AACS2383 Extra Questions

Q1. Discuss whether or not each of the following activities is a data mining task.

(a) Dividing the customers of a company according to their gender. (SheeYeap)

Yes, because it makes classification on data

No, it is a simple database query.

(b) Dividing the customers of a company according to their profitability. (Tan Kang Hong)

No, this is an accounting calculation, followed by the application of a threshold.

(c) Computing the total sales of a company.(Ng Eason)

No,because the total sales of a company are able to be calculated by simple calculation.

(d) Sorting a student database based on student identification numbers.(li yuet)

No, because this is only a simple database query.

(e) Predicting the outcomes of tossing a (fair) pair of dice. (Chun Wai)

No, because the outcomes of tossing a fair pair of dice can be calculated by probability.

(f) Predicting the future stock price of a company using historical records. (Cecilia Kong Xin Ru)

Yes,it is a task of data mining. Historical records of stock price can be used to create a predictive model called regression, one of the predictive modeling tasks that is used for continuous variables.

(g) Monitoring the heart rate of a patient for abnormalities.（yuhong）

Yes,because this data mining task is called anomaly detection.by observing the heart rate of the patient ,this data mining task can identify the abnormalities if the characteristics of the heart rate are different from normal observations.

(h) Monitoring seismic waves for earthquake activities.(Christopher Eng Kok Hong)

yes,as it is collecting data and interesting patterns and knowledge that is potentially useful for the future.

(i) Extracting the frequencies of a sound wave.(Chi Pui Mun)

no, as it is just an electrical engineering subfield that focuses on analysing, modifying such as sound.

Q2. Suppose that you are employed as a data mining consultant for an Internet search engine company. Describe how data mining can help the company by giving specific examples of how techniques, such as clustering, classification, association rule mining, and anomaly detection can be applied. (Cheng cai jie)

**Clustering**

Use to search through information that is exactly similar. Inputting a list of variables that classify the main idea that is being searched. Clustering will take the list of variables and match them with other main topics that contain the same variables.

**Classification**

Help by finding specific pieces of information more quickly by organizing that information in meaningful ways and identifying the categories that the information belongs within to search through that information classified.

**Association rule mining**

Showing attribute-value conditions that occur frequently together in a given set of data. It could append additional information in its result based on the keywords entered by the user.

**Anomaly detection**

Identifies data point, event, or observations that deviate from a dataset’s normal behaviour.

Q3. Classify the following attributes as binary, discrete, or continuous. Also classify them as qualitative (nominal or ordinal) or quantitative (interval or ratio). Some cases may have more than one interpretation, so briefly indicate your reasoning if you think there may be some ambiguity.

Example: Age in years. Answer: Discrete, quantitative, ratio

(a) Time in terms of AM or PM. (Chong Hao Yong)

* binary, quantitative, ratio(binary, qualitative, nominal)

(b) Brightness as measured by a light meter. (Kuah Jia Yu)

* Continuous, quantitative, ratio

(c) Brightness as measured by people’s judgments. (Sean Loi Yit Seng)

* Discrete, Qualitative, Ordinal

(d) Angles as measured in degrees between 0◦and 360◦. (Wong Yew Zun)

* Continuous, quantitative, ratio

(e) Bronze, Silver, and Gold medals as awarded at the Olympics. (Wong Weng Cheng Mavis)

* Discrete, qualitative, ordinal

(f) Height above sea level. (Ho Zhi Wen)

quantitative, interval

(g) Number of patients in a hospital. (Chong Hao Yong)

* Discrete, Quantitative, ratio

(h) ISBN numbers for books. (Look up the format on the Web.) (Chong Win Yan)

* discrete, qualitative, nominal

i) Ability to pass light in terms of the following values: opaque, translucent, transparent. (Nicholas Yue Qin Nam)

* Discrete, Qualitative, Ordinal

(j) Military rank. (Wong Yew Zun)

* Discrete, qualitative, ordinal

(k) Distance from the center of campus. (Lee Jing Jet)

