Assignment #5

Jason Maynard
U30503758

Problem #1)

a) Ockham's razor: A consistent hypothesis agrees with all the data. How do we choose from the multiple options for consistent data?

* Choose the simplest option.

Main advantages: Choosing a simple optim makes selecting an optim much easier. This relieves the computational complexity of learning.

"We will want to use the hypothesis (h) after we have learned it." and Computing h(x) when h is a linear function is guaranteed to be fast, while computing an arbitrary Turing machine program is not even guaranteed to terminate. For these reasons, most work on learning has focused on simple representations." — pg. 697 Text

To summarize, Choosing a more simple representation makes initial computation and future use of the function more easy.

Main disadvantages: Any time we make simplifying assumptions we loose the ability to make an exact fit and model the data perfectly. Simplifying the function will always result in a loss of accuracy. This is like sampling digital music to a CD or compressing a photo. There will always be some type of loss when taking an analog signal and making it "digital". There will be loss in the simplification.

b) Main limitations of decision tree learning:

A decision tree is still a model or simplification of the hypothesis space. Because of this simplification, there will be loss of granularity. Some decision trees can not represent the full function without being exponentially large. Some functions can be represented by decision trees while others can not due to the size. The hypothesis space is just too large in many cases.

1.) Choosing the appropriate attribute selection measure
2.) How deep to grow the decision tree
3.) Handling continuous attributes
4.) Handling training data w/ missing attributes
5.) Handling attributes w/ different costs
6.) Improving computational efficiency.

Jason Maynard
U30503758

Bind	Style	Color	Well Known	Length	Classification
Hard	Novel	NO	Popular	Long	P
Soft	Text	NO	Popular	Long	P
Soft	Novel	NO	Popular	Short	n
Hard	Text	Color	Popular	Short	P
Hard	Photo	Color	unknown	short	P
Soft	Text	NO	unknown	Short	n
Hard	Photo	color	Popular	Long	P
Soft	Novel	color	unknown	Short	n

8

Step 1: Determine high level entropy

$P = 5$ $n = 3$ $H(c) = -\left(\frac{5}{8}\right)\log_2\left(\frac{5}{8}\right) - \left(\frac{3}{8}\right)\log_2($

$H(c) = .95$

Step 2: Calculate info gain for different splits. (5 different splits)

1) Attribute (Bind)

Bind

Hard $(4/8)$

Novel	NO	popular	long	P
text	Color	pop	Short	P
photo	Color	unknown	Short	n
photo	Color	pop	Long	P

$H_{bind}(hard) = 0$ # pure split

Soft $(4/8)$

text	NO	pop	Long	P
Novel	NO	pop	Short	n
text	NO	unk	Short	n
Novel	Color	unk	Short	n

$H_{bind}(soft) = -.25\log_2(.25) - .75\log_2(.75)$

$= .81$

$Gain_{bind} = .95 - .5(0) - .5(.81)$

$= .545$

2) Attribute (style)

Style

Novel $(3/8)$

Position	Hardcover	Nocolor	popular	Long
Neg	Soft	Nocolor	popular	Short
Neg	Soft	Color	unknown	Short

$p = .33$ $n = .66$ $H_{Style}(Novel) = .92$

text $(3/8)$

P	Soft	Nocolor	popular	Long
P	Hard	Color	popular	Short
n	Soft	Nocolor	unknown	Short

$p = .66$ $n = .33$ $H_{Style}(text) = .92$

Photo journal $(2/8)$

P	Hard	Color	unknown	Short
P	Hard	Color	popular	Long

$p = 1.0$ $n = 0.0$ $H_{Style}(photo) = 0$

$Gain_{Style} = .95 - \left(\frac{3}{8}\right).92 - \left(\frac{3}{8}\right).92 - \frac{2}{8}(0)$

$= .216$

3) Attribute (color)

Color

No Color (4/8)

Hard	Novel	popular	long	p
Soft	text	popular	long	p
Soft	Novle	popular	short	n
Soft	text	unknown	short	n

p = .5
n = .5

$H_{Color}(No) = 1.0$

color (4/8)

Hard	text	popular	short	p
Hard	photo	unknown	short	p
Hard	photo	popular	long	p
Soft	Novel	unknown	short	n

p = .75
n = .25

$H_{Color}(Color) = .81$

$Gain_{Color} = .95 - .5(1) - .5(.8$
$= .05$

4) Attribute (well known)

well known?

popular (5/8)

