Mamoutou

Sangare

V00010526

SENG 474 ASSIGNMENT 1

1- **Construct the root and the first level of a decision tree for the contact lenses data. Use the ID3 algorithm. Show the details of your construction. Then, check your solution with Weka (the data file is included with Weka).**

Please see the end for weka output for verification purpose for the first tree exercises

P 1

Assignment 1

1) Attribute Age = young

info $([4,2,2]) = \text{entropy}\left(\frac{4}{8}, \frac{2}{8}, \frac{2}{8}\right) = \text{entropy}\left(\frac{1}{2}, \frac{1}{4}, \frac{1}{4}\right)$

$= -\frac{1}{2}\log_2\left(\frac{1}{2}\right) - \frac{1}{4}\log_2\left(\frac{1}{4}\right) - \frac{1}{4}\log_2\left(\frac{1}{4}\right)$

$= 1.5$

Attribute Age = pre-presbyopic

info $([5,2,1]) = \text{entropy}\left(\frac{5}{8}, \frac{2}{8}, \frac{1}{8}\right) = \text{entropy}\left(\frac{5}{8}, \frac{1}{4}, \frac{1}{8}\right)$

$= -\frac{5}{8}\log_2\left(\frac{5}{8}\right) - \frac{1}{4}\log_2\left(\frac{1}{4}\right) - \frac{1}{8}\log_2\left(\frac{1}{8}\right)$

$= 1.300$

Attribute Age = presbyopic

info $([6,1,1]) = \text{entropy}\left(\frac{6}{8}, \frac{1}{8}, \frac{1}{8}\right) = \text{entropy}\left(\frac{3}{4}, \frac{1}{8}, \frac{1}{8}\right)$

$= -\frac{3}{4}\log_2\left(\frac{3}{4}\right) - \frac{1}{8}\log_2\left(\frac{1}{8}\right) - \frac{1}{8}\log_2\left(\frac{1}{8}\right)$

$= 1.061$

Expected info

info $([4,2,2],[5,2,1],[6,1,1]) = 1.5\left(\frac{8}{24}\right) + 1.3\left(\frac{8}{24}\right) + 1.061\left(\frac{8}{24}\right)$

$= \boxed{1.287}$

Attribute Spectacle-prescrip = myope

info $([7,2,3]) = -\frac{7}{12}\log_2\left(\frac{7}{12}\right) - \frac{2}{12}\log_2\left(\frac{2}{12}\right) - \frac{3}{12}\log_2\left(\frac{3}{12}\right)$

$= 1.384$

Spectacle-prescrip = hypermetrope

info $([8,3,1]) = \text{Entropy}\left(\frac{8}{12}, \frac{3}{12}, \frac{1}{12}\right)$

$= -\frac{8}{12}\log_2\left(\frac{8}{12}\right) - \frac{3}{12}\log_2\left(\frac{3}{12}\right) - \frac{1}{12}\log_2\left(\frac{1}{12}\right)$

$$\text{entropy}\left(\frac{8}{12}, \frac{3}{12}, \frac{1}{12}\right) = 1.189$$

$$\text{Expected info} = \frac{12}{24}(1.384) + \frac{12}{24}(1.189)$$

So $\text{info}([7,2,3],[8,3,1]) = \boxed{1.2865}$

Attribute astigmation

Astigmation = yes

$$\text{info}([8,4,0]) = \text{entropy}\left(\frac{8}{12}, \frac{4}{12}, \frac{0}{12}\right) = \text{entropy}\left(\frac{2}{3}, \frac{1}{4}, 0\right)$$

$$= -\frac{2}{3}\log_2\left(\frac{2}{3}\right) - \frac{1}{4}\log_2\left(\frac{1}{4}\right) - 0$$

$$= 0.890$$

Astigmation = no

$$\text{info}([7,5,0]) = \text{entropy}\left(\frac{7}{12}, \frac{5}{12}, 0\right)$$

$$= -\frac{7}{12}\log_2\left(\frac{7}{12}\right) - \frac{5}{12}\log_2\left(\frac{5}{12}\right) - 0$$

