ECEN 758 Data Mining and Analysis
Decision Tree Example: [ZM]Ch 19

Nick Duffield

Department of Electrical & Computer Engineering

Texas A&M Institute of Data Science

Q2. Given Table 19.3, construct a decision tree using a purity threshold of 100%. Use information gain as the split point evaluation measure. Next, classify the point (Age=27,Car=Vintage).

Table 19.3. Data for Q2: Age is numeric and Car is categorical. Risk gives the class label for each point: high (H) or low (L)

Point	Age	Car	**Risk**
$\mathbf{x}_1$	25	Sports	L
$\mathbf{x}_2$	20	Vintage	H
$\mathbf{x}_3$	25	Sports	L
$\mathbf{x}_4$	45	SUV	H
$\mathbf{x}_5$	20	Sports	H
$\mathbf{x}_6$	25	SUV	H

Possible Split Points

Point	Age	Car	**Risk**
$\mathbf{x}_1$	25	Sports	L
$\mathbf{x}_2$	20	Vintage	H
$\mathbf{x}_3$	25	Sports	L
$\mathbf{x}_4$	45	SUV	H
$\mathbf{x}_5$	20	Sports	H
$\mathbf{x}_6$	25	SUV	H

- Age:
  numerical, use midpoints: 22.5, 35
- Car:
  - categorical, use {Sports}, {Vintage}, {SUV}.
  - Size 2 sets are complementary
    - Car in {Sports} ⇔Car not in {SUV, Vintage}

Evaluate Info Gain

Point	Age	Car	**Risk**
$\mathbf{x}_1$	25	Sports	*L*
$\mathbf{x}_2$	20	Vintage	*H*
$\mathbf{x}_3$	25	Sports	*L*
$\mathbf{x}_4$	45	SUV	*H*
$\mathbf{x}_5$	20	Sports	*H*
$\mathbf{x}_6$	25	SUV	*H*

- Information gain for each split:
  - $H(D) - H(D_Y, D_N) = H(D) - (n_Y/n)\,H(D_Y) - (n_N/n)\,H(D_N)$
- Computing H(D)
  - $P(L) = 2/6 = 1/3$
  - $P(H) = 1 – P(L) = 2/3$,
  - $H(D) = - P(L) \log_2 P(L) – P(H) \log_2 P(H)$

    $= - (1/3) \log_2 (1/3) – (2/3) \log_2 (2/3)$

    $= 0.9183$

Point	Age	Car	**Risk**
$\mathbf{x}_1$	25	Sports	L
$\mathbf{x}_2$	20	Vintage	H
$\mathbf{x}_3$	25	Sports	L
$\mathbf{x}_4$	45	SUV	H
$\mathbf{x}_5$	20	Sports	H
$\mathbf{x}_6$	25	SUV	H

Info gain: $H(D) - (n_Y/n)\, H(D_Y) - (n_N/n)\, H(D_N)$

$H(D_Y) = - P_Y(H) \log_2 P_Y(H) - P_Y(L) \log_2 P_Y(L)$

$H(D_N) = - P_N(H) \log_2 P_N(H) - P_N(L) \log_2 P_N(L)$

Split	**n_Y**	**$P_Y(H)$**	**$P_Y(L)$**	**n_N**	**$P_N(H)$**	**$P_N(L)$**	**Info Gain**
Age <= 22.5							
Age <= 35							
Car = Sports							
Car = Vintage							
Car = SUV							

Age <= 22.5

YES

NO

Point	Age	Car	**Risk**
$\mathbf{x}_1$	25	Sports	*L*
$\mathbf{x}_2$	20	Vintage	*H*
$\mathbf{x}_3$	25	Sports	*L*
$\mathbf{x}_4$	45	SUV	*H*
$\mathbf{x}_5$	20	Sports	*H*
$\mathbf{x}_6$	25	SUV	*H*

Info gain: $H(D) - (n_Y/n)\, H(D_Y) - (n_N/n)\, H(D_N)$

$H(D_Y) = - P_Y(H) \log_2 P_Y(H) - P_Y(L) \log_2 P_Y(L)$

$H(D_N) = - P_N(H) \log_2 P_N(H) - P_N(L) \log_2 P_N(L)$

Split	**n_Y**	**$P_Y(H)$**	**$P_Y(L)$**	**n_N**	**$P_N(H)$**	**$P_N(L)$**	**Info Gain**
Age <= 22.5							
Age <= 35							
Car = Sports							
Car = Vintage							
Car = SUV							