Continuous ,quantitative ,ratio

(l) Density of a substance in grams per cubic centimeter. (Aaron Wong Zyi Syen)

-Continuous,quantitative,ratio

(m) Coat check number. (When you attend an event, you can often give your coat to someone who, in turn, gives you a number that you can use to claim your coat when you leave.) (Nicholas Yue Qin Nam)

- Discrete, Qualitative, Nominal

Q4. Give a suitable business example and a business activity that can use data mining to assist with the business activity. Then identify a data mining function and discuss how the data mining function can be applied to assist with the chosen business activity. After that briefly, discuss whether a data query or simple statistical analysis can assist with the data mining activity in your example. (Kae Lun)

-Marketing (Association rules, Sequential rules, Clustering, Classification)Basket Analysis

The marketing industry usually makes use of basket analysis to make assumptions about the customer’s behavior like items that are more likely to be purchased together while sequential rules, association rule and classification will be used in sorting products to convene the customers. In this state, data query can be utilized to retrieve data from the databases for analytical aims.

-Bank Industry(Anomaly Analysis, Fraud Detection)

The bank company uses anomaly detection to investigate the market risk. For example, they will use statistical analysis to predict the likelihood of repayment from the borrowers and fraud detection is being applied to on hold suspicious transactions that will cause a loss to the customer or bank before confirming the legitimation of the transactions.

Q5. There are few issues involving data quality assessment such as accuracy, completeness, and consistency.

For each of the above-mentioned **THREE (3)** issues, discuss: how data quality assessment can depend on the intended use of the data. Support your answers with relevant examples.

Answer:(Jessie Liew)

Accuracy - higher accuracy, higher data quality assessment. More reliable. Example, M is male, not female.

Completeness - higher completeness, higher data quality assessment. less error, more knowledge can be obtained. Example, if everyone fills in their age in a survey, statistics about age range can be generated, therefore more discovery can be obtained.

Consistency - higher consistency, higher data quality assessment. less error occurs. Example, when you will be happy if your wifi speed is consistent when you are playing online games, if not, you will go angry.

Q6. Discuss all dimensions that can be used to assess the quality of data.(Yu Hong)

Completeness, Timeliness, Validity, Consistency, Uniqueness and Accuracy

**Accuracy**-indicates the data store are the correct value.Example, the birthday have different format depending on the region such as 12/02/2021, Feb 12, 2021 and 02/12/2021 which will cause the incorrect value being input into a database

**Completeness**-defines that the dataset is free of missing values and unavailable data.For instance, some of the attributes might have empty or missing values because of certain reasons like privacy violation and uncertainty. These missing values must be handled by filling in or removing the tuple in order to continue further analysis.

**Consistency**-the data storing format sometimes might lead to ambiguity.Hence, it is a need to ensure all data are consistent in the format for example gender can be stored as ‘male’ or ‘female’, ‘m’ or ‘f’ and in binary format.Other than that, when dealing with merging of databases in data integration, attributes or classes will have inconsistency in naming convention.Example, the birthday can be represented in ‘DOB’ or ‘Date Of Birth’.

**Accessibility**-the degree of availability and how easily the data can be accessed or retrieved.Besides, it refers to the ease of mining information and knowledge from the data.

Q7.Assume the following data is given: {22, 12, 61, 57, 30, 1, 32, 37, 37, 68, 42, 11, 25, 7, 8, 16}.

Rearrange:{1,7,8,11,12,16,22,25,30,32,37,37,42,57,61,68}

(i) Apply data discretization by binning the data into 4 bins using equal-depth and equi-width binning, respectively.

Equal-Depth : 16/4=4(size/frequency of 4)

Bin 1:1,7,8,11

Bin 2:12,16,22,25

Bin 3:30,32,37,37

Bin 4:42,57,61,68

Equal-Width:(68-1)/4=16.75(interval of approximately 17)

Bin 1(1-18):1,7,8,11,12,16

Bin 2(18-35):22,25,30,32

Bin 3(35-52):37,37,42

Bin 4(52-69):57,61,68

(ii) Distinguish the two binning methods. Give an appropriate example application for each of the binning methods.

