data tuples whose class labels are not known.

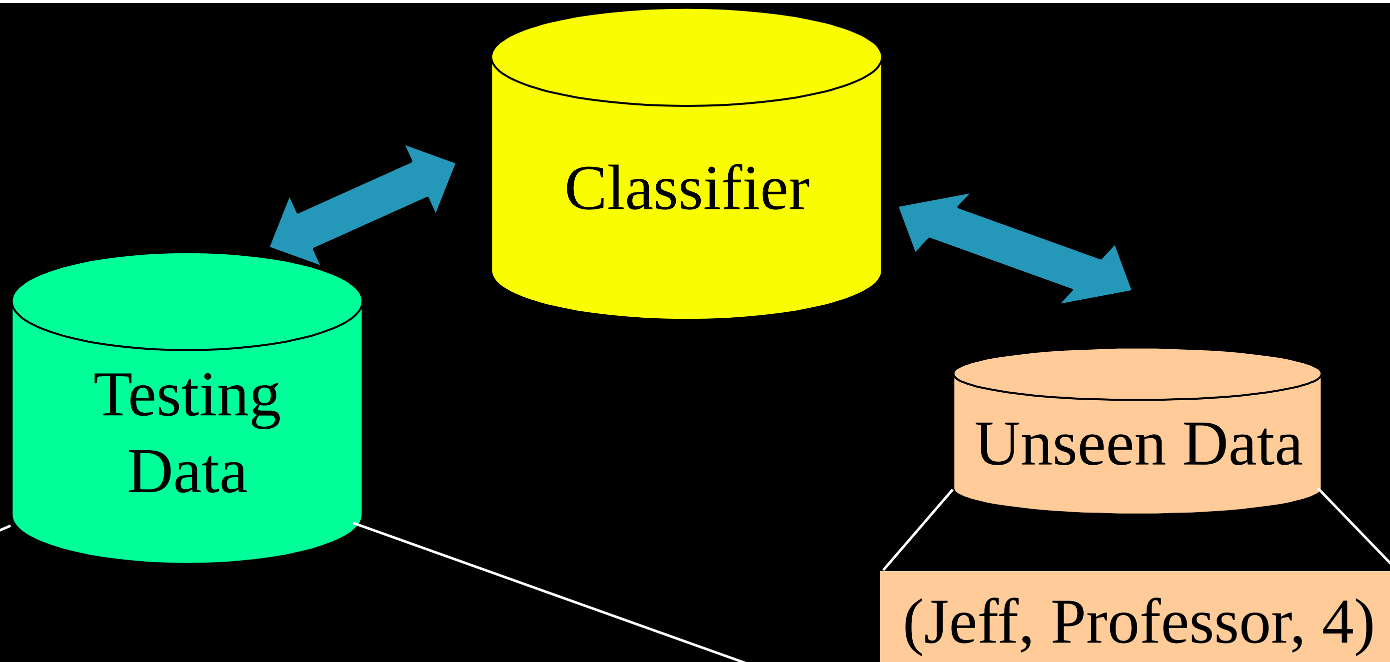
Process (1): Model Construction



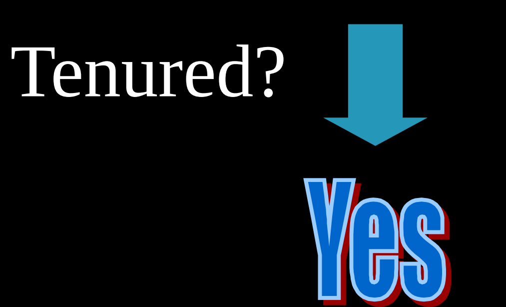
IF rank = ‘professor’
OR years > 6
THEN tenured = ‘yes’

NAME	RANK	YEARS	TENURED
Mike	Assistant Prof	3	no
Mary	Assistant Prof	7	yes
Bill	Professor	2	yes
Jim	Associate Prof	7	yes
Dave	Assistant Prof	6	no
Anne	Associate Prof	3	no