Age <= 22.5

YES

NO

Point	Age	Car	**Risk**
$\mathbf{x}_1$	25	Sports	L
$\mathbf{x}_2$	20	Vintage	H
$\mathbf{x}_3$	25	Sports	L
$\mathbf{x}_4$	45	SUV	H
$\mathbf{x}_5$	20	Sports	H
$\mathbf{x}_6$	25	SUV	H

Info gain: $H(D) - (n_Y/n)\,H(D_Y) - (n_N/n)\,H(D_N)$

$H(D_Y) = -P_Y(H)\log_2 P_Y(H) - P_Y(L)\log_2 P_Y(L)$
$H(D_N) = -P_N(H)\log_2 P_N(H) - P_N(L)\log_2 P_N(L)$

Split	$\mathbf{n_Y}$	$\mathbf{P_Y(H)}$	$\mathbf{P_Y(L)}$	$\mathbf{n_N}$	$\mathbf{P_N(H)}$	$\mathbf{P_N(L)}$	**Info Gain**
Age <= 22.5	2	1	0	4	1/2	1/2	0.2516
Age <= 35							
Car = Sports							
Car = Vintage							
Car = SUV							

Car = Sports

YES

NO

Point	Age	Car	**Risk**
$\mathbf{x}_1$	25	Sports	L
$\mathbf{x}_2$	20	Vintage	H
$\mathbf{x}_3$	25	Sports	L
$\mathbf{x}_4$	45	SUV	H
$\mathbf{x}_5$	20	Sports	H
$\mathbf{x}_6$	25	SUV	H

Info gain: $H(D) - (n_Y/n)\, H(D_Y) - (n_N/n)\, H(D_N)$

$H(D_Y) = - P_Y(H) \log_2 P_Y(H) - P_Y(L) \log_2 P_Y(L)$

$H(D_N) = - P_N(H) \log_2 P_N(H) - P_N(L) \log_2 P_N(L)$

Split	**n_Y**	**$P_Y(H)$**	**$P_Y(L)$**	**n_N**	**$P_N(H)$**	**$P_N(L)$**	**Info Gain**
Age <= 22.5	2	1	0	4	1/2	1/2	0.2516
Age <= 35							
Car = Sports							
Car = Vintage							
Car = SUV							

Car = Sports

YES

NO

Point	Age	Car	Risk
$\mathbf{x}_1$	25	Sports	L
$\mathbf{x}_2$	20	Vintage	H
$\mathbf{x}_3$	25	Sports	L
$\mathbf{x}_4$	45	SUV	H
$\mathbf{x}_5$	20	Sports	H
$\mathbf{x}_6$	25	SUV	H

Info gain: $H(D) - (n_Y/n)\, H(D_Y) - (n_N/n)\, H(D_N)$

$H(D_Y) = - P_Y(H) \log_2 P_Y(H) - P_Y(L) \log_2 P_Y(L)$

$H(D_N) = - P_N(H) \log_2 P_N(H) - P_N(L) \log_2 P_N(L)$

Split	n_Y	$P_Y(H)$	$P_Y(L)$	n_N	$P_N(H)$	$P_N(L)$	Info Gain
Age <= 22.5	2	1	0	4	1/2	1/2	0.2516
Age <= 35							
Car = Sports	3	1/3	2/3	3	1	0	0.4592
Car = Vintage							
Car = SUV							

Point	Age	Car	**Risk**
$\mathbf{x}_1$	25	Sports	L
$\mathbf{x}_2$	20	Vintage	H
$\mathbf{x}_3$	25	Sports	L
$\mathbf{x}_4$	45	SUV	H
$\mathbf{x}_5$	20	Sports	H
$\mathbf{x}_6$	25	SUV	H