Binning with equal width is the most straightforward but the outliers may dominate presentation and skewed data is not handled well.

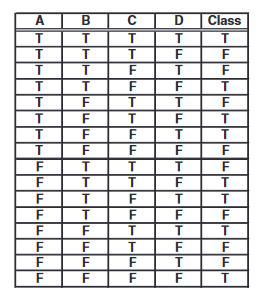
Equal depth,good data scaling but managing categorical attributes can be tricky.

(iii) If you know that the data represent ages of persons, what kind of binning method would you then use?

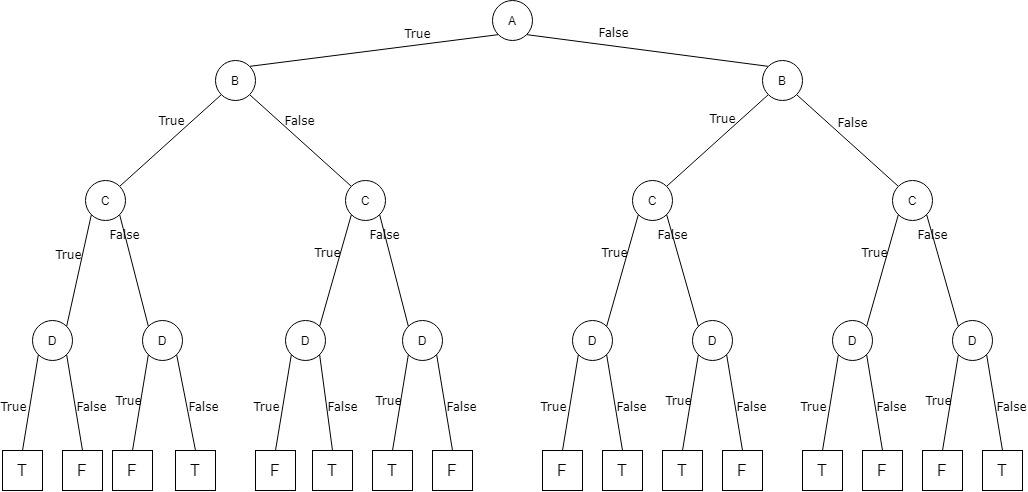
i will use equal depth as it have better data scaling.?

Equal-width binning because the age can be organized in a uniform range and the exact values can be seen in every bin.Add on, outliers might be removed using this binning method.Can estimate the skewness or not?

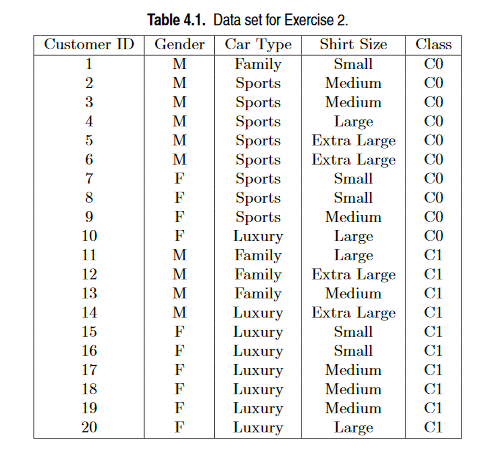
| Q8. Draw the full decision tree for the parity function of four Boolean attributes,A,B, C, and D. Is it possible to simplify the tree? |
| --- |



Answer:



Q9.Consider the training examples shown in Table 4.1 for a binary classification problem.



(a) Compute the Gini index for the overall collection of training examples.

C0 : 10/20

C1 : 10/20

1 - [(10/20)² + (10/20)²] = 0.5

(b) Compute the Gini index for the Customer ID attribute.

0

(c) Compute the Gini index for the Gender attribute.

M : 10/20, 6/10 in C0, 4/10 in C1

F : 10/20, 4/10 in C0, 6/10 in C1

Gini(M) = 1 - [(6/10)² + (4/10)²] = 0.48

Gini(F) = 1 - [(4/10)² + (6/10)²] = 0.48

Gini(Gender) = (10/20)(0.48) + (10/20)(0.48) = 0.48

(d) Compute the Gini index for the Car Type attribute using a multiway split.

