SEng 474 / CSc 578D

Data Mining – Fall 2016

Assignment 1 - Solutions

1.a) Construct the root and the first level of a decision tree for the contact lenses data using the ID3 algorithm.

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)

Astigmatismyes: 4/8/0 (12)

no: 7/5/0 (12)

Tear-prod-ratenormal: 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. **NOTE:** Entropy is calculated with log base 2 ($\log_2$).

Age entropies:entropy(2/4/2) = (-(2/8)*$\log_2$(2/8)-(4/8)* $\log_2$(4/8)-(2/8)* $\log_2$(2/8)) = 1.5 entropy(1/5/2) = (-(1/8)* $\log_2$(1/8)-(5/8)* $\log_2$(5/8)-(2/8)* $\log_2$(2/8))= 1.299 entropy(1/6/1) = (-(1/8)* $\log_2$(1/8)-(6/8)* $\log_2$(6/8)-(1/8)* $\log_2$(1/8))= 1.061 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)* $\log_2$(3/12)-(7/12)* $\log_2$(7/12)-(2/12)* $\log_2$(2/12))= 1.384 entropy(1/8/3) = (-(1/12)* $\log_2$(1/12)-(8/12)* $\log_2$(8/12)-(3/12)*

$\log_2(3/12)$)= 1.1887 avg_entropy = (12/24)* 1.384 + (12/24)* 1.1887 = 1.28635 bits

Astigmatism entropies:entropy(4/8/0) = (-(4/12)* $\log_2(4/12)$-(8/12)* $\log_2(8/12)$-0)= .918 entropy(7/5/0) = (-(7/12)* $\log_2(7/12)$-(5/12)* $\log_2(5/12)$-0)= .9799 avg_entropy = (12/24)* .918 + (12/24)* .9799 = .94895 bits

Tear-prod-rate entropies:entropy(4/3/5) = (-(4/12)* $\log_2(4/12)$-(3/12)* $\log_2(3/12)$-(5/12)* $\log_2(5/12)$)= 1.555 entropy(0/0/12) = (0-0-0)= 0avg_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 is actually a leaf because all of the instances going to that branch are "contact-lenses = none". For the second branch (Tear-prod-rate=normal) we have the following data instances:

age

pre-presbyopic presbyopic youngyoung pre-presbyopic presbyopic presbyopic pre-presbyopic pre-presbyopic presbyopic young

young

spectacle-prescripmyope yes myope yes hypermetrope yes myope yes hypermetrope yes hypermetrope yes myope no hypermetrope no myope no hypermetrope no hypermetrope no myope no

We have the following choices to split 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)

Astigmatismyes: 4/2/0 (6)

no: 1/5/0 (6)

Age entropies:entropy(2/2/0) = (-(2/4)* $\log_2$(2/4)-(2/4)* $\log_2$(2/4)-0)= .999 entropy(1/1/2) = (-(1/4)* $\log_2$(1/4)-(1/4)* $\log_2$(1/4)-(2/4)* $\log_2$(2/4))= 1.5 entropy(1/2/1) = (-(1/4)* $\log_2$(1/4)-(2/4)* $\log_2$(2/4)-(1/4)* $\log_2$(1/4))= 1.5 avg_entropy = (4/12)* .999 + (4/12)* 1.5 + (4/12)* .1.5 = 1.333 bits

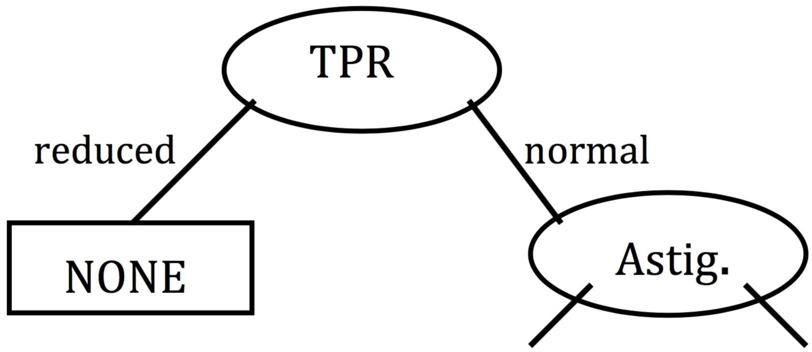
astigmatism tear-prod-rate contact-lenses

normal hard normal hard normal hard normal hard normal none normal none normal none normal soft normal soft normal soft normal soft normal soft

Spectacle-Prescription entropies:entropy(1/2/3) = (-(1/6)* $\log_2$(1/6)-(2/6)* $\log_2$(2/6)-(3/6)* $\log_2$(3/6))= 1.459 entropy(3/1/2) = (-(3/6)* $\log_2$(3/6)-(1/6)* $\log_2$(1/6)-(2/6)* $\log_2$(2/6))= 1.459 avg_entropy = (6/12)* 1.459 + (6/12)* 1.459 = 1.459 bits

Astigmatism entropies:entropy(4/2/0) = (-(4/6)* $\log_2$(4/6)-(2/6)*log2(2/6)-0)= .918 entropy(1/5/0) = (-(1/6)* $\log_2$(1/6)-(5/6)* $\log_2$(5/6)-0)= .65 avg_entropy = (12/24)* .918 + (12/24)*.65 = .784 bits

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



b) There were multiple things we were hoping to see. Here is a list of the most important:

1) sklearn library does not use ID3 to build its decision tree but an algorithm named CART (classification and regression tree).

2) CART builts binary trees, which imply that classes and attributes having more than two possible values will be evaluated under a series of combinations. For our data, it evaluates “none” vs. “soft+hard”, “soft” vs. “none+hard” and “hard” vs. ”none+soft”; and similarly on attributes who have more than two values.For instance the root calculation is:entropy(20/4) = (-(20/24)* $\log_2$(20/24)-(4/24)*$\log_2$(4/24))= 0.65 entropy(9/15) = (-(9/24)* $\log_2$(9/24)-(15/24)*$\log_2$(15/24))= 0.9544 entropy(19/5) = (-(19/24)* $\log_2$(19/24)-(5/24)*$\log_2$(5/24))= 0.7383 avg_entropy = (0.65 + 0.9544 + 0.7383)/3 = 0.7809 bits

3) The way our data was encoded transformed our attributes into a larger set of binary attributes, and entropy was calculated on each.

2.

Calculate the probabilities needed for Naïve Bayes using the contact lenses dataset. Classify: "*prepresbyopic, hypermetrope, yes, reduced, ?*" using your calculated probabilities.

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 or 5.4%

P(Soft|E) = .00097181729834791059 * 21.85409502694201713124 = .02123818758692129938 ~ 0.021 or 2.1%

P(None|E) = .04233665784652961530 * 21.85409502694201713124 = .92522934370118859406 ~ 0.925 or 92.5%

Classified as "None".

NOTE: smoothing needs to be applied to all classes for consistency. Otherwise the P(E) or alpha (1/P(E)) becomes biased.