Abelardo Riojas
ISC 4221C Prof. Quaife
Fall 2019

Assignment 3

1. Consider the set of training data shown in the table below for a binary classification problem (here + or -). Note that for each record there are three attributes, two of which are binary and one which is continuous.

Record	1	2	3	4	5	6	7	8	9
A	T	T	T	F	F	F	F	T	F
B	T	T	F	F	T	T	F	F	T
C	1.0	6.0	5.0	4.0	7.0	3.0	8.0	7.0	5.0
Class	+	+	-	+	-	-	-	+	-

a. Determine the entropy of this collection of training examples

4/9 are of class plus, 5/9 are of class minus .

Entropy(t) = $\sum_{i=1}^{k} p(i|t)log_2 p(i|t)$

= $-((4/9)\log_2(4/9) + (5/9)\log_2(5/9)) = 0.99107606$

b. What the information gains (based on entropy measure) for attributes A and B? Which one provides the largest information gain?

A	+	-
T	3	1
F	1	4

Entropy(A|T) = -(((3/4)log2(3/4) + (1/4)log2(1/4)) = 0.811278124

Entropy(A|F) = -(((1/5)log2(1/5) + (4/5)log2(4/5)) = 0.721928095

Δ = 0.99107606 - ((4/9)*0.811278124 + (5/9)* 0.721928095) = 0.229436841

B	+	-
T	2	3
F	2	2

Entropy(B|T) = -((2/5)log2(2/5) + (3/5)log2(3/5)) = 0.970950594

Entropy(B|F) = -((2/4)log2(2/4) + (2/4)log2(2/4)) = 1

Δ = 0.99107606 - ((5/9)*0.970950594 + (4/9)* 1) = 0.00721461889

Class A gives the largest information gain between classes A and B.

c.

C	+	-
≥ 3	1	1
btw	2	2
>6	1	2

Entropy(C|$\geq$3) = -((1/2)log2(1/2) + (1/2)log2(1/2)) = 1

Entropy(C|btw) = -((2/4)log2(2/4) + (2/4)log2(2/4)) = 1

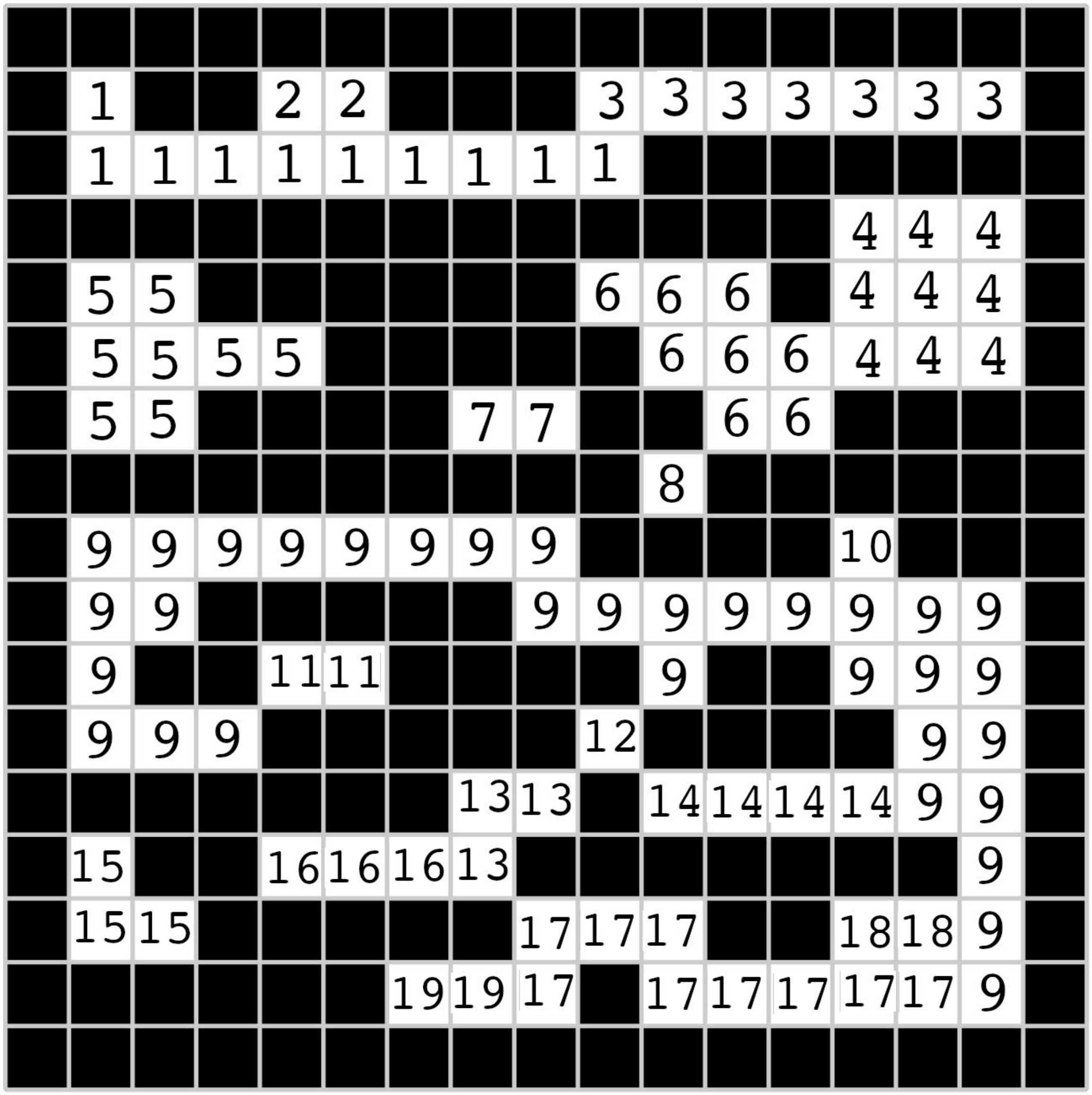
Entropy(C|>6) = -((1/3)log2(1/3) + (2/3)log2(2/3)) = 0.918295834

