Lecture 20

Tuesday, October 25, 2022 12:07 PM

Till hunt's algo in Midsem
current slides till pg. 45

Split Conditions & outcomes

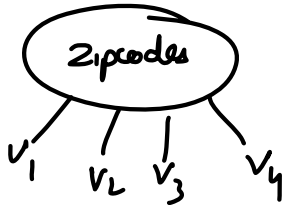
Ordinal, Nominal, Continuous, Binary

Nominal attributes → Gender → Male, Female

Binary

Multiway split

Zipcodes → v_1, v_2, v_3, v_4



Binary split

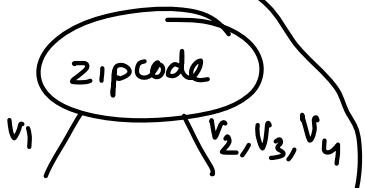
2^{k-1}

k = unique numbers

$k = 4$

$2^{4-1} - 1 = 7$

v_1	$v_2\ v_3\ v_4$
v_2	$v_1\ v_3\ v_4$
v_3	$v_1\ v_2\ v_4$
v_4	$v_1\ v_2\ v_3$
$v_1 v_2$	$v_3 v_4$
$v_1 v_4$	$v_2 v_3$
$v_1 v_3$	$v_2 v_4$



Ordinal attributes

Binary

Multiway split

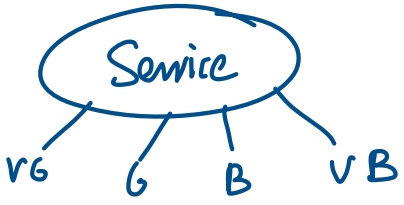
Service → very good, good, bad, very bad

Order property must be maintained while grouping values for binary split.

VG	G,B,VB	possible splits, maintain order property ✓✓
VG, G	B, VB	
VG, G, B	VB	

VG, B G,VB → doesn't maintain order ✗✗

for multiway split, its the same



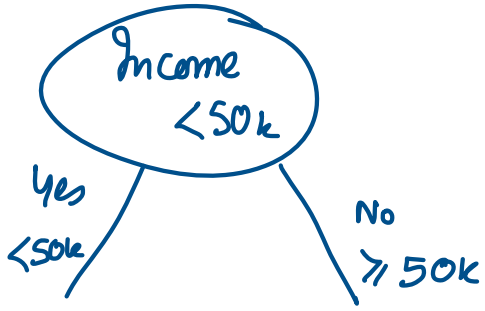
Binary Attributes

binary splits & no multiway, cuz all we have is 2 ✓

Continuous Attributes

Binary split

Split value → (v)



(*) we get v value by checking it with all the values in the column & find which one splits it the best by using certain measures

Multiway Split

10k → 100k

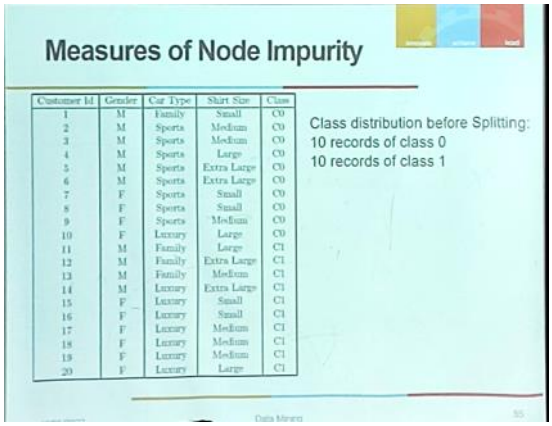
< 10k, 10k-25k, 25k-50k, 50k-80k, >= 80k

(*) split it into the ranges & select the one with most efficiency.

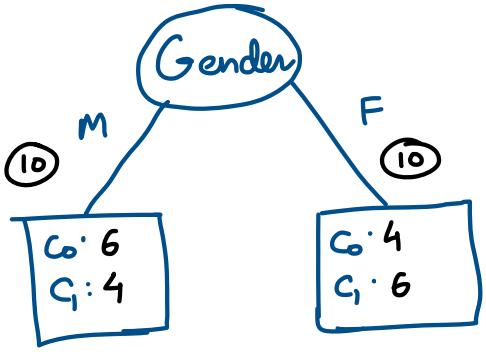
Measures of Node Impurity

class distribution — before splitting
                   — after splitting

class — C_0
      — C_1



(1) Gender



C_0	5
C_1	5

* This is uniform class distribution

MOST IMPURE SPLIT

C_0	10
C_1	0

* Most skewed distribution

PUREST SPLIT

{IDEAL}

is that a word?

What we need is the measure of the skewedness of the split

Impurity Measure

t → node

i → class

P(i/t) → fraction of records belonging to i @ node t

c → no. of classes

* Entropy $= -\sum_{i=0}^{c-1} p_i \log_2 p_i$

Gini $= 1 - \sum_{i=0}^{c-1} p_i^2$

} calculated in example.

(*) Gini = $1 - \sum_{i=0}^{c-1} p_i^2$

(*) Classification Error = $1 - \max_i \{p_i\}$

in example, don't panic (xD)

$\log_2(5/6) = \frac{\log(5/6)}{\log 2}$

Measures of Node Impurity

Node N_1	Count
Class=0	0
Class=1	6

Gini $= 1 - (0/6)^2 - (6/6)^2 = 0$
Entropy $= -(0/6)\log_2(0/6) - (6/6)\log_2(6/6) = 0$
Error $= 1 - \max[0/6, 6/6] = 0$

Node N_2	Count
Class=0	1
Class=1	5

Gini $= 1 - (1/6)^2 - (5/6)^2 = 0.278$
Entropy $= -(1/6)\log_2(1/6) - (5/6)\log_2(5/6) = 0.650$
Error $= 1 - \max[1/6, 5/6] = 0.167$

Node N_3	Count
Class=0	3
Class=1	3

Gini $= 1 - (3/6)^2 - (3/6)^2 = 0.5$
Entropy $= -(3/6)\log_2(3/6) - (3/6)\log_2(3/6) = 1$
Error $= 1 - \max[3/6, 3/6] = 0.5$

Node 1

$C_0 \rightarrow 0$

$C_1 \rightarrow 6$

$P_0 \rightarrow 0/6 = 0$

$P_1 \rightarrow 6/6 = 1$

Entropy $= -[0 \log 0 + 1 \log 1] = 0$

Gini $= 1 - [0^2 + 1^2] = 0$

Gini = $1 - [0^2 + 1^2] = 0$

Classification Error = $1 - \max\{0, 1\} = 1 - 1 = 0$

Node 2

$C_0 \rightarrow 1 \quad P_0 = 1/6$

$C_1 \rightarrow 5 \quad P_1 = 5/6$

Entropy = $-\left[\frac{1}{6}\log 6 + \frac{5}{6}\log \frac{5}{6}\right] = 0.65$

Gini = $1 - \left[\left(\frac{1}{6}\right)^2 + \left(\frac{5}{6}\right)^2\right] = 1 - \frac{26}{36} = 0.278$

Classification Error = $1 - \max\{1/6, 5/6\} = 1/6$

Homework {Verify with slides}

Node 3