Process (2): Using the Model in Prediction



NAME	RANK	YEARS	TENURED
Tom	Assistant Prof	2	no
Merlisa	Associate Prof	7	no
George	Professor	5	yes
Joseph	Assistant Prof	7	yes



Supervised vs. Unsupervised Learning

- Supervised learning (classification)
  - Supervision: The training data (observations, measurements, etc.) are accompanied by labels indicating the class of the observations
  - New data is classified based on the training set
- Unsupervised learning (clustering)
  - The class labels of training data is unknown
  - Given a set of measurements, observations, etc. with the aim of establishing the existence of classes or clusters in the data

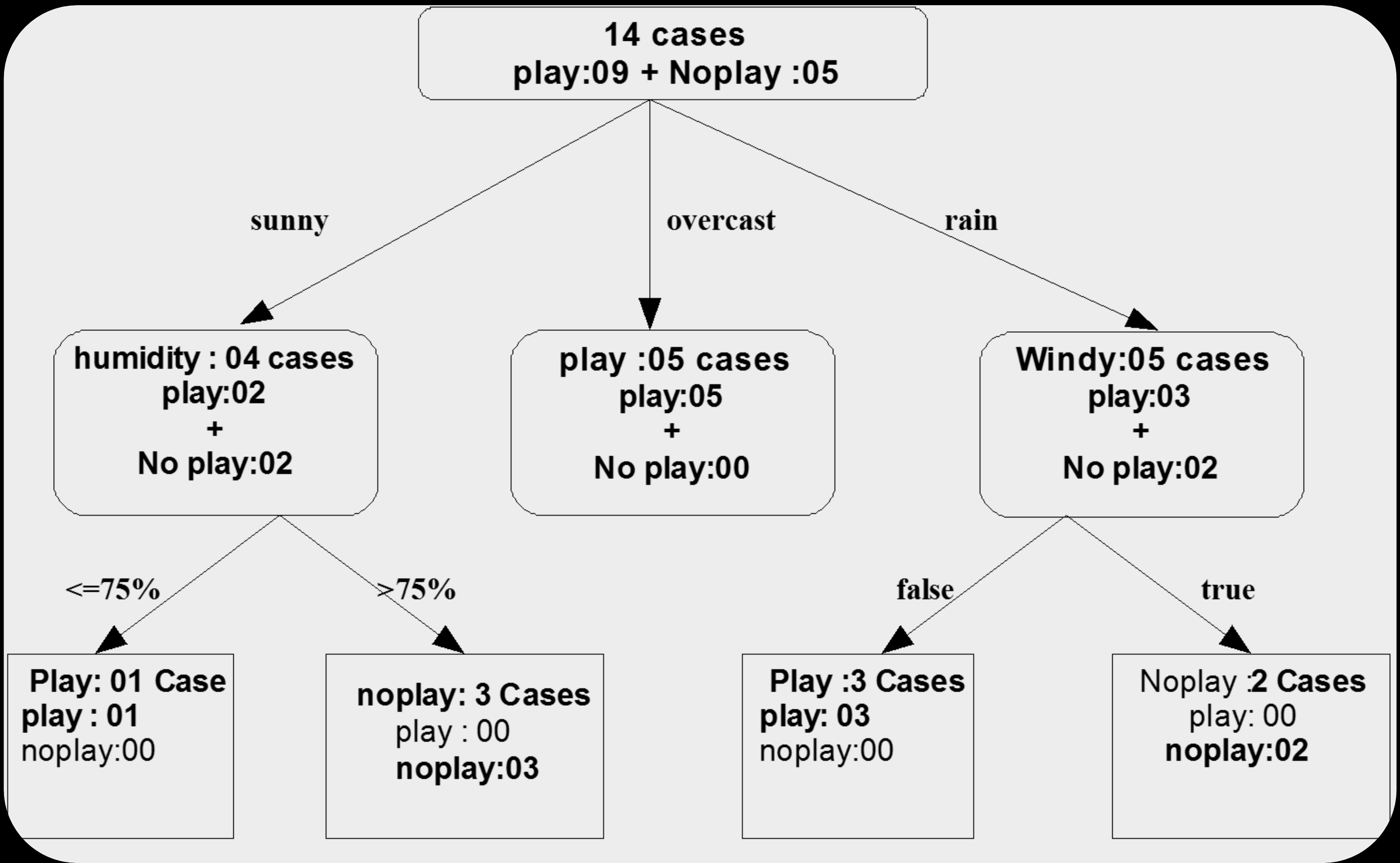
Algorithm for Decision Tree Induction

- Basic algorithm (a greedy algorithm)
  - Tree is constructed in a top-down recursive divide-and-conquer manner. At start, all the training datasets are at the root
  - Attributes are categorical (if continuous-valued, they are discretized in advance)
  - Dataset 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

Example: Decision Tree Induction

Weather Data Set					
Sl. No	Outlook	Temperature (° C)	Humidity (%)	Windy	**Class**
1	Sunny	75	70	True	**Play**
2	Sunny	83	90	True	**No play**
3	Sunny	87	85	false	**No play**
4	Sunny	76	95	false	**No play**
5	rain	71	80	True	**No play**
6	rain	65	70	True	**No play**
7	rain	75	80	false	**Play**
8	rain	75	80	false	**Play**
9	rain	68	95	false	**Play**
10	Overcast	73	90	True	**Play**
11	Overcast	82	78	false	**Play**
12	Overcast	64	65	True	**Play**
13	Overcast	81	75	false	**Play**
14	Overcast	80	74	false	Play

Example: Decision Tree Induction



Top-Down Tree Construction

BuildTree(Node *t*, Training database *Dt*,
Split Selection Method ***S***)

(1) Apply ***S*** to *D* to find splitting criterion

(2) **if** (*t* is not a leaf node)

(3) Create children nodes of *t*