Family : 4/20, ¼ in C0, ¾ in C1

Sports : 8/20, 8/8 in C0, 0/8 in C1

Luxury : 8/20, ⅛ in C0, ⅞ in C1

Gini(Family) = 1 - [(1/4)² + (3/4)²] = 0.375

Gini(Sports) = 1 - [(8/8)² + (0/8)²] = 0

Gini(Luxury) = 1 - [(1/8)² + (7/8)²] = 0.21875

Gini(Car Type) = (4/20)(0.375) + (8/20)(0) + (8/20)(0.21875) = 0.1625

(e) Compute the Gini index for the Shirt Size attribute using multiway split.

Small : 5/20,3/5 in C0,2/5 in C1

Medium : 7/20,3/7 in C0,4/7 in C1

Large : 4/20, 2/4 in C0,2/4 in C1

Extra Large : 4/20,2/4 in C0,2/4 in C1

Gini(Small) = 1 - [(⅗)²+(⅖)²] = 0.48

Gini(Medium) = 1-[(3/7)²+(4/7)²]=0.4898

Gini(Large) = 1 - [(2/4)2+ (2/4)2 ] = 0.5

Gini(Extra Large) = 1 - [(2/4)2] + [(2/4)2] = 0.5

Gini(Shirt Size) = (5/20)(0.48) +(7/20)(0.4898) + (4/20)(0.5) + (4/20)(0.5)=0.49143

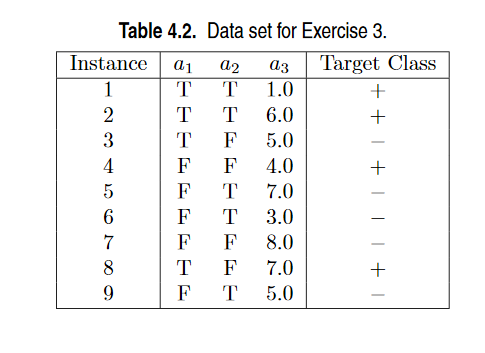
(f) Which attribute is better,Gender,Car Type,orShirt Size?

-Car Type attribute. It has the lowest gini index among other attributes. (lower gini,higher gain)(1-gini=gain)

(g) Explain why Customer ID should not be used as the attribute test condition even though it has the lowest Gini.

-Customer ID has no predictive power due to the fact that they have assigned value.

Q10. Consider the training examples shown in Table 4.2 for a binary classification problem.



1. What is the entropy of this collection of training examples with respect to the positive class?

P(+) = 4/9

P(-) = 5/9

Entropy = - [ (4/9)log2(4/9) + (5/9)log2(5/9) ]

= 0.9911

= 0.9911

There are 4 positive examples, and 5 negative examples. Thus, P(+) is 4/9 and P(-) is 5/9. The entropy of the training examples is 0.9911.

1. What are the information gains of a1 and a2 relative to these training examples?

For attribute a1, the corresponding counts and probabilities are:

| a1 | + | - | Total |
| --- | --- | --- | --- |
| T | 3 | 1 | 4 |
| F | 1 | 4 | 5 |

Entropy(T) = - [ (3/4)log2(3/4) + (1/4)log2(1/4) ]

Entropy(F) = - [ (1/5)log2(1/5) + (4/5)log2(4/5) ]

Entropy(a1) = (4/9)[ - [ (3/4)log2(3/4) + (1/4)log2(1/4) ] ]

* (5/9)[ - [ (1/5)log2(1/5) + (4/5)log2(4/5) ] ]

= (4/9)[ 0.8113 ] + (5/9)[ 0.7219 ] = 0.7616

= 0.7616

Gain(a1) = 0.9911 - 0.7616 = 0.2294

The entropy for a1 is (4/9)[ - [ (3/4)log2(3/4) + (1/4)log2(1/4) ] ] + (5/9)[ - [ (1/5)log2(1/5) + (4/5)log2(4/5) ] ] = 0.7616.