Info gain: $H(D) - (n_Y/n)\, H(D_Y) - (n_N/n)\, H(D_N)$

$H(D_Y) = -\, P_Y(H) \log_2 P_Y(H) - P_Y(L) \log_2 P_Y(L)$

$H(D_N) = -\, P_N(H) \log_2 P_N(H) - P_N(L) \log_2 P_N(L)$

Split	**n_Y**	**$P_Y(H)$**	**$P_Y(L)$**	**n_N**	**$P_N(H)$**	**$P_N(L)$**	**Info Gain**
Age <= 22.5	2	1	0	4	1/2	1/2	0.2516
Age <= 35	5	3/5	2/5	1	1	0	0.1092
Car = Sports	3	1/3	2/3	3	1	0	0.4592
Car = Vintage	1	1	0	5	3/5	2/5	0.1092
Car = SUV	2	1	0	4	1/2	1/2	0.2516

Point	Age	Car	**Risk**
$\mathbf{x}_1$	25	Sports	L
$\mathbf{x}_2$	20	Vintage	H
$\mathbf{x}_3$	25	Sports	L
$\mathbf{x}_4$	45	SUV	H
$\mathbf{x}_5$	20	Sports	H
$\mathbf{x}_6$	25	SUV	H

Info gain: $H(D) - (n_Y/n)\,H(D_Y) - (n_N/n)\,H(D_N)$

$H(D_Y) = -\,P_Y(H)\log_2 P_Y(H) - P_Y(L)\log_2 P_Y(L)$

$H(D_N) = -\,P_N(H)\log_2 P_N(H) - P_N(L)\log_2 P_N(L)$

Split	**n_Y**	**$P_Y(H)$**	**$P_Y(L)$**	**n_N**	**$P_N(H)$**	**$P_N(L)$**	**Info Gain**
Age <= 22.5	2	1	0	4	1/2	1/2	0.2516
Age <= 35	5	3/5	2/5	1	1	0	0.1092
Car = Sports	3	1/3	2/3	3	1	0	0.4592
Car = Vintage	1	1	0	5	3/5	2/5	0.1092
Car = SUV	2	1	0	4	1/2	1/2	0.2516

First Split: Car = Sports

Frist Split:
Car = Sports

YES

NO

Point	Age	Car	**Risk**
$\mathbf{x}_1$	25	Sports	L
$\mathbf{x}_2$	20	Vintage	H
$\mathbf{x}_3$	25	Sports	L
$\mathbf{x}_4$	45	SUV	H
$\mathbf{x}_5$	20	Sports	H
$\mathbf{x}_6$	25	SUV	H

Car = Sports?

