CARLSON
SCHOOL OF MANAGEMENT
UNIVERSITY OF MINNESOTA

PREDICTIVE ANALYTICS
MSBA 6420

Homework #1

Danny Moncada
(put your name above)

Total grade: ______ out of ___100___ points

There are 5 numbered questions. Please answer them all and submit your assignment as a single PDF file by uploading it to the HW1 submission on the course website. (Use the data mining software wherever it is helpful.)

1) (12 points) Choose the data technology (Q, U, or S) that is most appropriate for each of the following business questions/scenarios.

Q – SQL Querying
U – Unsupervised Learning
S – Supervised Learning

a) Q I want to know which of my customers are the most profitable. (historical data)

b) Q I need to get data on all my on-line customers who were exposed to the special offer, including their registration data, their past purchases, and whether or not they purchased in the 15 days following the exposure. (historical data)

d) U I would like to segment my customers into groups based on their demographics and prior purchase activity. I am not focusing on improving a particular task, but would like to generate ideas. (segmentation)

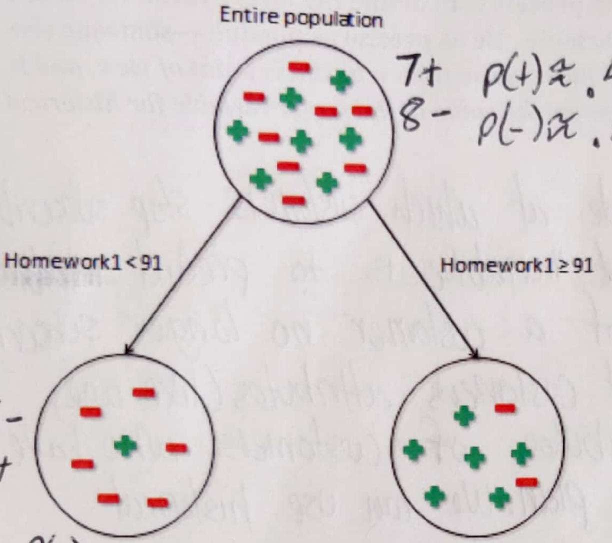
e) S I have a budget to target 10,000 existing customers with a special offer. I would like to identify those customers most likely to respond to the special offer. (classification)

0 1

f) U I want to know what characteristics differentiate my profitable customers with unprofitable ones. (segmentation)

g) S When the donor will back to the platform to donate again? (classification/ outcome)

2) **(14 points) The following figures show a simple classification problem. In particular, there are two types of students: (+) students who attend office hours and (–) students who do not attend office hours; and we have recorded the performance of the students in two assignments (i.e. Homework1 and Homework2) in order to investigate whether the performance of students on different assignments provides information about our target variable. How would you select to partition the students in an informative way? Show the detailed computations.**