Therefore, the information gain for a1 is 0.9911 - 0.7616 = 0.2294.

For attribute a2, the corresponding counts and probabilities are:

| a2 | + | - | Total |
| --- | --- | --- | --- |
| T | 2 | 3 | 5 |
| F | 2 | 2 | 4 |

Entropy(T) = -[ (⅖)log2(⅖) + (⅗)log2(⅗) ] = 0.97095

Entropy(F) = -[ (2/4)log2(2/4) + (2/4)log2(2/4) ] = 1

Entropy(a2) = (5/9)(-[ (⅖)log2(⅖) + (⅗)log2(⅗) ])+(4/9)( -[ (2/4)log 2(2/4) + (2/4)log2(2/4) ])

(5/9)(0.97095)+(4/9)(1)=0.983861111111111111111111111111

=0.9839

Gain(a2) = 0.9911 - 0.98386 = 0.00724

= 0.0072

(c) For a3, which is a continuous attribute, compute the information gain for every possible split.

| a3 | Class Label | Split Point | Entropy | Info Gain |
| --- | --- | --- | --- | --- |
| 1.0 | + | 2.0 | 0.8484 | 0.1427 |
| 3.0 | - | 3.5 | 0.9885 | **0.0026** |
| 4.0 | + | 4.5 | 0.9183 | 0.0728 |
| 5.0 | - | 5.5 | 0.9839 | 0.0072 |
| 5.0 | - |
| 6.0 | + | 6.5 | 0.9728 | 0.0183 |
| 7.0 | - | 7.5 | 0.8889 | 0.1022 |
| 7.0 | + |
| 8.0 | - |  |  |  |

Entropy(<2.0) = - [1/9[ (1/1)log2(1/1) + 0 ] + 8/9[ (3/8)log2(3/8) + (5/8)log2(5/8) ]]

= (-1/9)[ (1/1)log2(1/1) + 0 ]

+ (-8/9)[ (3/8)log2(3/8) + (5/8)log2(5/8) ]

= 0.8484

Info Gain = 0.9911 - 0.8484 = 0.1427

Entropy(<3.5) = - [2/9[(1/2)log2(1/2)+(1/2)log2(1/2)]+7/9[(4/7)log2(4/7)+(3/7)log2(3/7)]

= 0.98851

Info Gain = 0.9911 - [- [2/9[(1/2)log2(1/2)-(1/2)log2(1/2)]+7/9[(4/7)log2(4/7)+(3/7)log2(3/7)]]

= **0.00259**

Entropy(<4.5)= -[3/9[(2/3)log2(2/3)+(1/3)log2(1/3)]+[6/9[(2/6)log2(2/6)+(4/6)log2(4/6)]]

=0.91830

info Gain= 0.9911 - 0.91830 = **0.0728**

Entropy(<5.5) = -[5/9[(2/5)log2(2/5)+(3/5)log2(⅗)]+4/9[(2/4)log2(2/4)+(2/4)log2(2/4)]]

= - (-0.539417-0.4444444) =0.9839

Info Gain = 0.9911 -0.9839 = 0.0072

Entropy(<6.5) = - [6/9[(3/6)log2(3/6) + (3/6)log2(3/6)] + 3/9[(1/3)log2(1/3) + (2/3)log2(2/3)]]

= -(-0.666667 + (- 0.306099)) = 0.9728

Info Gain = 0.9911 - 0.9728 = **0.0183**

Entropy (<7.5)=-[8/9 [ (4/8)log2(4/8)+(4/8)log2(4/8) ] +1/9 [ 0+(1/1)log2(1) ]]

Info Gain = 0.9911 - 0.88889 = **0.1022**

**The best split for a3 occurs at split point equals to 2.**

1. What is the best split (among a1,a2,and a3) according to the information gain?

-->According to the information gain, a1 produces the best split as it has the highest information gain.