$$= 0.980$$

Expected info

$$\text{info}([8,4,0],[7,5,0]) = \frac{12}{24}(0.890) + \frac{12}{24}(0.980)$$

$$= \boxed{0.935}$$

Attribute tear-prod-rate

tear-prod-rate = reduce

$$\text{info}([12,0,0]) = \text{entropy}\left(\frac{12}{12}, 0, 0\right)$$

$$= -1\log_2(1) - 0 - 0 = 0$$

tear-prod-rate = normal

$$\text{info}([3,5,4]) = \text{entropy}\left(\frac{3}{12}, \frac{5}{12}, \frac{4}{12}\right)$$

$$= -\frac{3}{12}\log_2\left(\frac{3}{12}\right) - \frac{5}{12}\log_2\left(\frac{5}{12}\right) - \frac{4}{12}\log_2\left(\frac{4}{12}\right) = 1.555$$

Expected infor is P2

$$info([12,0,0],[3,5,4]) = \frac{12}{24}(0) + \frac{12}{24}(1.555)$$
$$= \boxed{0.7775}$$

Since tear-prod-rate has the smallest entropy, it is the root of the tree.

tear-prod-rate

Now we continue to split for tear-prod-rate ~~is~~ is normal. Using a smaller data set

Attribute Age

Age = young

$$info([2,2,0]) = entropy\left(\frac{2}{4},\frac{2}{4},0\right) = 1$$

Age = pre-presbyopic

$$info([2,1,1]) = entropy\left(\frac{2}{4},\frac{1}{4},\frac{1}{4}\right)$$
$$= -\frac{1}{2}\log_2\left(\frac{1}{2}\right) - \frac{1}{4}\log_2\left(\frac{1}{4}\right) - \frac{1}{4}\log_2\left(\frac{1}{4}\right)$$
$$= 1.5$$

Age = presbyopic

$$info([2,1,1]) = entropy\left(\frac{2}{4},\frac{1}{4},\frac{1}{4}\right) = 1.5$$

Expected info is

$$info([2,2,0],[2,1,1],[2,1,1]) = \frac{4}{12}\times 1 + \frac{4}{12}(1.5) + \frac{4}{12}(1.5)$$
$$= \boxed{1.333}$$

Attribute = spectacle-prescrip

Spectacle-prescrip = myope

$info([2,3,1]) = entropy\left(\frac{2}{6}, \frac{3}{6}, \frac{1}{6}\right)$

$= -\frac{1}{3}\log_2\left(\frac{1}{3}\right) - \frac{1}{2}\log_2\left(\frac{1}{2}\right) - \frac{1}{6}\log_2\left(\frac{1}{6}\right)$

$= 1.459$

~~info~~ Spectacle-prescrip = hypermetrope

$info([3,1,2]) = entropy\left(\frac{3}{6}, \frac{1}{6}, \frac{2}{6}\right) = 1.459$

Expected info is $\frac{6}{12}(1.459) + \frac{6}{12}(1.459) = \boxed{1.459}$

Attribute: astigmatism

astigmatism = no

$info([5,0,1]) = entropy\left(\frac{5}{6}, 0, \frac{1}{6}\right)$

$= -\frac{5}{6}\log_2\left(\frac{5}{6}\right) - 0 - \frac{1}{6}\log_2\left(\frac{1}{6}\right) = 0.650$

astigmatism = yes

$info([4,0,2]) = entropy\left(\frac{4}{6}, 0, \frac{2}{6}\right)$

$= -\frac{2}{3}\log_2\left(\frac{2}{3}\right) - \frac{1}{3}\log_2\left(\frac{1}{3}\right) = 0.918$