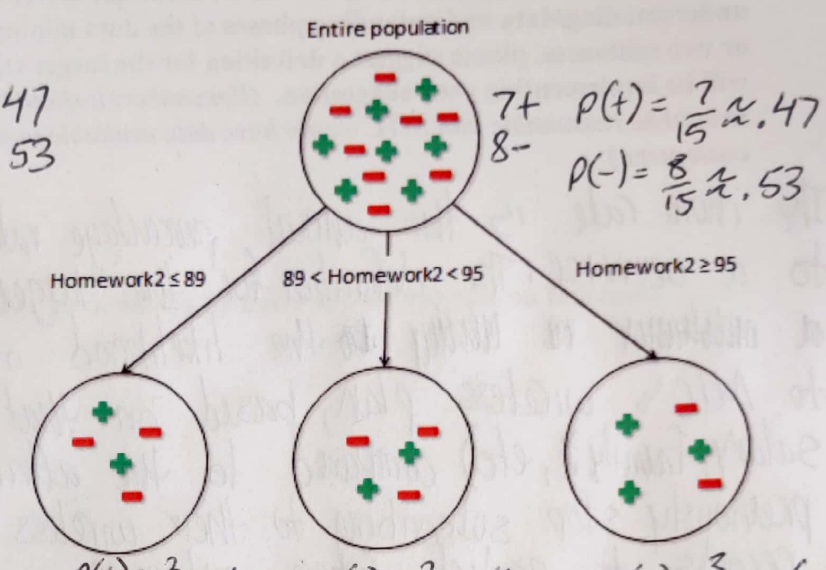


7+ $p(+) \approx .47$
8- $p(-) \approx .53$

6-
1+

$p(-) \approx \frac{6}{7} \approx .85$
$p(+) = \frac{1}{7} \approx .15$

$p(+) = \frac{6}{8} \approx .75$
$p(-) = \frac{2}{8} \approx .25$



7+ $p(+) = \frac{7}{15} \approx .47$
8- $p(-) = \frac{8}{15} \approx .53$

$p(+) = \frac{2}{5} = .4$
$p(-) = \frac{3}{5} = .6$

$p(+) = \frac{2}{5} = .4$
$p(-) = \frac{3}{5} = .6$

$p(+) = \frac{3}{5} = .6$
$p(-) = \frac{2}{5} = .4$

HW#1

Entropy(parent) = $-[0.47 \cdot \log_2(0.47) + 0.53 \cdot \log_2(0.53)]$

Entropy(parent) = $-[0.47 \cdot -1.1 + 0.53 \cdot -0.9]$

Entropy(parent) ≈ 0.99 (impure)

Entropy(left child) $\approx -[0.85 \cdot \log_2(0.85) + 0.15 \cdot \log_2(0.15)]$

Entropy(left child$_{<91}$) ≈ 0.609

Entropy(right child) $\approx -[0.75 \cdot \log_2(0.75) + 0.25 \cdot \log_2(0.25)]$

Entropy(right child$_{\geq 91}$) ≈ 0.811

Information Gain HW#1 $= .99 - [p(<91) \cdot .609 + p(\geq 91) \cdot .811]$

IG $= .99 - [0.467 \cdot .609 + .533 \cdot .811]$

$\boxed{\text{IG} \approx .273}$ HW#1

The goal of a decision tree is to maximize information gain. In this case, I would partition student using HW#1.

HW #2

Entropy(parent) = $-[0.47 \cdot \log_2(0.47) + 0.53 \cdot \log_2(0.53)]$
$\approx .99$ (impure)

Entropy(left child) = $-[.4 \cdot \log_2(.4) + .6 \cdot \log_2(.6)]$
$\approx .97$

Entropy(middle child) $\approx .97$
Entropy(right child) $\approx .97$
these are the same splits as the left child.

Information gain HW#2 $= .99 - [p(\leq 89) \cdot .97 + p(89 < x < 95) \cdot .97 + p(>95) \cdot .97]$

HW#2 IG $= .99 - [.333 \cdot .97]^3$

$\boxed{\text{IG} = 0.0297}$ #HW#2

worse information gain than HW#1

3) **(12 points) MTC (MegaTelCo) has decided to use supervised learning to address its problem of churn in its wireless phone business. As a consultant to MTC, you realize that a main task in the business understanding/data understanding phases of the data mining process is to define the target variable. In one or two sentences, please suggest a definition for the target variable. Be as precise as possible—someone else will be implementing your suggestion. *(Remember: it should make sense from a business point of view, and it should be reasonable that MTC would have data available to know the value of the target variable for historical customers.)***

The churn rate is the annual percentage rate at which customers stop subscribing to a service. The definition for the target variable is to predict ~~whether~~ ~~a customer is likely to~~ the likelihood of a customer no longer subscribing to MTC's wireless plan, based on that customers attributes (like age, salary, family #, etc) compared to the attributes of customers who have previously stop subscribing to their wireless plan. We can use historical records to predict future outcomes.

4) **(12 points) A predictive model has been applied to a test dataset and has classified 87 records as fraudulent (31 correctly so) and 953 as non-fraudulent (919 correctly so).**

- **Present the confusion matrix for this scenario.**

- **Calculate the error rate and accuracy rate.**