HOMEWORK 3

1.) Decision trees:

Multiway - split

Compare degree of impurity of parent node with children nodes

$$Gini(t) = 1 - \sum_{i=0}^{c-1} [p(i/t)]^2$$

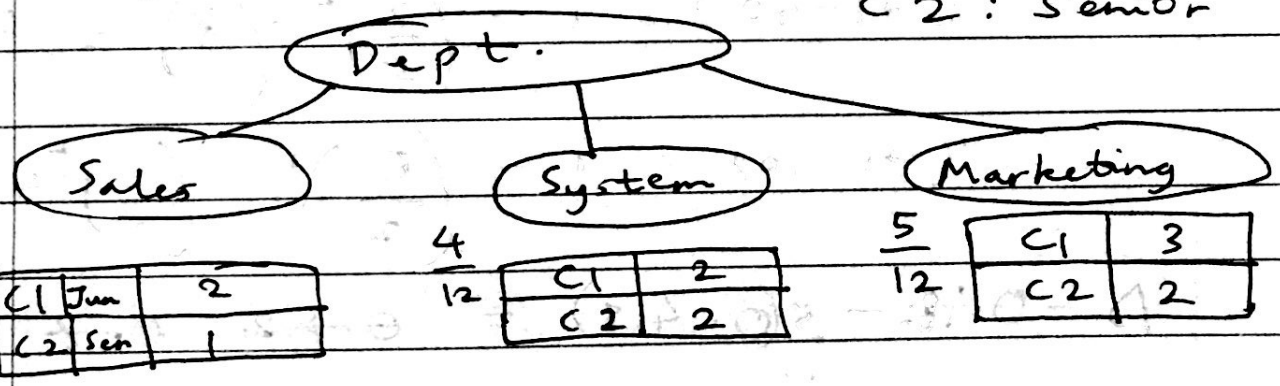
$$\Delta = I(parent) - \sum_{j=1}^{k} \frac{N(v_j)}{N} I(v_j)$$

Choose case with higher Δ gain

a.) Split with Dept:

C1: Junior

C2: Senior



Gini index sales $= 1 - \left(\frac{2}{3}\right)^2 - \left(\frac{1}{3}\right)^2 = 0.444$

system $= 1 - \left(\frac{2}{4}\right)^2 - \left(\frac{2}{4}\right)^2 = 0.5$

marketing $= 1 - \left(\frac{3}{5}\right)^2 - \left(\frac{2}{5}\right)^2 = 0.48$

Scanned by CamScanner

Gini Parent $= \frac{12}{12}$

C1	7
C2	5

$= 1 - \left(\frac{7}{12}\right)^2 - \left(\frac{5}{12}\right)^2$

$= 0.486$

$\Delta = 0.486 - \frac{3}{12}(0.444) - \frac{4}{12}(0.5) - \frac{5}{12}(0.48)$

$= \cancel{0.523}\ 0.009$

Split with Age:

Age

21-30

$\frac{5}{12}$

C1	5
C2	0

31-40

$\frac{5}{12}$

C1	2
C2	3

41-50

$\frac{2}{12}$

C1	0
C2	2

Gini $(21-30) = 1 - \left(\frac{5}{5}\right)^2 - \left(\frac{0}{5}\right)^2 = 0$

$(31-40) = 1 - \left(\frac{2}{5}\right)^2 - \left(\frac{3}{5}\right)^2 = 0.48$

$(41-30) = 1 - \left(\frac{0}{2}\right)^2 - \left(\frac{2}{2}\right)^2 = 0$

$\Delta = 0.486 - \frac{5}{12}(0.48) = \cancel{0.52}\ \cancel{0.8}$

$= 0.286$

Scanned by CamScanner

Split with Salary.

Salary

Low		Medium		High	
$\frac{3}{12}$	C1	3	$\frac{6}{12}$	C1	4	$\frac{3}{12}$	C1	0
	C2	0		C2	2		C2	2

$$\text{Gini (low)} = 1 - \left(\frac{3}{3}\right)^2 - \left(\frac{0}{3}\right)^2 = 0$$

$$\text{(Medium)} = 1 - \left(\frac{4}{6}\right)^2 - \left(\frac{2}{6}\right)^2 = 0.444$$

$$\text{(High)} = 1 - \left(\frac{0}{2}\right)^2 - \left(\frac{2}{2}\right)^2 = 0$$

$$\Delta = 0.486 - \frac{6}{12}(0.444) = \cancel{0.277}\ 0.264$$

Highest Gain with Split w.r.t. Age So we choose that.

Now for every node we try and split w.r.t. Dept and salary.

Dept.

21-30

Sales		Systems		Marketing	
$\frac{1}{5}$	C1	1	$\frac{2}{5}$	C1	2	$\frac{2}{5}$	C1	2
	C2	0		C2	0		C2	0

Scanned by CamScanner

Salary

21-30

Low		Medium		High	
C1	2	C1	3	C1	0
C2	0	C2	0	C2	0
$\frac{2}{5}$		$\frac{3}{5}$		0	

Gini (Sales) = 1 −

We can see here that Gini gain Δ will be same in both cases.

Also since we know all in age group 21-30 are Juniors.

Same is case with 41-50 all Seniors. So we can stop further iterations for them.

For 31-40

Dept

31-40

Sales		Systems		Marketing	
C1	1	C1	0	C1	1
C2	1	C2	1	C2	1

Gini (sales) $= 1 - \left(\frac{1}{2}\right)^2 - \left(\frac{1}{2}\right)^2 = 0.5$

(Systems) $= 1 - \left(\frac{1}{1}\right)^2 = 0$

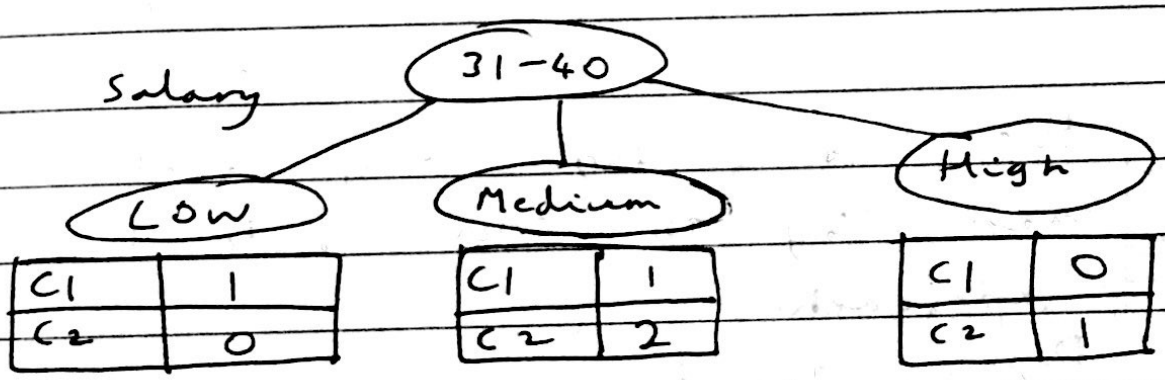
Scanned by CamScanner

$$(\text{Marketing}) = 1 - \left(\frac{1}{2}\right)^2 - \left(\frac{1}{2}\right)^2 = 0.5$$

$$\text{Parent Gini} = 1 - \left(\frac{2}{5}\right)^2 + \left(\frac{3}{5}\right)^2 = 0.48$$

$$\Delta = 0.48 - \frac{2}{5}(0.5) - \frac{2}{5}(0.5)$$

$$= 0.08$$



$$\text{Gini (low)} = 0$$

$$(\text{Medium}) = 1 - \left(\frac{1}{3}\right)^2 - \left(\frac{2}{3}\right)^2$$

$$= 0.444$$

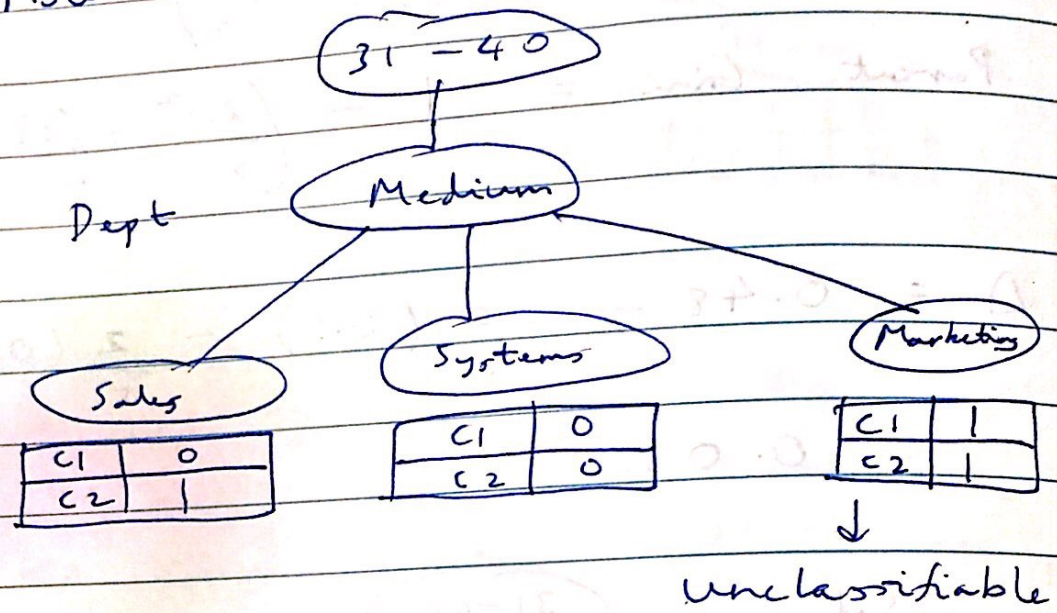
$$(\text{High}) = 0$$

$$\Delta = 0.48 - \frac{3}{5}(0.444)$$

$$= 0.213$$

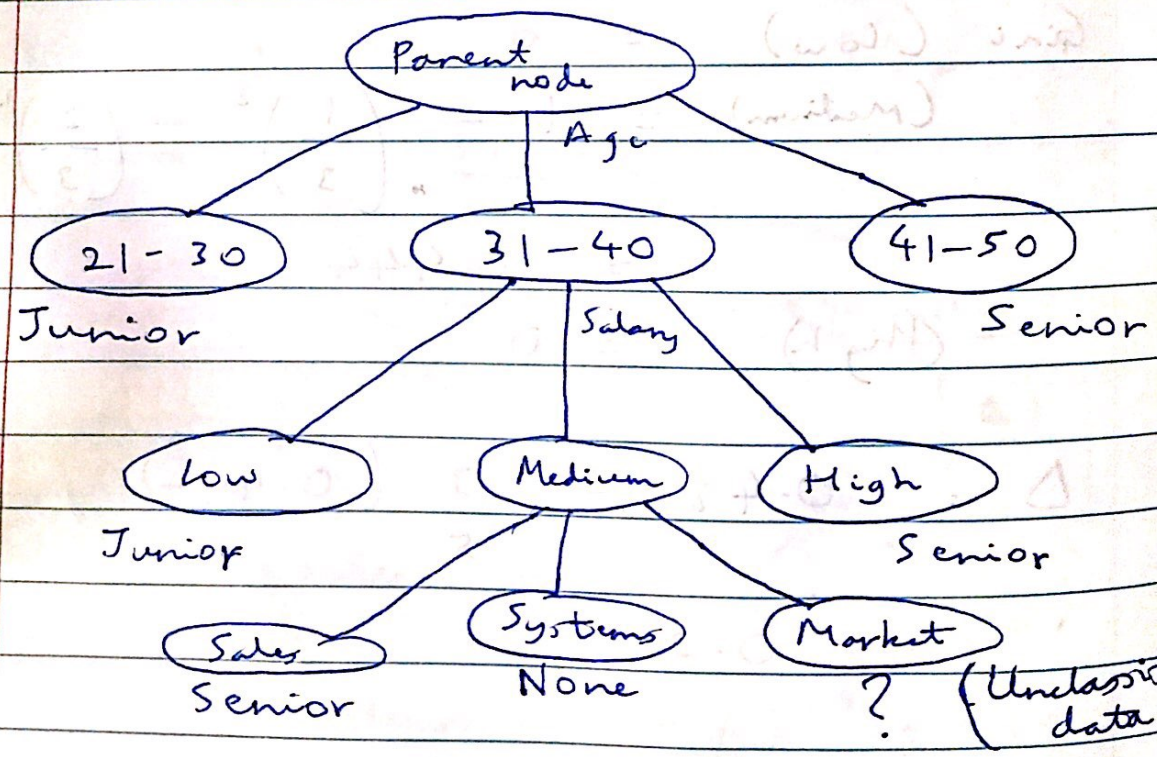
Scanned by CamScanner

Now we come to



We can assign any label here according to some rule-based approach.