(4) Partition *D* into children partitions

(5) Recurse on each partition

(6) **endif**

Attribute Selection Measures

- Select the attribute with the highest information gain
- Let p_i be the probability that an arbitrary tuple in D belongs to class C_i, estimated by $|C_{i, D}|/|D|$
- Expected information (entropy) needed to classify a tuple in D:

$$Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i)$$

- Information needed (after using A to split D into v partitions) to classify D:

$$Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times I(D_j)$$

- Information gained by branching on attribute A

$$Gain(A) = Info(D) - Info_A(D)$$

Case Study

The Sunburn Data Set					
Name	**Hair**	**Height**	**Weight**	**Lotion**	**Class label**
Sai	blonde	average	Light	No	Sunburned
Sachin	blonde	tall	Average	Yes	No sunburn
Ram	brown	short	Average	Yes	No sunburn
Rahim	blonde	short	Average	No	Sunburned
John	red	average	Heavy	No	Sunburned
Vicky	brown	tall	Heavy	No	No sunburn
Sara	brown	average	Heavy	No	No sunburn
Rani	blonde	short	Light	Yes	No sunburn

Partition the database on "Hair Color"

Sl.No	Name	Hair colour	Height	Weight	Lotion	Class label	Remark
1	Sai	Blonde	Average	Light	No	Sunburned	**sunburn case : 02 nosunburn case : 02**
2	Sachin	Blonde	Tall	Average	Yes	No sunburn	
4	Rahim	Blonde	Short	Average	No	Sunburned	
8	Rani	Blonde	Short	Light	Yes	No sunburn	

Sl.No	Name	Hair colour	Height	Weight	Lotion	Class label	Remark
3	Ram	Brown	Short	Average	Yes	No sunburn	**sunburn case : 00 nosunburn case : 03**
6	Vicky	Brown	Tall	Heavy	No	No sunburn	
7	Sara	Brown	Average	Heavy	No	No sunburn	

Sl.No	Name	Hair colour	Height	Weight	Lotion	Class label	Remark
5	John	Red	Average	Heavy	No	Sunburned	sunburn case : 01 nosunburn case : 00

Attribute	Instances	No of partitions
Height	{1,5,7}, {2,6},{3,4,8)	03
Weight	{1,8},{2,3,4}, {5,6,8}	03
Lotion	{1,4,5,6,7},{2,3,8}	02

