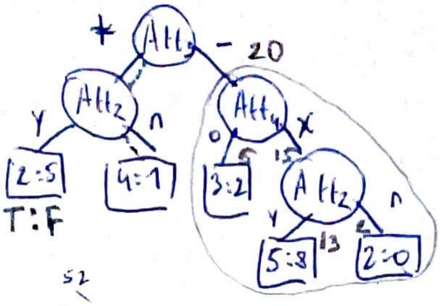
ML - DM mid2. exe:




T

260

1. Decision trees

200: $p(T): \frac{4}{5} = \underline{0.8}$ ✓

$p(F): \frac{1}{5} = 0.2$

52

201: $p(T): \frac{\cancel{5}}{20} \cdot \frac{3}{\cancel{5}} + \frac{15}{\cancel{20}_4} \cdot \frac{\cancel{5}}{13} = 0.15 + 0.29 = 0.44$

$p(F): \frac{\cancel{5}}{20} \cdot \frac{2}{\cancel{5}} + \frac{15}{20} \cdot \frac{8}{13} =$ ~~0.025 + 0.46 = 0.52~~ $0.1 + 0.46 = \underline{\underline{0.56}}$

202: $p(T): \frac{\cancel{13}}{15} \cdot \frac{5}{\cancel{13}} + \frac{2}{15} \cdot 1 = 0.33 + 0.13 = 0.46$

$p(F): \frac{\cancel{13}}{15} \cdot \frac{8}{\cancel{13}} + \frac{2}{\cancel{13}} \cdot 0 = 0.53$

Using Decision Rules (PRISM

2. i rule: IF (Att2 = n) AND (Att4 = 0) AND (Att3 = -) THEN T

we do not count the row containing this values, delete it.

choose biggest → $\boxed{Att_2(n)}$: [3/5] ~ 0.6 ✓
- $Att_2(y)$: [1/5] ~ 0.2
- $Att_3(+)$: [3/6] ~ 0.5
- $Att_3(-)$: [1/4] ~ 0.25
- $Att_4(x)$: [2/5] ~ 0.4
- $Att_4(0)$: [2/5] ~ 0.4

Total: 4/10

number of class=T ; total number of rows

consider rows that have $Att_2 = n$

- $\boxed{Att_3 = +}$ [2/3]
- $Att_3 = -$ [1/2]
- $Att_4 = x$ [1/2]
- $Att_4 = 0$ [2/3]

Total: 3/5

n where class=T ; total n

for $Att_3 = +$

- $Att_4 = x$ [0/0]
- $\boxed{Att_4 = 0}$ [2/3]

number of rows where $Att_4 = 0$ & class = T. ; number of rows with $Att_4 = 0$

ii rule: IF (Att2 = n) AND (Att3 = +) AND (Att4 = 0) THEN T

| | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| 1 | 0 | 2 | 7 | 8 | 7 |
| 2 | | 0 | 5 | 6 | 5 |
| 3 | | | 0 | 1 | 2 |
| 4 | | | | 0 | 1 |
| 5 | | | | | 0 |

→ 1

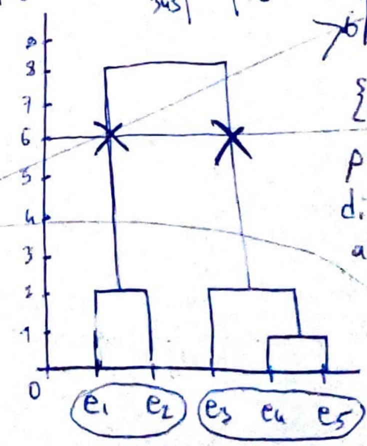
| | 1 | 2 | 3 | 45 |
|---|---|---|---|---|
| 1 | 0 | 2 | 7 | 8 |
| 2 | | 0 | 5 | 6 |
| 3 | | | 0 | 2 |
| 45 | | | | 0 |

→ 2

| | 12 | 3 | 45 |
|---|---|---|---|
| 12 | 0 | 7 | 8 |
| 3 | | 0 | 2 |
| 45 | | | 0 |

→ 3

| | 12 | 345 |
|---|---|---|
| 12 | 0 | 8 |
| 345 | | 0 |




6) 2 groups: {e1, e2}, {e3, e4, e5}

pick biggest distance, as at the end of this arrow

using:
- hierarchical aglomerative
- manhattan distance
- complete linkage (when joining, take the biggest distance)

if using single linkage, take smallest distance

1, 2, 2, 8 biggest diff.