Y

x_1	25	L
x_3	25	L
x_5	20	H

Purity = 2/3 < 1
SPLIT AGAIN

N

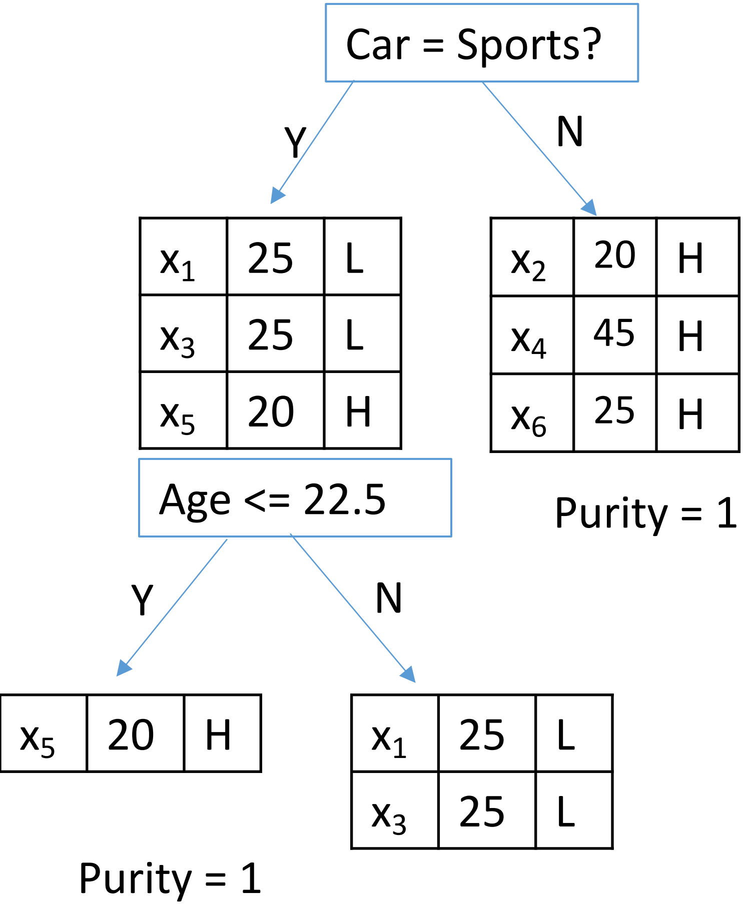
x_2	20	H
x_4	45	H
x_6	25	H

Purity = 1
LEAF NODE

Only Possible Split:
Age <= 22/5

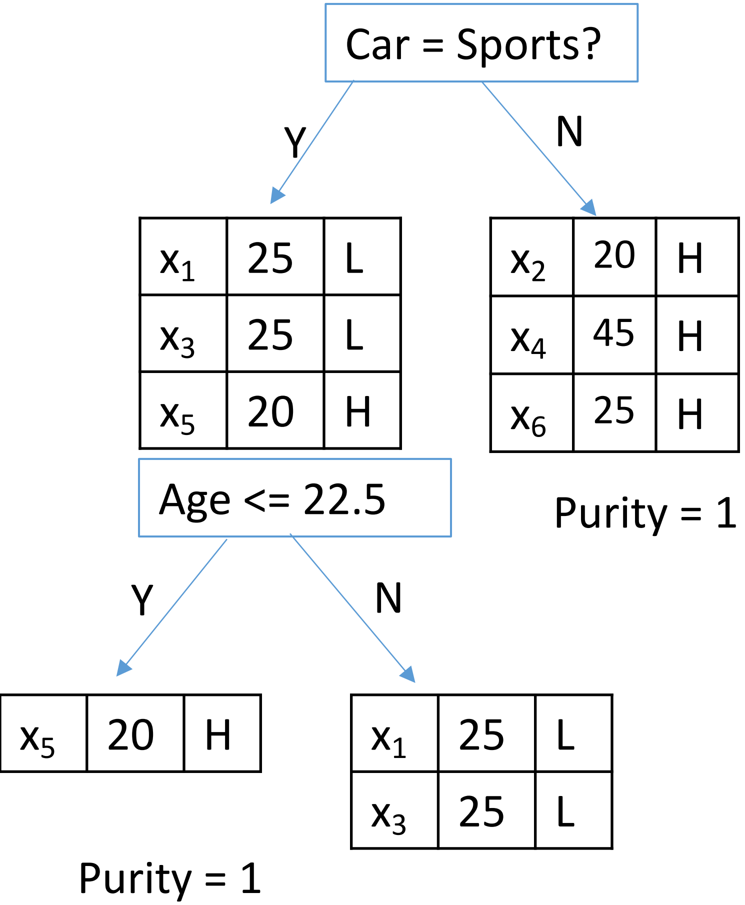
First Split: Car = Sports

Point	Age	Car	**Risk**
$\mathbf{x}_1$	25	Sports	*L*
$\mathbf{x}_2$	20	Vintage	*H*
$\mathbf{x}_3$	25	Sports	*L*
$\mathbf{x}_4$	45	SUV	*H*
$\mathbf{x}_5$	20	Sports	*H*
$\mathbf{x}_6$	25	SUV	*H*

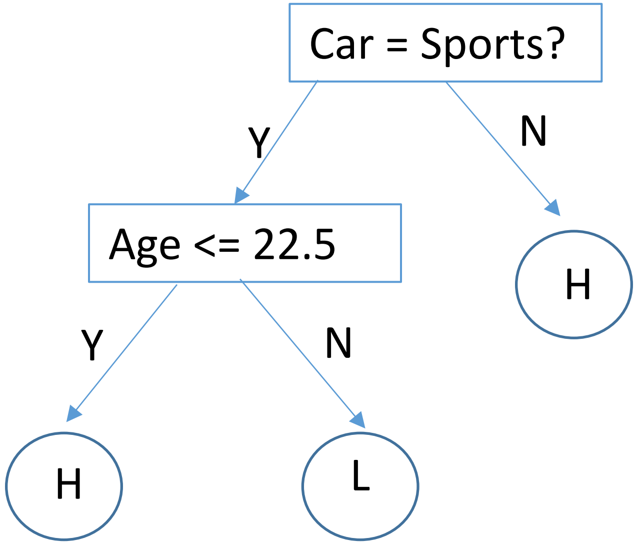


First Split: Car = Sports

Point	Age	Car	**Risk**
$\mathbf{x}_1$	25	Sports	L
$\mathbf{x}_2$	20	Vintage	H
$\mathbf{x}_3$	25	Sports	L
$\mathbf{x}_4$	45	SUV	H
$\mathbf{x}_5$	20	Sports	H
$\mathbf{x}_6$	25	SUV	H