Entropy for dataset D

Entropy = $$Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i)$$

- Class P: Sunburn = 3
- Class P: No Sunburn = 5

E (P_{SB}, P_{NoSB}) = $$Info(D) = -P_{SB} \log_2 P_{SB} - P_{NoSB} \log_2 P_{NoSB}$$

E (P_{SB}, P_{NoSB}) = - (3/8)$\log_2$ (3/8) - (5/8)$\log_2$ (5/8)

= - (0.375) log(0.375) / log 2 - (0.625) log(0.625) / log 2

= – (0.375) x (–0.425) /0.3 – 0.625 x (–0.204)/ 0.3

= 0.53 +0.425 = **0.955**

∴ Entropy for data set D = 0.955

Entropy for attribute "Hair Color"

Hair colour	Total no of cases	No. of sunburn cases	No. of nosunburn cases
Blonde	4	2	2
Brown	3	0	3
Red	1	1	0
Total	8	3	5

Entropy of attribute hair colour in data set D :

$$E(D_{Blonde}) = -(2/4)\log_2(2/4) - (2/4)\log_2(2/4) = \mathbf{1}$$

$$E(D_{Brown}) = -(0/3)\log_2(0/3) - (3/3)\log_2(3/3) = \mathbf{0}$$

$$E(D_{Red}) = -(1/1)\log_2(1/1) - (0/1)\log_2(0/1) = \mathbf{0}$$

The **probabilities** of hair colours blonde, brown, red are 4/8 , 3/8 and 1/8 respectively

Gain for Attribute Hair Color

$$Gain(A) = Info(D) - Info_A(D)$$

$$Gain(A) = E(D) - \sum_{i=1}^{n} p(D_i)\log_2(p_i)$$

$\therefore$ Gain = 0.955 – (4/8 x 1 + 1/8 x 0 + 3/8 x 0)

= 0.955 – 0.5 = 0.45

$\therefore$ Average entropy for attribute hair colour = 0.5

$\therefore$ Gain for attribute hair colour = 0.45

Entropy for attribute "Height"

Height	Total no of cases	No. of sunburn cases	No. of nosunburn cases
Short	03	01	02
Tall	02	00	02
average	03	01	02
Total	08	03	05

Entropy of attribute height in data set D

$E(D_{short}) = -(1/3)\log_2(1/3) - (2/3)\log_2(2/3) = \mathbf{0.918}$

$E(D_{tall}) = -(0/2)\log_2(0/2) - (2/2)\log_2(2/2) = \mathbf{0}$

$E(D_{Average}) = -(2/3)\log_2(2/3) - (1/3)\log_2(1/3) = \mathbf{0.918}$

The probabilities of short, tall , average are 3/8 , 2/8 and 3/8 respectively

Gain $= 0.955 - (3/8 \times 0.918 + 2/8 \times 0 + 3/8 \times .918 = \mathbf{0.267}$

∴ Average entropy for attribute height = 0.688

∴ Gain for attribute height = 0.267

Entropy for attribute "Weight"

Weight	Total no of cases	No. of sunburn cases	No. of nosunburn cases
Light	02	01	01
Average	03	01	02
Heavy	03	01	02
Total	**08**	**03**	**05**

Entropy of attribute Weight in data set D

$E(D_{light}) = -(1/2)\log_2(1/2) - (1/2)\log_2(1/2) =$ **1.0**

$E(D_{average}) = -(1/3)\log_2(1/3) - (2/3)\log_2(2/3) =$ **0.918**

$E(D_{heavy}) = -(1/3)\log(1/3) - (2/3)\log(2/3) =$ **0.918**

The probabilities of light, average and heavy are 2/8 , 3/8 and 3/8

Gain $= 0.955 - (2/8 \times 1 + 3/8 \times 0.918 + 3/8 \times 0.918) =$ **0.017**

∴ Average entropy for attribute weight = 0.938

∴ Gain for attribute height = 0.017

Entropy for attribute “Lotion”

Lotion	Total no of cases	No. of sunburn cases	No. of nosunburn cases
Yes	03	00	03
No	05	03	02
Total	08	03	05