1. What is the best split (between a1 and a2) according to the classification error rate?

-a1: error rate = 2/9

-a2: error rate = 4/9

-->Therefore, according to the classification error rate, a1 produces the best split.(a1 lower error rate)

1. What is the best split (between a1 and a2) according to the Gini index.

a1

| a1 | + | - | Total |
| --- | --- | --- | --- |
| T | 3 | 1 | 4 |
| F | 1 | 4 | 5 |

T - 4/9,3/4 in “+”,1/4 in “-”

F - 5/9,1/5 in “+”,4/5 in “-”

Gini(T) = 1 - [(3/4)^2+(1/4)^2] = 0.375

Gini(F) = 1 - [(1/5)^2+(4/5)^2] = 0.32

Gini(a1) = (4/9)(0.375)+(5/9)(0.32) = 0.3444

a2

T - 5/9,2/5 in “+”, 3/5 in “-”

F - 4/9, 2/4 in “+”, 2/4 in “-”

Gini(T) = 1 - [(2/5)^2+(3/5)^2] = 0.48

Gini(F) = 1 - [(2/4)^2+(2/4)^2] = 0.5

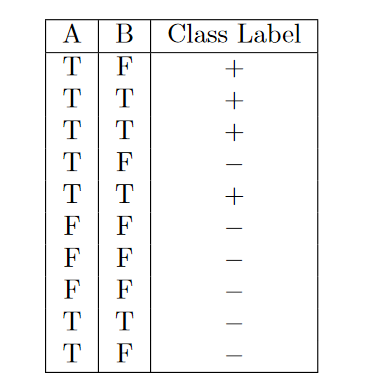
Gini(a2) = (5/9)(0.48)+(4/9)(0.5) = 0.4889

Gini(a1) = 0.3444

Gini(a2) = 0.4889

-->Therefore, according to the Gini Index, a1 is the best split as it has the lowest gini index which is 0.3444 compared to a2.

Q11. Consider the following data set for a binary class problem.



1. Calculate the information gain when splitting on A and B.Which attribute would the decision tree induction algorithm choose?

P(+)=4/10

P(-)=6/10

Entropy = -[(4/10)log2(4/10)+(6/10)log2(6/10)]

= 0.9710

Entropy(A) = (7/10)[-[(4/7)log2(4/7)+(3/7)log2(3/7)]]+(3/10)[-[(0/3)log2(0/3)+(3/3)log2(3/3)]]

= 0.6897

Information Gain(A)=0.9710 - 0.6897 = 0.2813

Entropy(B) = (4/10)[- [(3/4)log2(3/4) + (1/4)log2(1/4)]] + (6/10)[-[(1/6)log2(1/6) + (5/6)log2(5/6)]] = 0.324511 + 0.390013 = **0.7145**

Information Gain(B)=0.9710 - 0.7145 = **0.2565**

1. Calculate the gain in the Gini index when splitting on A and B. Which attribute would the decision tree induction algorithm choose?

Overall Gini index before splitting:

1 - [(4/10)2 + (6/10)2] = 0.48

Gini index after splitting on A:

Gini(A=T) = 1 - [(4/7)2 + (3/7)2] = 0.4898

Gini(A=F) = 1 - [(0/3)2 + (3/3)2] = 0

Gini(A) = (7/10)(0.4898) + (3/10)(0) = 0.3429

Gain(A) = 0.48 - 0.3429 = 0.1371

Gini index after splitting on B:

Gini(B=T) = 1 - [(3/4)2 + (1/4)2] = 0.375

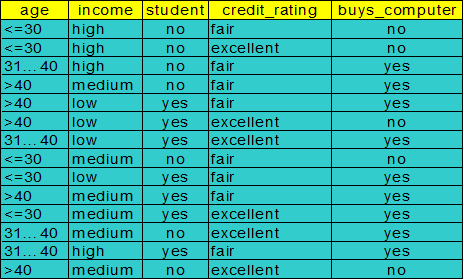
Gini(B=F) = 1 - [(1/6)2 + (5/6)2] = 0.2778

Gini(B) = (4/10)(0.375) + (6/10)(0.2778) = 0.3167

Gain(B) = 0.48 - 0.3167 = 0.1633

Therefore, attribute B will be chosen to split because of higher gain.