Δ = 0.99107606 - ((2/9)*1 + (4/9)* 1 + (3/9)*0.918295834) = 0.018310782

d. Determine a decision tree based upon choosing the largest information gain using the entropy measure and splitting attribute C as in part (c). Justify your choices

Class A has the highest gain. Consequently, we choose A as the choice of the first attribute to test. The next choice is choice B as the "true" leaf for B is homogenous for both the True A and False A case. If Class B returns false, then class C will determine the class of the result.

2. For this problem, use the two-dimensional image in Figure 1. Your goal is to label the white pixels in terms of connected regions.
   a. Recall the labelling algorithm we introduced in class. For each white pixel, we considered the neighboring pixels to the north and west.



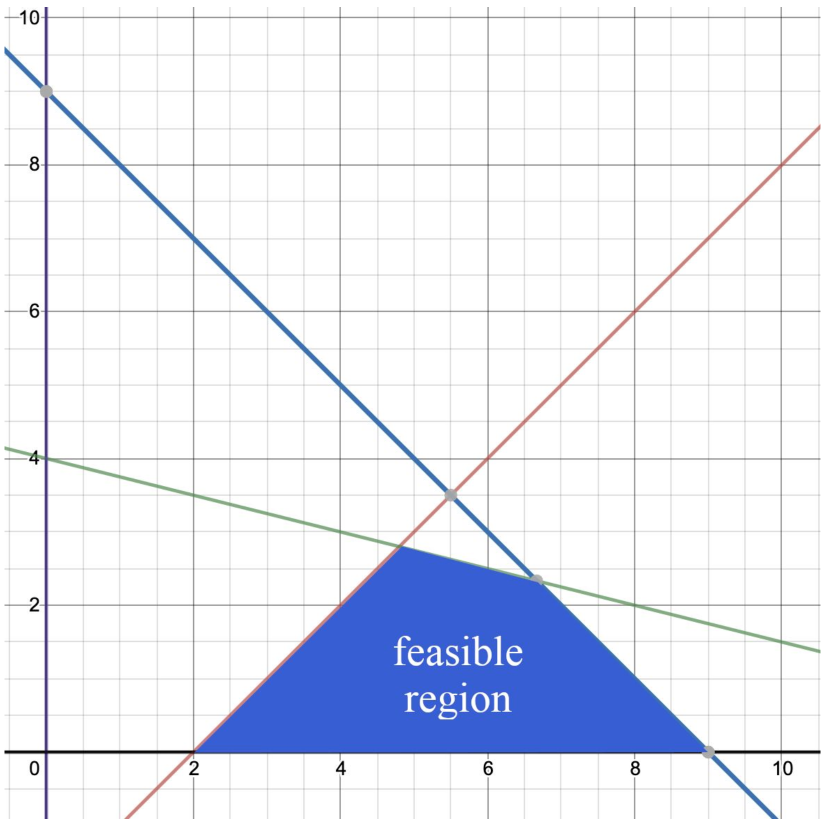
   b. Apply label readjustment by creating a table similar to the one we formed in class. Column 1: Label before readjustment Column 2: Label after readjustment

1	1	Stays
2	1	2 → 1
3	3	Stays

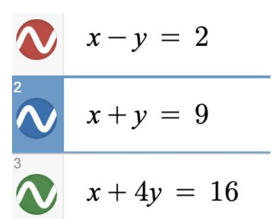
4	4	Stays
5	5	Stays
6	4	6 → 4
7	7	Stays
8	8	Stays
9	9	Stays
10	9	10 → 9
11	11	Stays
12	12	Stays
13	13	Stays
14	9	14 → 9
15	15	Stays
16	13	16 → 13
17	9	17 → 9
18	9	18 → 9
19	9	19 → 17 → 9

3. Maximize the linear function $z = 5x_1 + x_2$ with the constraints

$$\begin{aligned} x_1 - x_2 &\geq 2, \\ x_1 + x_2 &\leq 9, \\ x_1 + 4x_2 &\leq 16, \\ x_1, x_2 &\geq 0. \end{aligned}$$



with



Veriticies = (2.0,0.0), (4.8, 2.8), (6.667, 2.333), (9.0, 0)

Z values = 10, 26.8, 35.66667, 45

Maximum occurs at (9,0). Therefore $x_1 = 9$ and $x_2 = 0$ for max(z).