Entropy of attribute Lotion in data set D

$E(D_{yes}) \quad = \quad -(0/3)\log_2(0/3) - (3/3)\log_2(3/3) \quad = \quad 0$

$E(D_{no}) \quad = \quad -(3/5)\log_2(3/5) - (2/5)\log_2(2/5) \quad = \quad 0.96$

The probabilities of Yes and No are 3/8 and 5/8

Gain $=$ $0.955 - (3/8 \times 0 + 5/8 \times 0.96) = \mathbf{0.355}$

∴ Average entropy for attribute lotion = 0.60

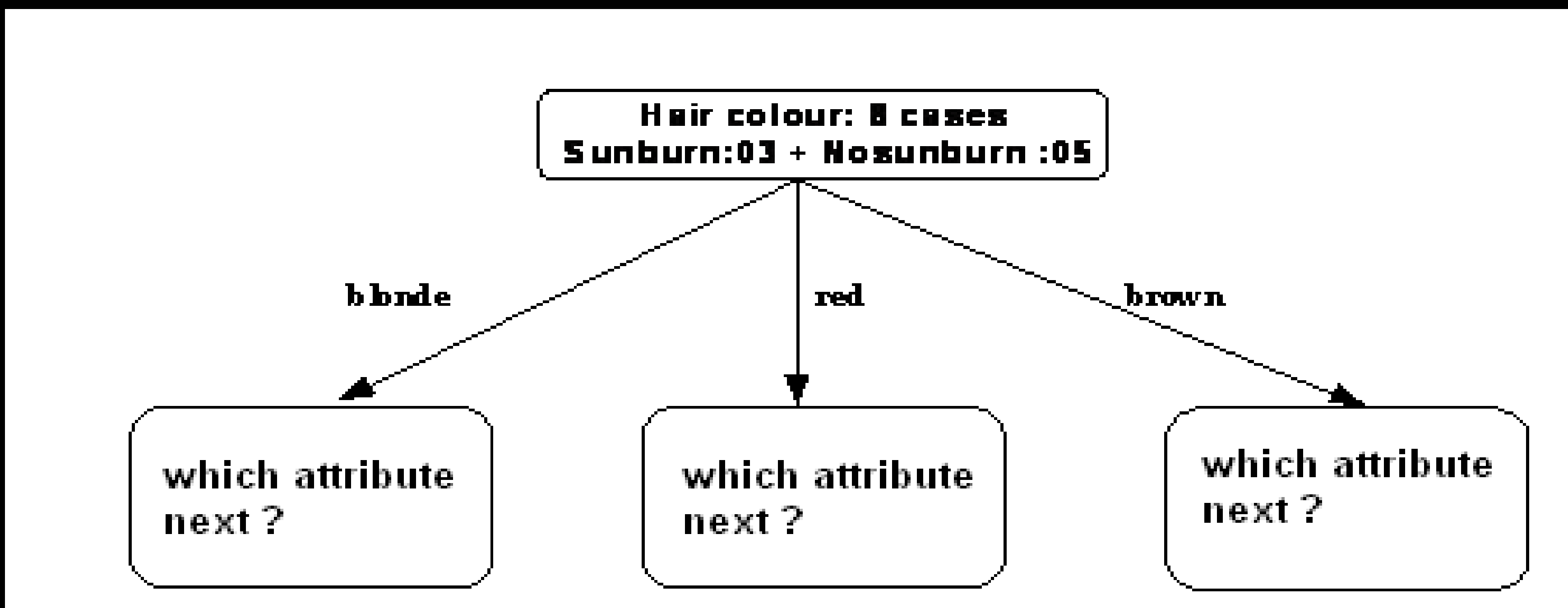
∴ Gain for attribute height = 0.355

Average Entropy and Gain

Attribute	Average Entropy	Gain
Hair colour	0.5	0.45
Height	0.688	0.267
Weight	0.938	0.017
Lotion	0.60	0.355

Constructing a Decision Tree

The attribute hair colour is selected as the root node since the entropy of hair colour is less compared the entropy values of other attribute. Also, the gain value of attribute hair colour is highest when compared to other gain values.



Selecting the Next Attribute for 1st Partition

Entropy calculations of height, weight and lotion for branch or partition "Blonde"

SL.No	Name	Hair colour	Height	Weight	Lotion	Class label	Remark
1	Sai	Blonde	Average	Light	No	Sunburned	**sunburn case : 02 nosunburn case : 02**
2	Sachin	Blonde	Tall	Average	Yes	No sunburn	
4	Rahim	Blonde	Short	Average	No	Sunburned	
8	Rani	Blonde	Short	Light	Yes	No sunburn	

- Class P: Sunburn = 2 Class P: No Sunburn = 2

E (P_{SB}, P_{NoSB}) = $Info(D) = -P_{SB}\log_2 P_{SB} - P_{NoSB}\log_2 P_{NoSB}$

E (PSB, PNoSB) = - (2/4) log2 (2/4) - (2/4) log2 (2/4) = 1

1^{st} Partition: Entropy for Attribute "Height"

Height	Total no of cases	No. of sunburn cases	No. of nosunburn cases
Average	01	01	00
Tall	01	00	01
Short	02	01	01
Total	04	02	02

Entropy of attribute height for branch blonde in data set P1

$E(D_{short}) = -(1/2)\log_2(1/2) - (1/2)\log_2(1/2) = \mathbf{1.0}$

$E(D_{tall}) = -(0/1)\log_2(0/1) - (1/1)\log_2(1/1) = \mathbf{0}$

$E(D_{Average}) = -(1/1)\log_2(1/1) - (0/1)\log_2(0/1) = 0$

The **probabilities** of short, tall , average are 2/4 , 1/4 and 1/4 respectively

$\therefore$ Gain $= 1.0 - (2/4 \times 1.0 + 1/4 \times 0 + 1/4 \times 0) = \mathbf{0.5}$

$\therefore$ Average entropy for attribute height for branch blonde= 0.5

$\therefore$ Gain for attribute height for branch blonde= 0.5

1st Partition: Entropy for Attribute "Weight"

Weight	Total no of cases	No. of sunburn cases	No. of nosunburn cases
Light	02	01	01
Average	02	01	01
Heavy	00	00	00
Total	**04**	**02**	**02**

Entropy of attribute Weight for branch blonde in data set P1

$E(D_{light}) = -(1/2)\log_2(1/2) - (1/2)\log_2(1/2) = \mathbf{1.0}$

$E(D_{average}) = -(1/2)\log_2(1/2) - (1/2)\log_2(1/2) = \mathbf{1.0}$

$E(D_{heavy}) = 0$

The **probabilities** of light, average , heavy are 2/4 , 2/4 and 0 respectively

$\therefore$ Gain = 1.0 – (2/4 x 1.0 + 2/4 x 1.0 + 0 x 0) =1– (0.5 + 0.5) **= 1–1 = 0**

$\therefore$ Average entropy for attribute weight for branch hair colour => blonde = 1.0

$\therefore$ Gain for attribute weight for branch hair colour => blonde = 0.0

1st Partition: Entropy for Attribute "Lotion"

Lotion	Total no of cases	No. of sunburn cases	No. of nosunburn cases
yes	02	00	02
No	02	02	00
Total	04	02	02