Additional:(two-way splits) \*for Gini Index



Gini(D) = 1 - [(9/14)^2 + (5/14)^2]

= 1 - 53/98

= 0.4592

| Subset | Yes | No | Total |
| --- | --- | --- | --- |
| Low | 3 | 1 | 4 |
| Medium | 4 | 2 | 6 |
| High | 2 | 2 | 4 |

After splitting: ~

{low, medium, high} **X not represent a split**

{low, medium} = **0.443**

{low, high} = 0.458

{medium, high} = 0.450

{low} = 0.450

{medium} = 0.458

{high} = **0.443**

{ } **X not represent a split**

Gini income E {low} = 4/14 (1 - [(3/4)^2 + (1/4)^2]) + 10/14 (1 - [(6/10)^2 + (4/10)^2])

= 0.450

Gini income E {medium,high}=10/14 (1 - [(6/10)^2 + (4/10)^2])+4/14 (1 - [(3/4)^2 + (1/4)^2])

=0.450

Gini income E {medium}= 6/14 (1 - [(4/6)^2 + (2/6)^2]) + 8/14 (1 - [(5/8)^2 + (3/8)^2])

= 0.458

Gini income E {low, high}= 8/14 (1 - [(5/8)^2 + (3/8)^2]) + 6/14 (1 - [(4/6)^2 + (2/6)^2])

= 0.458

Gini income E {high]=4/14 ( 1 - [ (2/4)^2 + (2/4)^2 ]) + 10/14 (1 - [(7/10)^2 + (3/10)^2])

= 0.443

Gini income E {low, medium}=10/14 (1 - [(7/10)^2 + (3/10)^2])+4/14 ( 1 - [ (2/4)^2 + (2/4)^2 ])

= 0.443

Reduction in Impurity:

Gain(income) = Gini(D) - Gini\_A(D)

= 0.459 - 0.443

= 0.016

Q12. Consider a training set that contains 100 positive examples and 400 negative examples. For each of the following candidate rules,

R1:A−→+ (covers 4 positive and 1 negative examples),

R2:B−→+ (covers 30 positive and 10 negative examples),

R3:C−→+ (covers 100 positive and 90 negative examples),

determine which is the best and worst candidate rule according to:

1. Rule accuracy.

accuracy(R) = ncorrect / ncovers

R1=⅘\*100% = 80%

R2 = 30/40 \* 100% = 75%

R3 = 100/190\*100% = 52.63%

1. FOIL’s information gain.

P0=100

N0=400

Information gain(R0,R1)=4(log2(⅘)-log2(100/500))

=8.0004

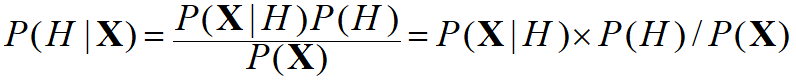
Information gain(R0,R2)=30(log2(30/40)-log2(100/500))

=57.21

Information gain(R0,R3)=100(log2(100/190)-log2(100/500))

=139.6

Q13 (Bayes Theorem C4 Part1 slide 32)

1. Suppose the fraction of undergraduate students who smoke is 15% and the fraction of graduate students who smoke is 23%. If one-fifth of the college students are graduate students and the rest are undergraduates,what is the probability that a student who smokes is a graduate student? P(G|S)

UG = Undergraduate student  
S = Smoker  
G = Graduate student

P(UG) = ⅘ , P(G) = ⅕ , P(S|UG)=0.15 , P(S|G)=0.23

P(S) = [P(S|UG) \* P(UG)] + [P(S|G) \* P(G)]

= (0.15 \* 0.8) + (0.23 \* 0.2)

= 0.166

P(UG|S)=P(S|UG)\*P(UG)/P(S)=(0.15\*4/5)/0.166=0.722891566

P(G|S)=P(S|G)\*P(G)/P(S)=(0.23\*1/5)/0.166=0.277108433

1. Given the information in part (a), is a randomly chosen college student more likely to be a graduate or undergraduate student?