Hard	Novel	NoColor	long	p
Soft	text	NoColor	long	p
Soft	Novel	NoColor	short	n
hard	text	color	short	p
hard	photo	color	long	p

p = 4/5 = .8
n = 1/5 = .2

$H_{wellknown}(popular) = .71$

unknown (3/8)

Hard	photo	Color	short	p
Soft	text	NoColor	short	n
Soft	novel	color	short	n

p = .33
n = .64

$H_{wellknown}(unknon) = .92$

$Gain_{wellknow} = .95 - (5/8).71 - (3/8).92$
$= .16$

5) Attribute (length)

length

Long (3/8)

hard	Novel	noColor	popular	p
Soft	text	noColor	Popular	p
hard	photo	color	Popular	p

$H_{length}(long) = 0$

short (5/8)

Soft	novel	noColor	popular	n
hard	text	Color	popular	p
hard	photo	Color	unknown	p
Soft	text	noColor	unknown	n
Soft	novel	Color	unknown	n

p = .4

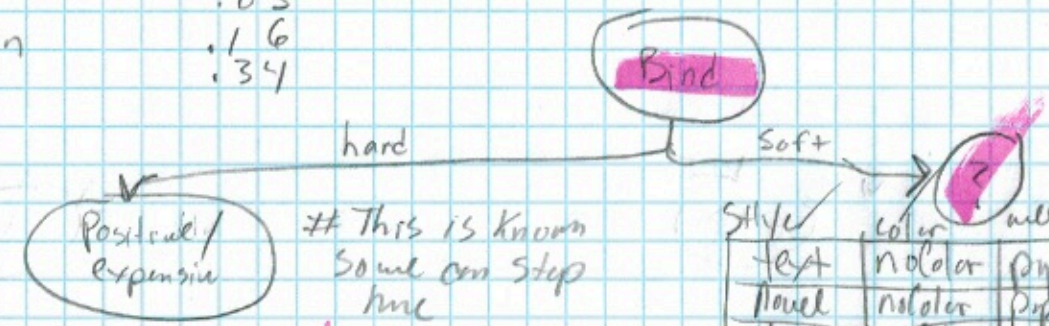
$H_{length}(short) = .97$

$Gain_{length} = .95 - (3/8)0 - (5/8).97$
$= .34$

Step 3: Break down the next level of splits given the first level split

Split	infoGained
Bind	.55
Style	.26
Color	.05
Well Known	.16
length	.34

This is a recursive step.
bind is best split



this is unknown so we need to recurse on the problem

Style	color	wellKnown	length	
text	noColor	popular	long	p
novel	noColor	popular	short	n
text	noColor	unknown	short	n
novel	Color	unknown	short	n

p = .25
n = .75

H_{bind} (soft) = .81

novel	noColor	popular	long	p
text	Color	popular	short	p
photo	Color	unknown	short	p
photo	Color	popular	long	p

H_{bind} (soft) is the starting entropy for the next level of calculations, Repeat step2 to calculate gain for different splits.

attribute(1,1) (soft)
bind

Style

text (2/4)

Color pretties?	Well Known?	length of book?	
noColor	popular	long	p
noColor	unknown	short	n

E_{style} (text) = 1.0

novel (2/4)

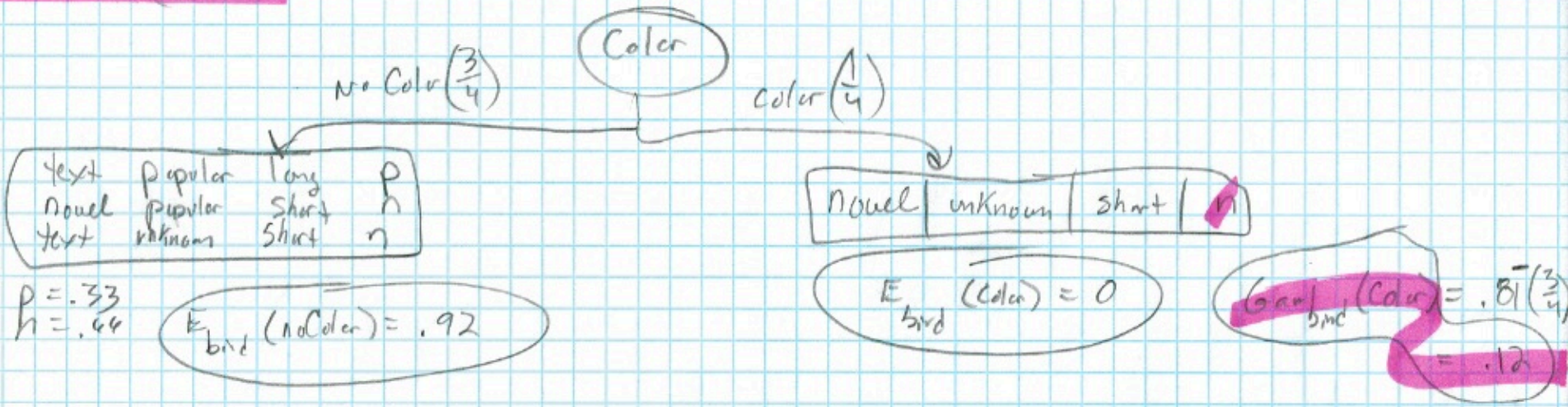
noColor	popular	short	n
Color	unknown	short	n