Entropy of attribute Lotion for branch blonde in data set P1

$E(D_{yes}) = -(0/2)\log_2(0/2) - (2/2)\log_2(2/2) = 0$

$E(D_{no}) = -(2/2)\log_2(2/2) - (0/2)\log_2(0/2) = 0$

The **probabilities** of yes and no are 2/4 , 2/4 respectively

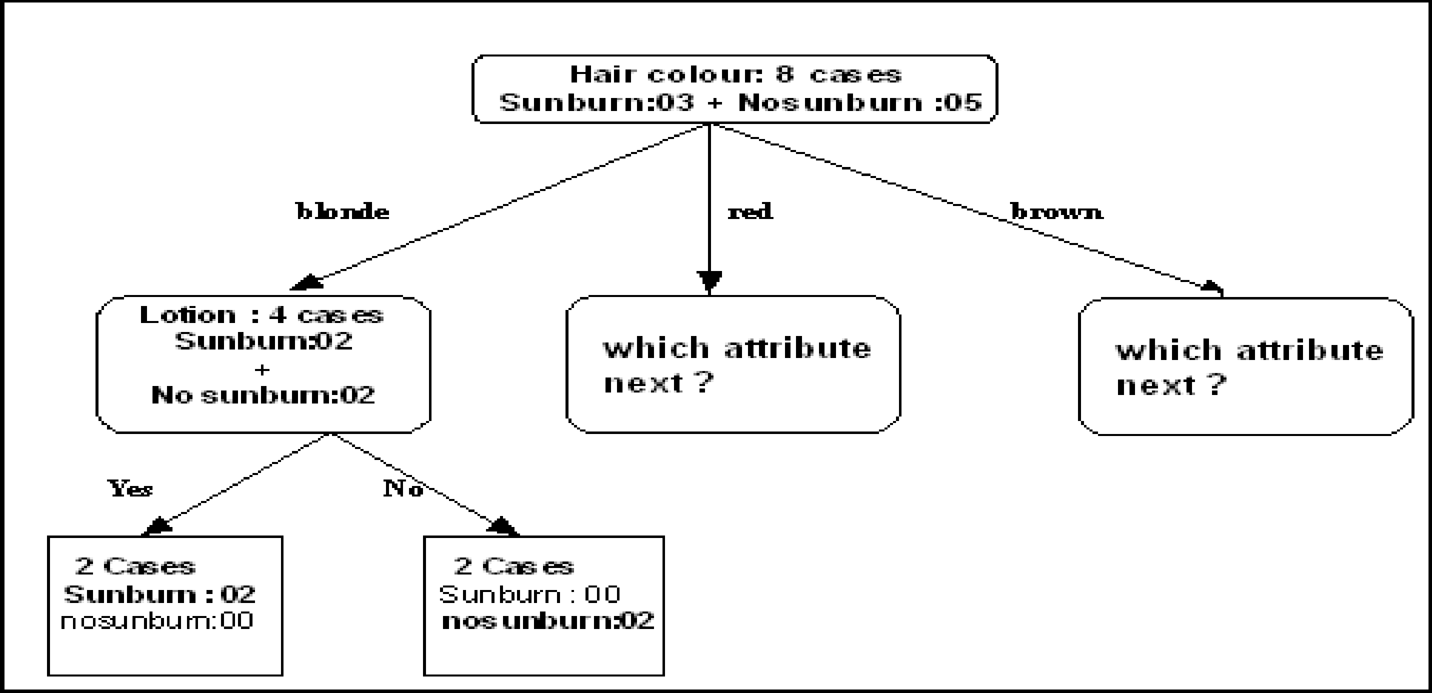
$\therefore$ Gain = 1.0 – (2/4 x 0.0+ 2/4 x 0.0)= 1– (0) = **1**

$\therefore$ Average entropy for attribute lotion for branch hair colour => blonde = 0.0