1 0 6

| TID | Products | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 22149 | A, | B, | C, | L, | P, | Q, | U, | X |
| 33277 | A, | B, | C, | | P, | R, | | X |
| 44305 | A, | | C, | | | T, | V, | X |
| 57493 | A, | | | | P, | | V, ... | Z |
| 55511 | A, | B, | | | P, | | V, | X |

4.

| ID | Att1 | 2 | 3 | 4 |
|---|---|---|---|---|
| 300 | ~~2015~~ | - | - | - |

if att4,3,2,1, class are different for every ~~val~~ id for 300, write 1, else 0, you get:

| big table | |
|---|---|
| | $\sqrt{0^2+1^2+1^2+0^2} = \sqrt{2}$ |
| 2004.401 | $\sqrt{1} = 1 \leftarrow$ |
| | $\sqrt{3}$ |
| 2015.735 | $0 \leftarrow$ |
| | $\sqrt{3}$ |
| | $\sqrt{4} = 2$ |
| 2020.000 | $\sqrt{1} = 1 \leftarrow$ |
| | $\sqrt{2}$ |
| | $\sqrt{2}$ |
| 2019.186 | $0 \leftarrow$ |
| | $\sqrt{3}$ |

find average

we pick 4 smallest because our $k = 4$.

this is if the class? or attributes are nominal. If they are numeric, pick most dominant ids

$$\frac{2004.401 + 2015.735 + 2020 + 2019.186}{4} = 2014.83275$$

- we divide by the number of closest neighbours, not with k.
- if attribute from small table has the same value for attribute from big table, then 0, 1 otherwise. ex:

| ~~No~~ ID | Att1 | Att2 | Att3 | Att4 | Class |
|---|---|---|---|---|---|
| 110 | 2008.7 | Y | - | 0 | F |
| 300 | | N | + | 0 | T |

$\sqrt{1^2 + 1^2 + 0^2 + 1^2} = \sqrt{3}$

repeat for each

5. Association rules.

a) K-itemsets, minimum support 60 % = 0.6

| 1-itemsets | 2-itemsets | 3-itemsets |
|---|---|---|
| A P | AB ~~BC~~ ~~CV~~ | ~~ABC~~ ~~ACV~~ ~~BCP~~ ~~BVX~~ |
| B V | AC BP CX | ABP ACX ~~BCV~~ ~~CPV~~ |
| C X | AP ~~BV~~ ~~PV~~ | ~~ABV~~ ~~APV~~ ~~BCX~~ ~~CPX~~ |
| | AV BX PX | ABX APX ~~BPV~~ ~~CVX~~ |
| | AX ~~CP~~ ~~VX~~ | ~~ACP~~ ~~AVX~~ BPX ~~PVX~~ |
| 6 | 9 | 5 |

4-itemsets

~~BPCX~~
~~APCX~~
~~ABCX~~
~~ABPC~~
ABPX

1

6 + 9 + 5 + 1 = 21 itemsets

b) $2^k - 2$

$6 \cdot 0 + 9 \cdot 2 + 5 \cdot 6 + 1 \cdot 14 = 62$ associat rules

# of itemsets

2 to the power of which itemset

multiply, number of k-itemsets, we got above 6 for 1-itemset, 9 for 2-itemset, with 2 to the power of k. ~~if for 6,~~

we multiply 6 with 0 because $2^1 - 2 = 0$, 9 with two because $2^2 - 2 = 2$.

c) {A, B, P}: minimum confidence = 0.8 (80 %)

$A \to BP$ $3/5 = 0.6$

$\boxed{B \to AP}$ $3/3 = 1$ ✓

$P \to AB$ $3/4 = 0.75$

$\boxed{BP \to A}$ $3/3 = 1$ ✓

$AP \to B$ $3/4 = 0.75$

$\boxed{AB \to P}$ $3/3 = 1$ ✓

$\boxed{\text{confidence}(A \to BP) = \frac{\sigma(\{A,B,P\})}{\sigma(\{A\})}}$

support

upper part is how many items have all A, B, P attributes, and lower part is number of occurences of the set left of the arrow.

d) $\text{Lift}(A \to BP) = \frac{\text{conf.}(A \to BP)}{\sigma(\{BP\})}$

$\boxed{B \to AP}$ $1/\frac{4}{5} = \frac{5}{4} = 1.25$

$BP \to A$: $1/\frac{3}{3} = \frac{1}{1} = 1$

$\boxed{AB \to P}$ $1/\frac{4}{5} = \frac{5}{4} = 1.25$ ✓

"enough" = lift > 1