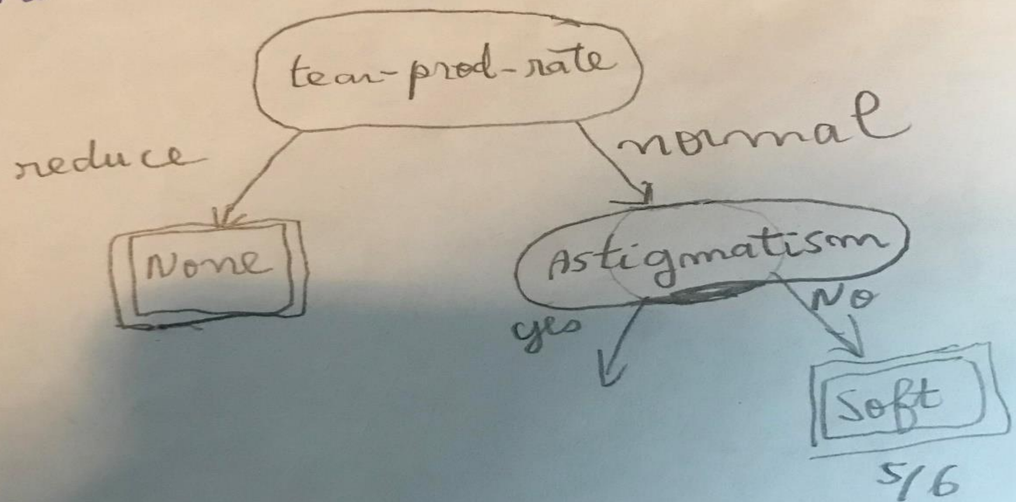
Expected information is

$info([5,0,1],[4,0,2]) = 0.650 \times \frac{6}{12} + 0.918 \times \frac{6}{12}$

$= \boxed{0.784}$

Since astigmatism is the smallest attribute, it is the winner!

The tree looks like this so far:



2_ Construct two rules using Prism

Rule we seek : If ? then Play = yes

Possible Test

outlook = Sunny	$\frac{2}{5}$	humidity = high	3/7
outlook = raining	$\frac{3}{5}$	humidity = Normal	$\frac{6}{7}$
outlook = overcast	$\frac{4}{4}$	windy = False	$\frac{6}{8}$
Temperature = hot	2/4	windy = True	$\frac{3}{6}$
Temperature = mild	4/6		
Temperature = Cool	3/4		

we choose outlook = overcast since it has the highest probability $(\frac{4}{4})$

Rule 1: $(\text{outlook} = \text{overcast}) \wedge (\cdots) \Rightarrow \text{Play} = \text{yes}$

Instances covered so far is:

outlook	Temperature	Humidity	Windy	Play
overcast	Hot	High	False	yes
overcast	Cool	Normal	True	yes
overcast	Mild	High	True	yes
overcast	Hot	Normal	False	yes

The rule is very accurate, getting 4 out of 4

so R_1 : outlook = overcast → yes

Second rule for recommending "yes" is built from instance not covered by R_1

Rule we seek :
If ? then "yes"

outlook = Sunny $\frac{2}{5}$
outlook = raining $\frac{3}{5}$
Temperature = hot $\frac{0}{2}$
Temperature = mild $\frac{3}{4}$
Temperature = Cool $\frac{2}{3}$
Humidity = High $\frac{1}{5}$
Humidity = Normal $\frac{4}{5}$
Windy = False $\frac{4}{6}$
Windy = True $\frac{1}{4}$

We pick humidity = Normal
R_2 = (humidity = Normal) ∧ (?) → yes
Instances covered so far =

outlook	Temperature	Humidity	windy	Play
Rainy	Cool	Normal	False	yes
Rainy	Cool	Normal	True	No
Sunny	Cool	Normal	False	yes
Rainy	Mild	Normal	False	yes
Sunny	Mild	Normal	True	yes

outlook = raining $\frac{2}{3}$
outlook = Sunny $\frac{2}{2}$
Temperature = Cool $\frac{2}{3}$
Temperature = Mild $\frac{2}{2}$

Windy = False $\frac{3}{3}$
Windy = True $\frac{1}{2}$
we pick windy = False

P5

Thus $R_2 = (\text{humidity} = \text{Normal} \wedge \text{windy} = \text{False}) \rightarrow \text{Play}$

Therefore

$$\boxed{\begin{array}{l}(\text{outlook} = \text{overcast}) \rightarrow \text{"yes"} \\ (\text{humidity} = \text{Normal} \wedge \text{Windy} = \text{False}) \rightarrow \text{"yes"}\end{array}}$$

We now do the same for play = No

Rule we seek; If ? then Play = No

Possible Test

outlook = Sunny	$\frac{3}{5}$
outlook = overcast	$\frac{0}{4}$
outlook = raining	$\frac{2}{5}$
Temperature = hot	$\frac{2}{4}$
Temperature = mild	$\frac{2}{6}$
Temperature = cool	$\frac{1}{4}$
humidity = high	4/7
humidity = Normal	$\frac{1}{7}$
Windy = False	$\frac{2}{8}$
windy = True	$\frac{3}{6}$