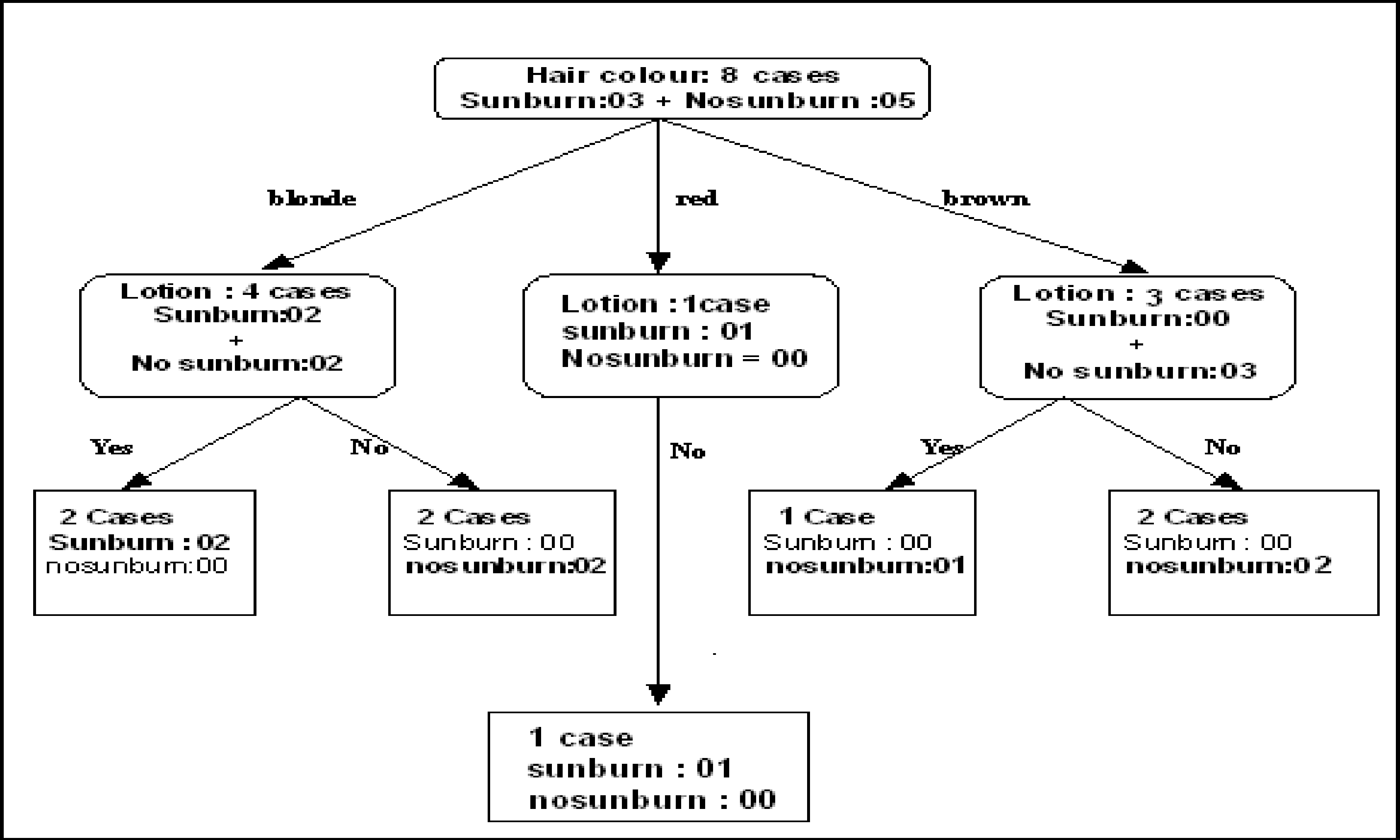
$\therefore$ Gain for attribute lotion for branch hair colour => blonde = 1.0

1st Partition: Average Entropy and Gain

Attribute	Average Entropy	Gain
Height	0.5	0.5
Weight	1.0	0.0
Lotion	0.0	1.0



Decision Tree



ASSIGNMENT -1: Vehicle Dataset

Cust_No	Age	Income	Student	Rating	Buy Vehicle
1	Young	High	No	Fair	No vehicle
2	Young	High	No	Good	No vehicle
3	Middle	High	No	Fair	Yes vehicle
4	Old	Medium	No	Fair	Yes vehicle
5	Old	Low	Yes	Fair	Yes vehicle
6	Old	Low	Yes	Good	No vehicle
7	Middle	Low	Yes	Good	Yes vehicle
8	Young	Medium	No	Fair	No vehicle
9	Young	Low	Yes	Fair	Yes vehicle
10	Old	Medium	Yes	Fair	Yes vehicle
11	Young	Medium	Yes	Good	Yes vehicle
12	Middle	Medium	No	Good	Yes vehicle
13	Middle	High	Yes	Fair	Yes vehicle
14	Old	Medium	No	good	No vehicle

ASSIGNMENT – 2: Computer Purchase

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

Thank you

QUESTIONS????