E_{style} (novel) = 0

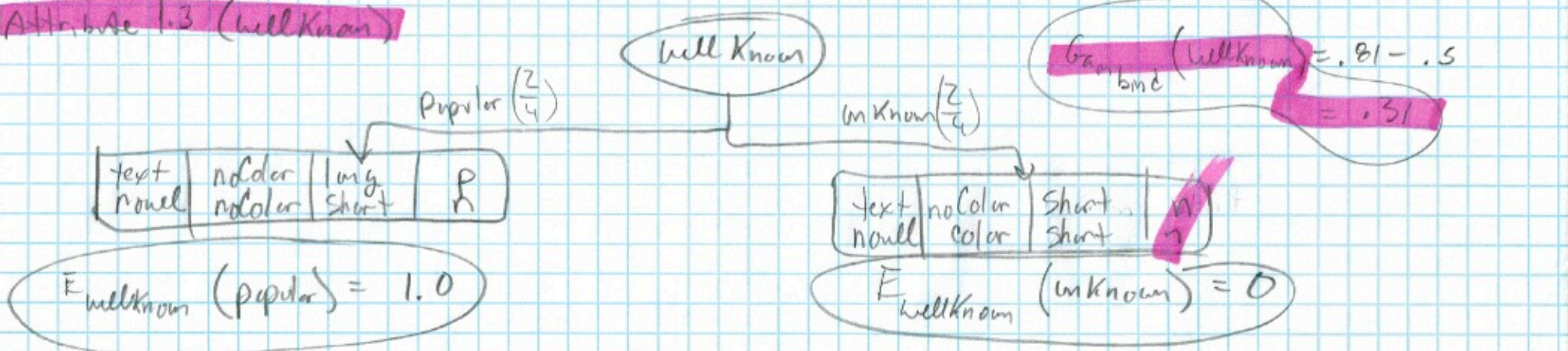
gain = .81 − .5(1)
= .31

$gain_{bind}$ (Style) = .31

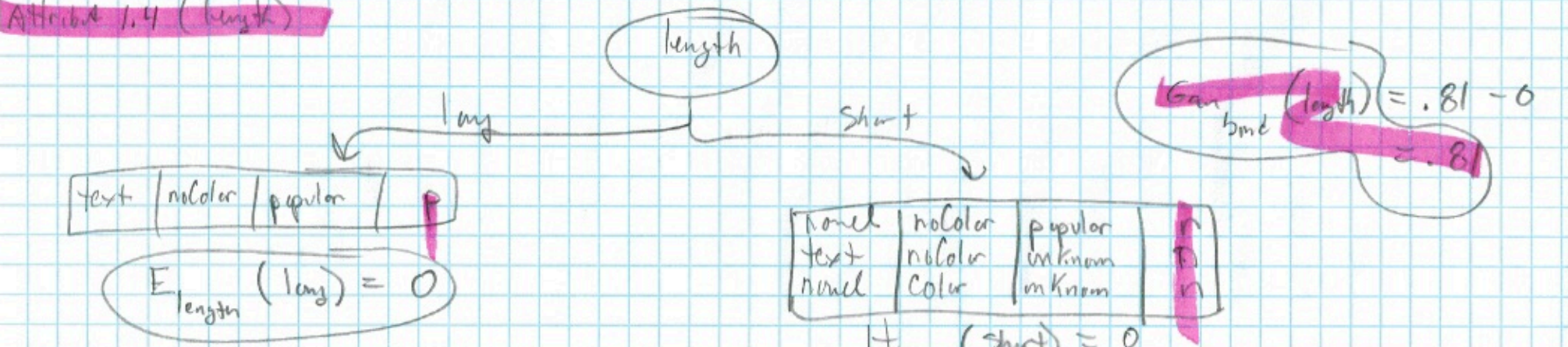
Attribute 1.2 (Color)