Decision tree:



Scanned by CamScanner

b.) Assuming unclassified as error

i.) Individual class accuracy :

Junior $= \frac{6}{7}$

Senior $= \frac{4}{5}$

ii.) Overall accuracy $= \frac{10}{12}$

c.)

Sales	21 - 30	High	Junior
Systems	21 - 30	Medium	Junior
Marketing	41 - 50	High	Senior
Marketing	31 - 40	Low	Junior

Scanned by CamScanner

2.) Naïve Bayes Classification :

a.) Assumption:

Attributes are conditionally independent given the class label y.

$$P(X \mid Y=y) = \prod_{i=1}^{d} \left(P(X_i \mid Y=y) \right)$$

Attribute set $X = \{X_1, X_2 \ldots\ldots X_d\}$

b.) $$P(Y \mid X) = \frac{P(Y) \prod_{i=1}^{d} P(X_i \mid Y)}{P(X)}$$

$$P(A=0 \mid +) = \frac{P(A=0) \cdot P(+ \mid A=0)}{P(+)}$$

$$= \frac{1/2 \cdot 2/5}{1/2}$$

$$= \frac{2}{5}$$

Scanned by CamScanner

$$P(A=0 \mid -) = \frac{P(A=0) \cdot P(- \mid A=0)}{P(-)}$$

$$= \frac{1/2 \cdot 3/5}{1/2}$$

$$= \frac{3}{5}$$

$$P(A=1 \mid +) = \frac{P(A=1) \cdot P(+ \mid A=1)}{P(+)}$$

$$= \frac{1/2 \cdot 3/5}{1/2}$$

$$= \frac{3}{5}$$

$$P(A=1 \mid -) = \frac{P(A=1) \cdot P(- \mid A=1)}{P(-)}$$

$$= \frac{1/2 \cdot 2/5}{1/2}$$

$$= \frac{2}{5}$$

Scanned by CamScanner

$$P(B=0|+) = \frac{P(B=0) \cdot P(+|B=0)}{P(+)}$$

$$= \frac{7/10 \cdot 4/7}{1/2}$$

$$= \frac{4}{5}$$

$$P(B=0|-) = \frac{7/10 \cdot 3/7}{1/2}$$

$$= \frac{3}{5}$$

$$P(B=1|+) = \frac{3/10 \cdot 1/3}{1/2}$$

$$= \frac{1}{5}$$

$$P(B=1|-) = \frac{3/10 \cdot 2/3}{1/2}$$

$$= \frac{2}{5}$$

Scanned by CamScanner

$$P(C=0|+) = \frac{1/10 \cdot 1}{1/2}$$

$$= \frac{1}{5}$$

$$P(C=0|-) = \frac{1/10 \cdot 0}{1/2} = 0$$

$$P(C=1|+) = \frac{9/10 \cdot 4/9}{1/2} = \frac{4}{5}$$

$$P(C=1|-) = \frac{9/10 \cdot 5/9}{1/2} = 1$$

$P(A=0|+) = 2/5$

$P(A=1|+) = 3/5$

$P(B=0|+) = 4/5$

$P(B=1|+) = 1/5$

$P(C=0|+) = 1/5$

$P(C=1|+) = 4/5$

$P(A=0|-) = 3/5$

$P(A=1|-) = 2/5$

$P(B=0|-) = 3/5$

$P(B=1|-) = 2/5$

$P(C=0|-) = 0$

$P(C=1|-) = 1$

Scanned by CamScanner

c.) $A = 0,\ B = 1,\ C = 0$

Method - I :

Let's check for $-$ & $+$

Considering the occurrence of this instance

If negative $-$; $P(-)$: A, B, C.

All are independent so.

$$P(-|A=0) \cdot P(-|B=1) \cdot P(-|C=0)$$

$$= \frac{P(-) \cdot P(A=0|-)}{P(A=0)} \cdot \frac{P(-) \cdot P(B=1|-)}{P(B=1)} \cdot \frac{P(-) \cdot P(C=0)}{P(C=0)}$$

$$= \frac{1/2 \cdot 3/5}{1/2} \cdot \frac{1/2 \cdot 2/5}{3/10} \cdot \frac{1/2 \cdot 0}{1/10}$$

$$= \frac{3}{5} \cdot \frac{2}{3} \cdot \frac{0}{9} = \frac{0}{9} = 0$$

If $+$; $P(+)$

$$P(+|A=0) \cdot P(+|B=1) \cdot P(+|C=1)$$

$$= \frac{2}{5} \cdot \frac{1}{3} \cdot \frac{4}{5} = \frac{8}{25}$$

Scanned by CamScanner

Since $P(+) > P(-)$

we can estimate that class label is "+".

d.) m- estimate approach:

$p = 0.5, \; m = 4$

$$p(x_i | y_j) = \frac{n_c + mp}{n + m}$$

$n \rightarrow$ total instances from class y_j

$n_c \rightarrow$ training examples from y_j that take on value x_i.