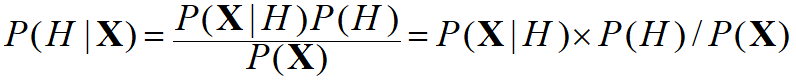
Undergraduate students. P(UG) = ⅘, is higher than P(G) = ⅕. Probability of undergraduate students is higher than the probability of graduate students.

1. Repeat part (b) assuming that the student is a smoker.

Undergraduate student,P(UG|S) = 0.7229, is higher than P(G|S) = 0.2771

Probability of undergraduate students is higher than the probability of graduate students.

1. Suppose 30% of the graduate students live in a dorm but only 10% of the undergraduate students live in a dorm. If a student smokes and lives in the dorm, is he or she more likely to be a graduate or undergraduate student? You can assume independence between students who live in a dorm and those who smoke.



D = students live in a dorm

P(D|G) = 0.3, P(D|UG) = 0.1

P(D) = [(P(D|UG)\*P(UG)] + [P(D|G)\*P(G)]

= (0.1\*4/5) + (0.3\*0.2)

= 0.14

P(D n S) = 0.14 \* 0.166

= 0.04284

P(D n S | G) = P(D|G) \* P(S|G)

= 0.3 \* 0.23

= 0.069

P(G | D n S) = [P(D n S | G)\*P(G)/ P(D n S)]

= 0.069 \* ⅕ /0.04284

= 0.3221

P(D n S | UG) = P(D|UG)\*P(S|UG)

=0.1\*0.15

= 0.015

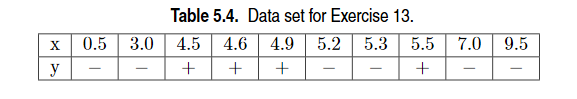
P(UG | D n S) = [P(D n S|UG)xP(UG)/P(D n S)]

= [(0.015)(⅘)/(0.04284)]

= 0.2801

He or she more likely to be a graduate student, because P(G|D n S) = 0.3221, is higher than P(UG|D n S) = 0.2801.

Q14 . Consider the one-dimensional data set shown in Table 5.4.



1. Classify the data point x=5.0 according to its 1-, 3-, 5-, and 9-nearestneighbors (using majority vote)

| x | y | difference  \*distance=difference\*10 (ex:0.1=1) | Weight  (100/distance^2) | Vote(w\*D) |
| --- | --- | --- | --- | --- |
| 0.5 | - | 4.5 | 0.049383 | 0.2222 |
| 3.0 | - | 2.0 | 0.250000 | 0.5 |
| 4.5 | + | 0.5 | 4.0 | 2.0 |
| 4.6 | + | 0.4 | 6.25 | 2.5 |
| 4.9 | + | 0.1 | 100 | 10 |
| 5.2 | - | 0.2 | 25 | 5 |
| 5.3 | - | 0.3 | 11.1111 | 3.3333 |
| 5.5 | + | 0.5 | 4.0 | 2.0 |
| 7.0 | - | 2.0 | 0.25 | 0.5 |
| 9.5 | - | 4.5 | 0.049383 | 0.2222 |

For data point 5.0,

1-NN = 4.9 ⇒ +

3-NN = 4.9,5.2,5.3 ⇒ -

5-NN = 4.6,4.9,5.2,5.3,5.5 ⇒ +

9-NN = 3.0,4.5,4.6,4.9,5.2,5.3,5.5,7.0,9.5 ⇒ -

1. Repeat the previous analysis using the distance-weighted voting approach described in Section 5.2.1.

1-NN=(0.1\*100)=10=>+

3-NN = (0.1\*100) = 10 or 5+3.33 = 8.33 ⇒ +

5-NN = (6.25\*0.4)+(100\*0.1)+(4.0\*0.5)=14.5 or (0.2\*25)+(0.3\*11.11) = 8.33 ⇒ +

9-NN = 2.0+2.5+10+2.0=16.5 or 0.22+0.5+5+3.33+0.5+0.22 = (9.77) ⇒ +