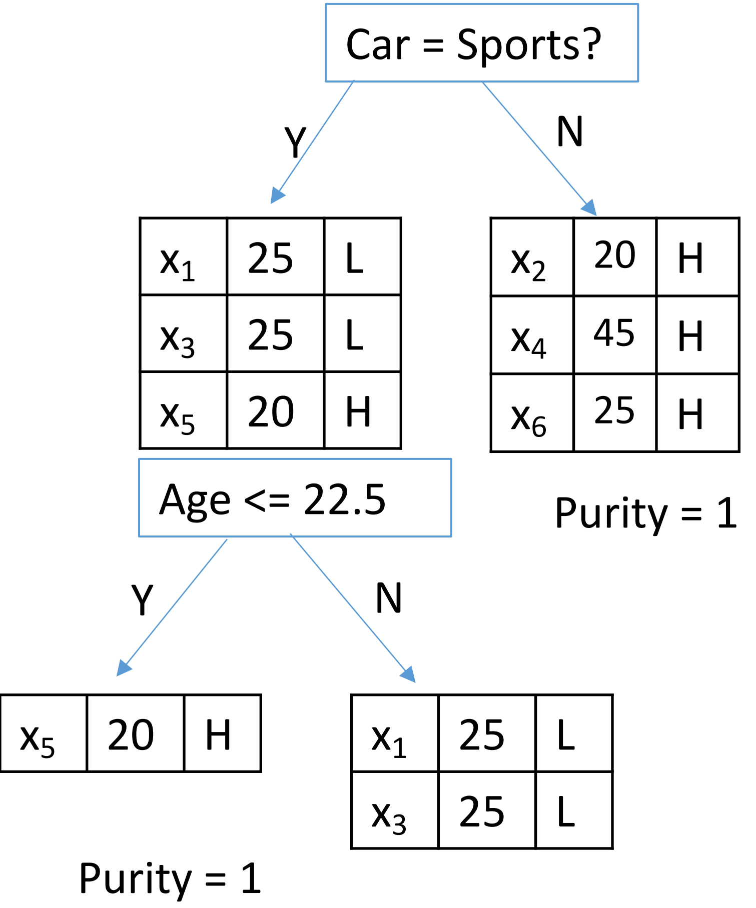


Final Classifier

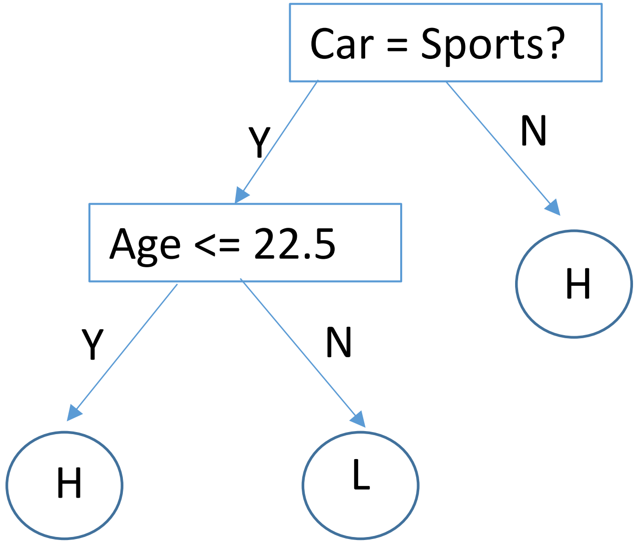


First Split: Car = Sports

Point	Age	Car	**Risk**
$\mathbf{x}_1$	25	Sports	L
$\mathbf{x}_2$	20	Vintage	H
$\mathbf{x}_3$	25	Sports	L
$\mathbf{x}_4$	45	SUV	H
$\mathbf{x}_5$	20	Sports	H
$\mathbf{x}_6$	25	SUV	H



Final Classifier



Classify: (Age = 27, Car = Vintage)
→ H: Since Car not = Sports