Assignment 1. Solutions

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

For the root we have the following choices:

Age
- young: 2/4/2 (8)
- pre-presbyopic: 1/5/2 (8)
- presbyopic: 1/6/1 (8)

Spectacle-Prescription
- myope: 3/7/2 (12)
- hypermetrope: 1/8/3 (12)

Astigmatism
- yes: 4/8/0 (12)
- no: 7/5/0 (12)

Tear-prod-rate
- normal: 4/3/5 (12)
- reduced: 0/0/12 (12)

x/y/z means that we have x instances of some class, y instances of another class, and z instances of yet another class. The order x/y/z doesn't matter for computing entropies.

In the following I am computing the entropies using the natural logarithm *l*(.) (i.e. the one with base *e*). In order to get the values for logarithms with base 2, we have to divide with *l*(2). The reason I am not using directly logarithms with base 2 is that the calculators don't provide us log with base 2.

Age entropies:
entropy(2/4/2) = (-(2/8)*l(2/8)-(4/8)*l(4/8)-(2/8)*l(2/8))/l(2) = 1.5
entropy(1/5/2) = (-(1/8)*l(1/8)-(5/8)*l(5/8)-(2/8)*l(2/8))/l(2) = 1.3
entropy(1/6/1) = (-(1/8)*l(1/8)-(6/8)*l(6/8)-(1/8)*l(1/8))/l(2) = 1.06
avg_entropy = (8/24)*1.5 + (8/24)*1.3 + (8/24)*1.06 = 1.287 bits

Spectacle-Prescription entropies:
entropy(3/7/2) = (-(3/12)*l(3/12)-(7/12)*l(7/12)-(2/12)*l(2/12))/l(2) = 1.384
entropy(1/8/3) = (-(1/12)*l(1/12)-(8/12)*l(8/12)-(3/12)*l(3/12))/l(2) = 1.1887
avg_entropy = (12/24)* 1.384 + (12/24)* 1.1887 = 1.28635 bits

Astigmatism entropies:
entropy(4/8/0) = (-(4/12)*l(4/12)-(8/12)*l(8/12)-0)/l(2) = .918

entropy(7/5/0) = (-(7/12)*l(7/12)-(5/12)*l(5/12)-0)/l(2) = .9799
avg_entropy = (12/24)* .918 + (12/24)* .9799 = .94895 bits
Tear-prod-rate entropies:
entropy(4/3/5) = (-(4/12)*l(4/12)-(3/12)*l(3/12)-(5/12)*l(5/12))/l(2) = 1.555
entropy(0/0/12) = (0-0-0)/l(2) = 0
avg_entropy = (12/24)* 1.555 + (12/24)*0 = **.7775 bits**

The smallest average entropy is for Tear-prod-rate, so we choose it for the root.
Now we have two branches (Tear-prod-rate=reduced) and (Tear-prod-rate=normal)
The first branch ends up to be a leaf because all the instances going that branch are "contact-lenses = none".

For the other branch (Tear-prod-rate=normal) we have the following data instances:

age	spectacle-prescrip	astigmatism	tear-prod-rate	contact-lenses
pre-presbyopic	myope	yes	normal	hard
presbyopic	myope	yes	normal	hard
young	hypermetrope	yes	normal	hard
young	myope	yes	normal	hard
pre-presbyopic	hypermetrope	yes	normal	none
presbyopic	hypermetrope	yes	normal	none
presbyopic	myope	no	normal	none
pre-presbyopic	hypermetrope	no	normal	soft
pre-presbyopic	myope	no	normal	soft
presbyopic	hypermetrope	no	normal	soft
young	hypermetrope	no	normal	soft
young	myope	no	normal	soft

We have the following choices to break further:

Age
- young: 2/2/0 (4)
- pre-presbyopic: 1/1/2 (4)
- presbyopic: 1/2/1 (4)

Spectacle-Prescription
- myope: 1/2/3 (6)
- hypermetrope: 3/1/2 (6)

Astigmatism
- yes: 4/2/0 (6)
- no: 1/5/0 (6)

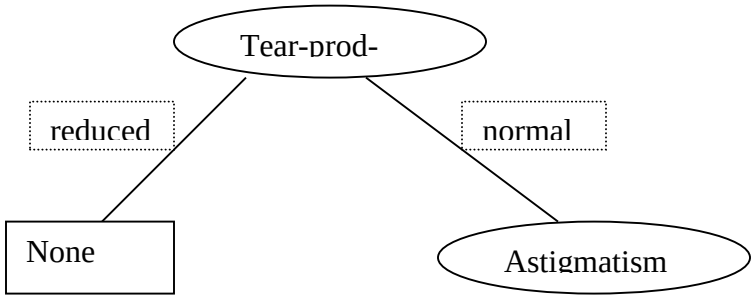
Age entropies:
entropy(2/2/0) = (-(2/4)*l(2/4)-(2/4)*l(2/4)-0)/l(2) = .999

entropy(1/1/2) = (-(1/4)*l(1/4)-(1/4)*l(1/4)-(2/4)*l(2/4))/l(2) = 1.5
entropy(1/2/1) = (-(1/4)*l(1/4)-(2/4)*l(2/4)-0)/l(2) = .999
avg_entropy = (4/12)* .999 + (4/12)* 1.5 + (4/12)* .999 = 1.166 bits