$m \rightarrow$ equivalent sample size.

$p \rightarrow$ user specified.

$$P(A=0|+) = \frac{2+2}{5+4} = \frac{4}{9}$$

$$P(A=0|-) = \frac{3+2}{5+4} = \frac{5}{9}$$

$$P(A=1|+) = \frac{3+2}{5+4} = \frac{5}{9}$$

$$P(A=1|-) = \frac{2+2}{5+4} = \frac{4}{9}$$

Scanned by CamScanner

$$P(B=0|+) = \frac{4+2}{5+4} = \frac{2}{3}$$

$$P(B=0|-) = \frac{3+2}{5+4} = \frac{5}{9}$$

$$P(B=1|+) = \frac{1+2}{5+4} = \frac{1}{3}$$

$$P(B=1|-) = \frac{2+2}{5+4} = \frac{4}{9}$$

$$P(C=0|+) = \frac{1+2}{5+4} = \frac{1}{3}$$

$$P(C=0|-) = \frac{0+2}{5+4} = \frac{2}{9}$$

$$P(C=1|+) = \frac{4+2}{5+4} = \frac{2}{3}$$

$$P(C=1|-) = \frac{5+2}{5+4} = \frac{7}{9}$$

Scanned by CamScanner

e.) $A = 0, B = 1, C = 0$

Method – II

We can choose to ignore the denominator $P(x)$ as it is not part of the condition, remaining constant.

If we assume $P(A = 0, B = 1, C = 0) = p$

then.

$$P(+ \mid A=0, B=1, C=0) = P(A=0, B=1, C=0 \mid +) \times P(+)/p$$

$$= P(A=0 \mid +) \times P(B=1 \mid +) \times P(C=0 \mid +) \times P(+)/p$$

$$= \frac{4/9 \times 1/3 \times 1/3 \times 1/2}{p} = 0.0247/p$$

$$P(- \mid A=0, B=1, C=0) = P(A=0 \mid -) \times P(B=1 \mid -) \times P(C=0 \mid -) \times P(-)/p$$

$$= 5/9 \times \frac{4}{9} \times \frac{2}{9} \times 1/2 \,/p = 0.0274/p$$

Scanned by CamScanner

$P(-) > P(+)$

$\therefore$ Class label is '−'.

f.) m-estimate probability approach is better as the entire posterior probability of an instance becomes zero if one of the conditional probabilities is zero, in case of basic approach, which may lead to misclassification. As seen in the example above.

Scanned by CamScanner

3) Holt's 1-Rule method

Rule for value 0 in A

Most frequent class c is '−'

Clause: If A = 0 class is '−'

Classification accuracy $\frac{4}{5}$

Rule for value 1 in A

Most frequent class c is '+'

Clause: If A = 1 class is '+'.

Classification accuracy $\frac{3}{5}$

Rule for value 0 in B

Most frequent class c is '−'

Clause: If B = 0 class is '−'

Classification accuracy = $\frac{4}{7}$

Rule for value 1 in B

Most frequent clause c is '−'

Clause: If B = 1 class is '−'

Classification accuracy = $\frac{2}{3}$

Scanned by CamScanner

Rule for $C = 0$
Most frequent class c is '–'
Clause: If $C = 0$ class is '–'
Classification accuracy $= \frac{3}{5}$

Rule for value 1 in C
Most frequent class c is '–'
Clause: If $C = 1$ class is '–'
Classification accuracy $= \frac{3}{5}$

i.) Rule for attribute 'A' is
If $A = 0$ class is '–', otherwise class is '+' ($A = 1$)

ii.) Rule for attribute 'B' is
If $B = 1$ class is '–' otherwise class is '+' ($B = 0$)

iii.) Rule for attribute 'C' is
If $C = 0$ class is '–' otherwise class is '+' ($C = 1$) [Actually undecided as accuracy is same]

Scanned by CamScanner

b.) 'A' is the best attribute.
as it has highest accuracy,
total error is minimum.

Scanned by CamScanner