True (Since it has the highe

We pick outlook = sunny

Thus, (outlook = sunny) $\wedge$ (...) $\rightarrow$ No

Instances covered so far

Outlook	Temperature	Humidity	Windy	Play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Sunny	Mild	High	False	No
Sunny	Cool	Normal	False	Yes
Sunny	Mild	Normal	True	Yes

Repeat the same process

Temperature = hot $\frac{2}{2}$

Temperature = Mild $\frac{1}{2}$

Temperature = Cool $\frac{0}{1}$

Humidity = high $\frac{3}{3}$

Humidity = Normal $\frac{0}{2}$

Windy = False $\frac{2}{3}$

Windy = True $\frac{1}{2}$

We pick humidity = high (Coverage $\frac{3}{3}$)

Therefore $\boxed{R_1 : (\text{outlook} = \text{Sunny}) \wedge (\text{humidity} = \text{high}) \rightarrow N}$

For the Second Rule (R_2) (with 11 instances)

outlook = overcast, $\frac{0}{4}$

outlook = sunny, $\frac{0}{2}$

outlook = Rainy, $\frac{2}{5}$

temperature = hot, $\frac{0}{2}$

temperature = Mild, $\frac{1}{5}$

temperature = cool $\frac{1}{4}$

Humidity = High $\frac{1}{4}$

Humidity = Normal $\frac{1}{7}$

Windy = False $\frac{0}{6}$

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

we can pick either outlook = rainy or windy is true. I will pick outlook = rainy $\left(\frac{2}{5}\right)$

So, if (outlook = rainy) $\wedge$ (...) $\rightarrow$ Play = No

The new data set is.

outlook	Temperature	Humidity	Windy	Play
Rainy	Mild	High	False	yes
Rainy	Cool	Normal	False	yes
Rainy	Cool	Normal	True	No
Rainy	Mild	Normal	False	yes
Rainy	Mild	High	True	No

we apply the same algorithm

Temperature = Mild, $\frac{1}{3}$

Temperature = Cool, $\frac{1}{2}$

Humidity = High, $\frac{1}{2}$

Humidity = Normal, $\frac{1}{3}$

Windy = False, $\frac{0}{3}$

Windy = True, $\frac{2}{2}$

we pick windy = True, thus R_2 becomes

$\boxed{((\text{Outlook} = \text{rainy}) \wedge (\text{windy} = \text{True})) \rightarrow \text{Play} = \text{No}}$

Final results by PRISM

(outlook = overcast) $\longrightarrow$ Play = "yes"

(humidity = Normal $\wedge$ windy = False) $\rightarrow$ Play = "yes"

(outlook = sunny $\wedge$ humidity = high) $\rightarrow$ Play = "No"

((outlook = rainy) $\wedge$ (windy = True)) $\rightarrow$ Play = "No"

Exercise 3

Classify using Naïve Bayes method the data item:

$$P(\text{lenses}=\text{hard}\mid E) = P(\text{Age}=\text{pre-presbyopic}\mid \text{lenses}=\text{hard})$$
$$\times P(\text{spectacle-prescrip}=\text{hypermetrope}\mid \text{lenses}=\text{hard})$$
$$\times P(\text{astigmatism}=\text{yes}\mid \text{lenses}=\text{hard})$$
$$\times P(\text{tear-prod-rate}=\text{reduce}\mid \text{lenses}=\text{hard})$$
$$= \frac{1/4 * 2/4 * \frac{4}{4} \times \frac{0}{4} \times \frac{4}{25}}{P(E)}$$

Apply Laplace estimator and adding k the number of possible attribute values.

then
$$P(\text{lenses}=\text{hard}\mid E) = \frac{\frac{1+1}{4+3} \times \frac{2+1}{4+2} \times \frac{4+1}{4+2} \times \frac{(0+1)}{4+2} \times \frac{4+1}{24+3}}{P(E)}$$
$$= \frac{\frac{2}{7} \times \frac{1}{2} \times \frac{5}{6} \times \frac{1}{6} \times \frac{5}{27}}{P(E)}$$

$$P(\text{lenses}=\text{soft}\mid E) = \frac{\frac{2+1}{5+1} \times \frac{3+1}{5+2} \times \frac{0+1}{5+2} \times \frac{0+1}{5+2} \times \frac{5+1}{24+3}}{P(E)}$$
$$\simeq \frac{\frac{1}{2} \times \frac{2}{3} \times \frac{1}{7} \times \frac{1}{7} \times \frac{2}{9}}{P(E)}$$

$$P(\text{lenseses}=\text{None}) \simeq \frac{\frac{5+1}{15+3} \times \frac{8+1}{15+2} \times \frac{8+1}{15+2} \times \frac{12+1}{15+2} \times \frac{15+1}{24+3}}{P(E)}$$
$$= \frac{\frac{1}{3} \times \frac{9}{17} \times \frac{9}{17} \times \frac{13}{17} \times 16/27}{P(E)}$$

Thus, $P(\text{lenses}=\text{hard}|E) = \dfrac{\frac{50}{13608}}{P(E)} \approx \dfrac{0.0037}{P(E)}$

$P(\text{lenses}=\text{soft}|E) = \dfrac{\frac{4}{2646}}{P(E)} \approx \dfrac{0.0015}{P(E)}$

$P(\text{lenses}=\text{None}|E) = \dfrac{\frac{16848}{397953}}{P(E)} = \dfrac{0.042}{P(E)}$

$P(E) = 0.042 + 0.0015 + 0.0037 = 0.0472$

Therefore $P(\text{lenses}=\text{hard}|E) = \dfrac{0.0037}{0.0472} \approx 8\%$

$P(\text{lenses}=\text{soft}|E) = \dfrac{0.0015}{0.0472} \approx 3\%$

$P(\text{lenses}=\text{None}) = \dfrac{0.042}{0.0472} \approx 89\%$

The date item is classified as None since it has the highest percentage.

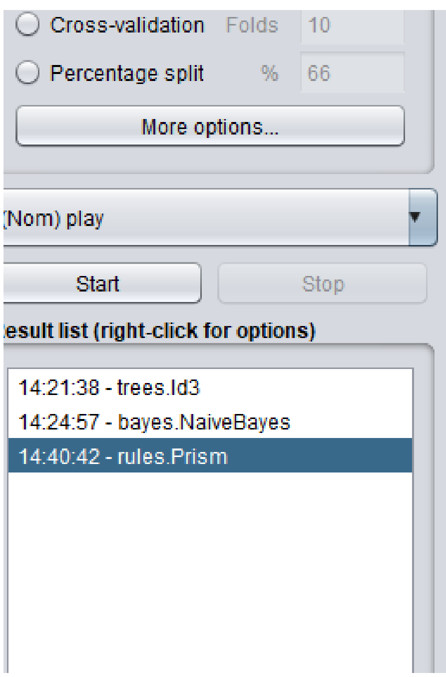
```
=== Classifier model (full training set) ===

Id3


tear-prod-rate = reduced: none
tear-prod-rate = normal
|  astigmatism = no
|  |  age = young: soft
|  |  age = pre-presbyopic: soft
|  |  age = presbyopic
|  |  |  spectacle-prescrip = myope: none
|  |  |  spectacle-prescrip = hypermetrope: soft
|  astigmatism = yes
|  |  spectacle-prescrip = myope: hard
|  |  spectacle-prescrip = hypermetrope
|  |  |  age = young: hard
|  |  |  age = pre-presbyopic: none
|  |  |  age = presbyopic: none

Time taken to build model: 0.02 seconds

=== Evaluation on training set ===
```




```
=== Classifier model (full training set) ===

Prism rules
----------
If outlook = overcast then yes
If humidity = normal
   and windy = FALSE then yes
If temperature = mild
   and humidity = normal then yes
If outlook = rainy
   and windy = FALSE then yes
If outlook = sunny
   and humidity = high then no
If outlook = rainy
   and windy = TRUE then no


Time taken to build model: 0 seconds

=== Evaluation on training set ===
```

Test data:

pre-presbyopic, hypermetrope, yes, reduced,none