Spectacle-Prescription entropies:
entropy(1/2/3) = (-(1/6)*l(1/6)-(2/6)*l(2/6)-(3/6)*l(3/6))/l(2) = 1.459
entropy(3/1/2) = (-(3/6)*l(3/6)-(1/6)*l(1/6)-(2/6)*l(2/6))/l(2) = 1.459
avg_entropy = (6/12)* 1.459 + (6/12)* 1.459 = 1.459 bits

Astigmatism entropies:
entropy(4/2/0) = (-(4/6)*l(4/6)-(2/6)*l(2/6)-0)/l(2) = .918
entropy(1/5/0) = (-(1/6)*l(1/6)-(5/6)*l(5/6)-0)/l(2) = .65
avg_entropy = (12/24)* .918 + (12/24)*.65 = **.784** bits

So, we choose astigmatism for the next level. The tree so far is:



2. Construct two rules using PRISM for the weather data. Show the details of your construction. Then, check your solution with Weka.

Rule we seek: **If ? then play=yes**
(Of course, we could have alternatively started with **If ? then play=no**. Actually, according to the heuristic mentioned in the PRISM pseudocode, we should start with the latter. However, in order to be compatible with WEKA, I am starting with the former. WEKA unfortunately ignores the heuristic. Both solutions will be considered fine when marking this question.)

Outlook=sunny	2/5
Outlook=overcast	4/4
Outlook=rainy	3/5
Temp=cool	3/4
Temp=mild	4/6

Temp=hot	2/4
Humidity=normal	6/7
Humidity=high	3/7
Windy=true	3/6
Windy=false	6/8

Best accuracy Outlook=overcast 4/4, so we choose it as a condition:

If Outlook=overcast then play=yes

We consider the instances covered by this rule, i.e. those that make the antecedent true. We observe that all them have play=yes, so we are done with the first rule.

Now, let's consider the instances not covered by this rule.

Outlook	Temperature	Humidity	Windy	Play
rainy	mild	high	TRUE	no
rainy	cool	normal	TRUE	no
rainy	mild	high	FALSE	yes
rainy	cool	normal	FALSE	yes
rainy	mild	normal	FALSE	yes
sunny	hot	high	FALSE	no
sunny	mild	high	FALSE	No
sunny	hot	high	TRUE	No
sunny	cool	normal	FALSE	yes
sunny	mild	normal	TRUE	yes

Rule we seek: **If ? then play=yes**

Outlook=sunny	2/5
Outlook=rainy	3/5
Temp=cool	2/3
Temp=mild	3/5
Temp=hot	0/2
Humidity=normal	4/5
Humidity=high	1/5
Windy=true	1/4
Windy=false	4/6

The best to choose is Humidity=normal 4/5, so we get

If Humidity=normal then play=yes

The covered instances (i.e. those that satisfy the antecedent) are:

outlook	temperature	humidity	windy	play
rainy	cool	normal	TRUE	no
rainy	cool	normal	FALSE	yes
sunny	cool	normal	FALSE	yes
rainy	mild	normal	FALSE	yes
sunny	mild	normal	TRUE	yes

It's a pretty good rule and we might want to stop here, but let's say we want full accuracy.

We want to extend the rule as: **If Humidity=normal and ? then play=yes**

Outlook=sunny	2/2
Outlook=rainy	2/3
Temp=cool	2/3
Temp=mild	2/2
Temp=hot	undefined
Windy=true	1/2
Windy=false	3/3

Between Temp=mild and Windy=false we choose Windy=false because it has greater coverage. So our second rule becomes:

If Humidity=normal and Windy=false then play=yes

We consider the instances covered by this rule, i.e. those that make the antecedent true. We observe that all them have play=yes, so we are done with the second rule.

3. (4 points) Classify using Naïve Bayes method (on contact lenses data) the data item: *pre-presbyopic, hypermetrope, yes, reduced, ?*

P(Hard|E) = (1+1)/(4+3) * (1+1)/(4+2) * (4+1)/(4+2) * (0+1)/(4+2) * (4+1)/(24+3)
= alpha * .00244953948657652361

P(Soft|E) = (2+1)/(5+3) * (3+1)/(5+2) * (0+1)/(5+2) * (0+1)/(5+2) * (5+1)/(24+3)
= alpha * .00097181729834791059

P(None|E) = (5+1)/(15+3) * (8+1)/(15+2) * (8+1)/(15+2) * (12+1)/(15+2) * (15+1)/(24+3)
= alpha * .04233665784652961530

alpha = 1/(.00244953948657652361 + .00097181729834791059 + .04233665784652961530) = 21.85409502694201713124

P(Hard|E) = .00244953948657652361 * 21.85409502694201713124
= .05353246871189010655
~ 0.054

P(Soft|E) = .00097181729834791059 * 21.85409502694201713124
= .02123818758692129938
~ 0.021

P(None|E) = .04233665784652961530 * 21.85409502694201713124
= .92522934370118859406
~ 0.925