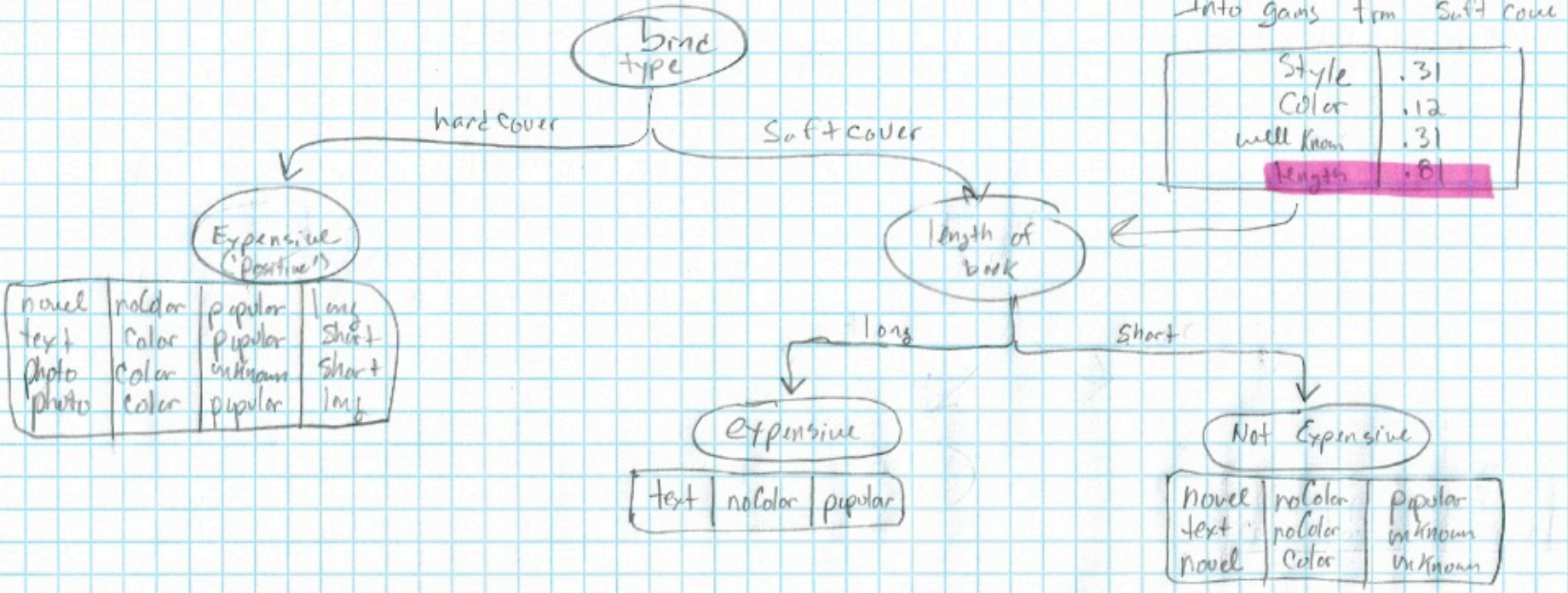
Attribute 1.3 (wellKnown)



Attribute 1.4 (length)



length is the next most important attribute and produces a pure set.



Artificial Intelligence

Name and ID _____

Assignment 5

Due date: November 30 at 11:59pm

Problem 1 (4 points)

Read Chapter 18 and answer the following questions:

a. What are the main advantages and drawbacks of using Ockham's razor in learning
b. What are the main limitation of the decision-tree learning?

Problem 2 (6 points)

This file contains eight training data and three test data, and the attributes refer to whether a book will be expensive at the local bookstore.

Classification	Bind type	Style of book	Color pictures?	Is the book well known?	Length of book
Positive	Hardcover	Novel	Nocolor	Popular	Long
Positive	Softcover	Textbook	Nocolor	Popular	Long
Negative	Softcover	Novel	Nocolor	Popular	Short
Positive	Hardcover	Textbook	Color	Popular	Short
Positive	Hardcover	Photojournal	Color	Unknown	Short
Negative	Softcover	Textbook	Nocolor	Unknown	Short
Positive	Hardcover	Photojournal	Color	Popular	Long
Negative	Softcover	Novel	Color	Unknown	Short

Use these data to construct a decision tree; you should compute the information gains to decide which attributes are more important. For each node of the tree, indicate the corresponding information gain.

Problem 3 (10 points)

Implement a program for building decision trees. It should read a file with training and test examples, use the training examples to build a tree, and then classify the text examples. The only required output is the classification of the test examples; it does not have to include the tree itself (if you output also the tree, you will have 5 points bonus). The input format is as follows: