CS 422

Data Mining

Homework Assignment 1

Ronit Rudra
A20379221
rrudra@hawk.iit.edu

Course Instructor
Prof. Jawahar Panchal

1 Textbook Questions

1.1 Chapter 1

1.1.1 Question 1

(a) This is not a data mining task as division of observations based on a particular attribute is a simple case of sorting the entries.

(b) As profitability of each customer is already known, the customers are subsetted based on bounds or ranges. Thus, this is not a data mining task.

(c) Computing the total sales of a company is a straightforward calculation. Hence, not a data mining task.

(d) This is the same as (a). Sorting a database is not a data mining task.

(e) Since the die is fair, probability of each outcome is exactly 1/6. This is a simple probability calculation. Furthermore, past data is not used to predict future outcomes. Thus, it is not a data mining task.

(f) Prediction of time series data such as stock prices is considered as a data mining task as historical data is gathered, analyzed and a model is developed which would hopefully predict future values with some degree of accuracy. ARMA, ARIMA are some of the methods which would be used to perform this task.

(g) Monitoring the heart rate for abnormalities would be considered as a data mining task. Since each person has different quantitative parameters for a healthy heart, it would be very difficult to set a standardized value for these. A model needs to be developed which can generalize how a normal heart should behave like. Any outliers would be flagged as abnormal behavior and would result in an alarm.

(h) This can be a data mining task. A seismic counter throws an alarm when waves of sufficient magnitude are detected. This can result in a lot of false positives as nearby subterranean drilling, explosions etc. could result in a shockwave massive enough to trigger the alarm. A model can be developed to learn how to classify the waves as authentic earthquakes or otherwise. The model could also be tailored for geo-specific areas by incorporating data on frequency, magnitude, duration etc. to better classify or predict earthquakes.

(i) Extracting the frequencies of a sound wave is a signal processing task and is not data mining as it involves a fixed operation which produces the exact intended result all the time.

1.2 Chapter 2

1.2.1 Question 2

(a) This is binary as there are only two values AM and PM. It is qualitative and ordinal as AM and PM are just labels to arrange forenoon and afternoon.

(b) Brightness measured by a light meter will give out a continuous stream of data which is quantifiable. Since its a quantity, it can be represented as a ratio.

(c) Asking people to measure brightness would result in them giving discrete interpretations. The answers would be qualitative as they would mention things like, dark, low light, dull, bright etc. and these labels would be ordinal as there is a fixed place for them in order of brightness(dark, dull, bright).

(d) Measurment of angles is continuous (subject to least count), quantitative and a ratio.

(e) Medals awarded are discrete (3 values), qualitative and ordinal (gold > silver > bronze).

(f) Height above sea level is a measurement thus is continuous and quantitative (ratio).

(g) The number of patients in a hospital is a discrete quantity (ratio).

(h) ISBN numbers are unique codes for a book (even different editions). This is a discrete, qualitative value and is nominal as there is no particular ordering between the different values.

(i) Opaque, translucent and transparent are discrete, qualitative ordinal values as there is a set order.

(j) Military rank is discrete, qualitative and ordinal (General > Lt. Gen > Maj. Gen > Brig. > Col. and so on)

(k) Distance from campus center would be a continuous, quantitative value. This would be a ratio or an interval depending on whether exact values are required or ranges such as within 500 feet, within a mile, 1 to 2 miles etc.

(l) Density in grams per cubic centimeter is continuous, quantitative and a ratio. On the other hand, it can also be considered discrete as at the interface between the material and the surroundings, there is a sharp discontinuous change in density. Thus, density being continuous or discrete is context specific.

(m) Coat check number is discrete, qualitative and nominal as the number is simply a numeric placeholder for the coat in no specific order. The number provided simply corresponds to a vacant rack where the coat is hung.

1.2.2 Question 7

Temporal autocorrelation is high for data which is measured closer in time. Daily temperatures have a high degree of autocorrelation for a particular location while amount of rainfall is highly unpredictable (even the best meteorological estimates have erroneous predictions) as it varies locally with time. Suppose, we have the data for both rainfall and temperature for a particular location over the span of a few years. Autocorrelation of the temperature and rainfall data for, let's say, year 1 and year 2 would yield a higher value for temperature as the the quantitative value of temperature would be closer than that of rainfall. For example, in year 1 for a particular day the temperature and rainfall are 28° celsius and 3 mm. For year 2 on the same day we could expect the temperature to be roughly the same but would have no idea on how much rainfall would occur.

1.3 Question 18

(a) x = 0101010001
y = 0100011000
The L1 or hamming distance is:

0101010001
0100011000
||||||||||
0001001001

Summing it up: the L1 distance is 3

The jaccard similarity is given by: J = number of matching presences/number of attributes not involved in 00 matches = $\frac{f_{11}}{(f_{01}+f_{10}+f_{11})}$ Thus $J = \frac{2}{(2+1+2)} = \frac{2}{5} = 0.5$

(b) Before comapring similarity of the different distance and similarity metrics, let us examine how each one is calculated. The Simple Matching Coefficient is the ratio of matches to the total number of bits i.e. $\frac{(f_{11}+f_{00})}{(f_{11}+f_{00}+f_{10}+f_{01})}$ The Hamming Distance is the total number of matches i.e $(f_{11} + f_{00})$ The Jaccard Similarity is number of 11 matches excluding 00 matches i.e $\frac{f_{11}}{(f_{01}+f_{10}+f_{11})}$ The Cosine similarity is the angle between the vectors i.e.

$$\frac{x \cdot y}{|x||y|}$$

Just by looking at the equations it is obseved that SMC and Hamming are similar as they involve both 00 and 11 matching.

On the other hand Jaccard ignores 00 matches. This is the same for cosine as 00 matches do not contribute to the angle calculation (addition of zero) Thus Jaccard and Cosine are similar.

(c) Since we are comparing the similarity between two genes (given that 1 and zero represent presence and absence of attributes respectively), it would be reasonable and intuitive to use the Jaccard similarity as hamming distance calculates dissimilarity between two vectors. Since two organisms of different species would not have the same genetic makeup, calculating the hamming distance would be useless as we already know they are different. The Jaccard similarity would give us an approximate idea of which genes they have in common.

(d) Now we are comparing two humans. Since two humans share 99.9% of the same genes, the jaccard similarity would yield a very high value. What we need is a metric to differentiate between the two. The hamming distance would be a more appropriate choice. (This question is the exact opposite of the previous one)

1.3.1 Question 19

For this question Cosine will be represented as C, Euclidean as E, Jaccard as J, and Correlation as R.

$$C = \frac{x \cdot y}{|x| \cdot |y|}$$

$$E = |x - y|$$

$$J = \frac{f_{11}}{(f_{01} + f_{10} + f_{11})}$$

$$R = \frac{s_{xy}}{s_y \cdot x_y}$$

$$s_{xy} = \frac{1}{n-1} \sum_{k=1}^{n} (x_k - \overline{x})(y_k - \overline{y})$$

$$s_x = \sqrt{\frac{1}{n-1} \sum_{k=1}^{n} (x_k - \overline{x})^2}$$

$$\overline{x} = \frac{1}{n} \sum_{k=1}^{n} x_k$$

(a) X = (1,1,1,1), Y = (2,2,2,2)

$$C = \frac{(1 \cdot 2 + 1 \cdot 2 + 1 \cdot 2 + 1 \cdot 2)}{(\sqrt{1+1+1+1}) \cdot \sqrt{4+4+4+4})} = \frac{8}{2 \cdot 4} = 1$$

$$s_{xy} = 0$$

$$s_x = 0 = s_y$$

$$R = \frac{0}{0}$$

$$E = \sqrt{(1-2)^2 + (1-2)^2 + (1-2)^2 + (1-2)^2} = 2$$

(b) X = (0,1,0,1), Y = (1,0,1,0)

$$C = \frac{(\cdot 1 + 1 \cdot 0 + 0 \cdot 1 + 1 \cdot 0)}{(\sqrt{0+1+0+1}) \cdot \sqrt{1+0+1+0})} = \frac{0}{2} = 0$$

$$E = \sqrt{(0-1)^2 + (1-0)^2 + (0-1)^2 + (1-0)^2} = 2$$

$$J = \frac{0}{2+2} = 0$$

$$\overline{x} = \frac{2}{4} = \frac{1}{2}$$

$$\overline{y} = \frac{2}{4} = \frac{1}{2}$$

$$s_x = \sqrt{\frac{1}{3}(0.25 + 0.25 + 0.25 + 0.25)} = \sqrt{\frac{1}{3}}$$

$$s_y = \sqrt{\frac{1}{3}(0.25 + 0.25 + 0.25 + 0.25)} = \sqrt{\frac{1}{3}}$$

$$s_{xy} = \frac{1}{3}[(0 - 0.5)(1 - 0.5) + (1 - 0.5)(0 - 0.5) + (0 - 0.5)(1 - 0.5) + (1 - 0.5)(0 - 0.5)] = \frac{-1}{3}$$

$$R = \frac{\frac{-1}{3}}{\sqrt{\frac{1}{3}} \cdot \sqrt{\frac{1}{3}}} = -1$$

(c) X = (0,-1,0,1), Y = (1,0,-1,0)

$$C = \frac{(0 + 0 + 0 + 0)}{\sqrt{1 + 1} \cdot \sqrt{1 + 1}} = 0$$

$$E = \sqrt{1 + 1 + 1 + 1} = 2$$

$$\overline{x} = \frac{1 - 1}{4} = 0$$

$$\overline{y} = \frac{1 - 1}{4} = 0$$

$$s_x = \sqrt{\frac{1}{3}(0 + 1 + 0 + 1)} = \sqrt{\frac{2}{3}}$$

$$s_y = \sqrt{\frac{1}{3}(1 + 0 + 1 + 0)} = \sqrt{\frac{2}{3}}$$

$$s_{xy} = \frac{1}{3}[(0 - 0)(1 - 0) + (-1 - 0)(0 - 0) + (0 - 0)(-1 - 0) + (1 - 0)(0 - 0)] = 0$$

$$R = \frac{0}{\sqrt{\frac{2}{3}} \cdot \sqrt{\frac{2}{3}}} = 0$$

(d) X = (1,1,0,1,0,1), Y = (1,1,1,0,0,1)

$$C = \frac{(1+1+0+0+0+1)}{\sqrt{1+1+0+1+0+1}\cdot\sqrt{1+1+1+0+0+1}} = \frac{3}{4} = 0.75$$

$$J = \frac{1+1+1}{1+1+1+1+0+1} = \frac{3}{5} = 0.6$$

$$\overline{x} = \frac{1+1+0+1+0+1}{6} = \frac{4}{6} = 0.67$$

$$\overline{x} = \frac{1+1+1+0+0+1}{6} = \frac{4}{6} = 0.67$$

$$s_x = \sqrt{\frac{1}{5}(4(0.33)^2 + 2(-0.67)^2)} = 0.51$$

$$s_y = \sqrt{\frac{1}{5}(4(0.33)^2 + 2(-0.67)^2)} = 0.51$$

$$s_{xy} = \frac{1}{5}[(0.33)(0.33) + (0.33)(0.33) + (-0.67)(0.33) + (0.33)(-0.67) + (-0.67)(-0.67) + (0.33)(0.33)]$$

$$R = \frac{0.06668}{0.51 \cdot 0.51} = 0.25$$

(e) X = (2,-1,0,2,0,-3), Y = (-1,1,-1,0,0,-1)

$$C = \frac{-2-1+0+0+3}{\sqrt{4+1+4+9}\cdot\sqrt{1+1+1+1}} = 0$$

$$\overline{x} = \frac{2-1+2-3}{6} = 0$$

$$\overline{y} = \frac{-1+1-1-1}{6} = \frac{-1}{3}$$

$$s_x = \sqrt{\frac{1}{5}(4+1+4+9)} = 1.9$$

$$s_y = \sqrt{\frac{1}{5}(1.35+1.77+0.21)} = 0.816$$

$$s_{xy} = \frac{1}{5}[(2)(-0.67) + (-1)(1.33) + (0)(-0.67) + (2)(0.33) + (0)(0.33) + (-3)(-0.67)] = 0$$

$$R = \frac{0}{1.9 \cdot 0.816} = 0$$

1.4 Chapter 3

1.4.1 Question 8

A boxplot or whisker plot visualizes the total range, interquartile range, median as well as any outliers a particular attribute vector may have. The box inside the plot (between the minima and maxima) shows how much the values are skewed around the median. The median is the 50th percentile and for symmetric distribution the datapoints are equally divided on both sides of the median. From figure 3.11 in the book, the sepal length is symmetric; the sepal width is squashed around the median but is symmetric although the datapoints are very close to the median; the petal length is higly skewed as most datapoints lie below the median; the petal width is mildly skewed around the median.

1.5 Chapter 4

1.5.1 Question 2

(a) The overall Gini index is 0.5 as both class C0 and C1 have 10 observations each.

(b) Total Gini Impurity = 0

Table 1: Customer ID Gini Impurity

CID	Gini
1	$1 - \frac{1}{1} - \frac{0}{1} = 0$
2	$1 - \frac{1}{1} - \frac{0}{1} = 0$
.	....
.	....
.	....
.	....
20	$1 - \frac{0}{1} - \frac{1}{1} = 0$

(c)

$$Male : 1 - (\frac{6}{10})^2 - (\frac{4}{10})^2 = 0.48$$
$$Female : 1 - (\frac{4}{10})^2 - (\frac{6}{10})^2 = 0.48$$
$$total : \frac{10}{20} \cdot (0.48) + \frac{10}{20} \cdot (0.48) = 0.48$$

Table 2: Gender Gini Impurity

Class	Male	Female
C0	6	4
C1	4	6

(d)

Table 3: Car Type Gini Impurity

Class	Family	Sports	Luxury
C0	1	8	1
C1	3	0	7

$$Family : 1 - (\frac{1}{4})^2 - (\frac{3}{4})^2 = 0.375$$
$$Sports : 1 - (\frac{8}{8})^2 - (\frac{0}{8})^2 = 0$$
$$Luxury : 1 - (\frac{1}{8})^2 - (\frac{7}{8})^2 = 0.219$$
$$total : \frac{4}{20} \cdot (0.375) + \frac{8}{20} \cdot (0) + \frac{8}{20} \cdot (0.219) = 0.163$$

(e)

Table 4: Shirt Size Gini Impurity

Class	Small	Meduim	Large	Extra Large
C0	3	3	2	2
C1	2	4	2	2

$$Small : 1 - (\frac{3}{5})^2 - (\frac{2}{5})^2 = 0.48$$
$$Medium : 1 - (\frac{3}{7})^2 - (\frac{4}{7})^2 = 0.49$$
$$Large : 1 - (\frac{1}{2})^2 - (\frac{1}{2})^2 = 0.5$$
$$ExtraLarge : 1 - (\frac{1}{2})^2 - (\frac{1}{2})^2 = 0.5$$
$$total : \frac{5}{20} \cdot (0.48) + \frac{7}{20} \cdot (0.49) + \frac{4}{20} \cdot (0.5) + \frac{4}{20} \cdot (0.5) = 0.49$$

(f) Car type is the best attribute for classification as it has the lowest gini impurity among the three.

(g) Customer ID is just a unique identifier and not an attribute. Every new observation is assigned a new and unique customer ID and does not contribute towards a classification attibute even though it's gini impurity is zero.

1.5.2 Question 3

The data provided is as follows:

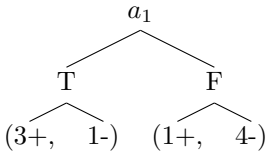
Table 5: Data

Instance	a_1	a_2	a_3	Target class
1	T	T	1.0	+
2	T	T	6.0	+
3	T	F	5.0	-
4	F	F	4.0	+
5	F	T	7.0	-
6	F	T	3.0	-
7	F	F	8.0	-
8	T	F	7.0	+
9	F	T	5.0	-

(a)

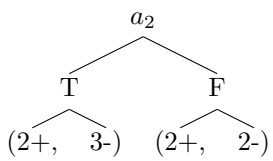
$$P(+) = \frac{4}{9} P(-) = \frac{5}{9}$$
$$Entropy(+) = -\frac{4}{9} log_2 \frac{4}{9} = 0.5199$$
$$Entropy(-) = -\frac{5}{9} log_2 \frac{5}{9} = 0.471$$
$$Total = 0.5199 + 0.471 = 0.9909$$

(b) Information gain is defined as the entropy difference between the parent and child. If we are to split on a_1 then,



$$Entropy(Left) = -\frac{3}{4} log_2 \frac{3}{4} - \frac{1}{4} log_2 \frac{1}{4} = 0.81127$$
$$Entropy(Right) = -\frac{1}{5} log_2 \frac{1}{5} - \frac{4}{5} log_2 \frac{4}{5} = 0.7218$$
$$Entropy(Total) = weightage\_left \times Entropy(Left) + weightage\_right \times Entropy(Right)$$
$$Entropy(Total) = \frac{4}{9} \times 0.81127 + \frac{5}{9} \times 0.7218 = 0.7615$$
$$Information\ Gain = 0.9909 - 0.7615 = 0.2294$$

Splitting along a_2,

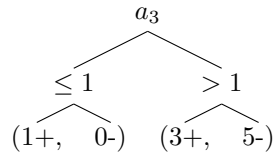


$$Entropy(Left) = -\frac{2}{5} log_2 \frac{2}{5} - \frac{3}{5} log_2 \frac{3}{5} = 0.9708$$

$$Entropy(Right) = -\frac{2}{4}log_2\frac{2}{4} - \frac{2}{4}log_2\frac{2}{4} = 1$$
$$Entropy(Total) = weightage\_left \times Entropy(Left) + weightage\_right \times Entropy(Right)$$
$$Entropy(Total) = \frac{5}{9} \times 0.9708 + \frac{4}{9} \times 1 = 0.9837$$
$$Information\ Gain = 0.9909 - 0.9837 = 0.0072$$

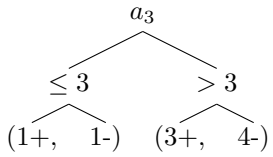
(c) For attribute a_3, which is a continuous variable, we split in multiple ways:

CASE 1:



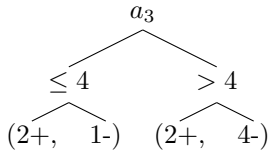
$$Entropy(Left) = -1log_2 1 = 0$$
$$Entropy(Right) = -\frac{3}{8}log_2\frac{3}{8} - \frac{5}{8}log_2\frac{5}{8} = 0.95$$
$$Entropy(Total) = weightage\_left \times Entropy(Left) + weightage\_right \times Entropy(Right)$$
$$Entropy(Total) = \frac{1}{9} \times 0 + \frac{8}{9} \times 0.95 = 0.8444$$
$$Information\ Gain = 0.9909 - 0.8444 = 0.1465$$

CASE 2:



$$Entropy(Left) = -\frac{1}{2}log_2\frac{1}{2} - \frac{1}{2}log_2\frac{1}{2} = 1$$
$$Entropy(Right) = -\frac{3}{7}log_2\frac{3}{7} - \frac{4}{7}log_2\frac{4}{7} = 0.985$$
$$Entropy(Total) = weightage\_left \times Entropy(Left) + weightage\_right \times Entropy(Right)$$
$$Entropy(Total) = \frac{2}{9} \times 1 + \frac{7}{9} \times 0.985 = 0.9885$$
$$Information\ Gain = 0.9909 - 0.9885 = 0.0024$$

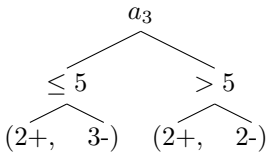
CASE 3:



$$Entropy(Left) = -\frac{2}{3}log_2\frac{2}{3} - \frac{1}{3}log_2\frac{1}{3} = 0.9183$$
$$Entropy(Right) = -\frac{2}{6}log_2\frac{2}{6} - \frac{4}{6}log_2\frac{4}{6} = 0.9183$$
$$Entropy(Total) = weightage\_left \times Entropy(Left) + weightage\_right \times Entropy(Right)$$

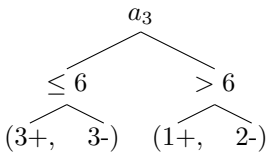
$$Entropy(Total) = \frac{3}{9} \times 0.9183 + \frac{6}{9} \times 0.9183 = 0.9183$$
$$Information\ Gain = 0.9909 - 0.9885 = 0.0726$$

CASE 4:



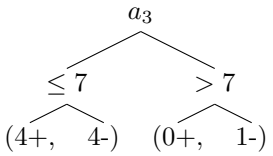
$$Entropy(Left) = -\frac{2}{5}log_2\frac{2}{5} - \frac{3}{5}log_2\frac{3}{5} = 0.9709$$
$$Entropy(Right) = -\frac{2}{4}log_2\frac{2}{4} - \frac{2}{4}log_2\frac{2}{4} = 1$$
$$Entropy(Total) = weightage\_left \times Entropy(Left) + weightage\_right \times Entropy(Right)$$
$$Entropy(Total) = \frac{5}{9} \times 0.9709 + \frac{4}{9} \times 1 = 0.9836$$
$$Information\ Gain = 0.9909 - 0.9836 = 0.0071$$

CASE 5:



$$Entropy(Left) = -\frac{3}{6}log_2\frac{3}{6} - \frac{3}{6}log_2\frac{3}{6} = 1$$
$$Entropy(Right) = -\frac{1}{3}log_2\frac{1}{3} - \frac{2}{3}log_2\frac{2}{3} = 0.9813$$
$$Entropy(Total) = weightage\_left \times Entropy(Left) + weightage\_right \times Entropy(Right)$$
$$Entropy(Total) = \frac{6}{9} \times 1 + \frac{3}{9} \times 0.9813 = 0.9727$$
$$Information\ Gain = 0.9909 - 0.9727 = 0.1813$$

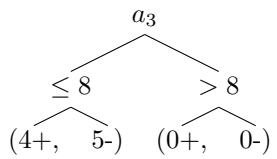
CASE 6:



$$Entropy(Left) = -\frac{4}{8}log_2\frac{4}{8} - \frac{4}{8}log_2\frac{4}{8} = 1$$
$$Entropy(Right) = -\frac{0}{1}log_2\frac{0}{1} - \frac{1}{1}log_2\frac{1}{1} = 0$$
$$Entropy(Total) = weightage\_left \times Entropy(Left) + weightage\_right \times Entropy(Right)$$
$$Entropy(Total) = \frac{8}{9} \times 1 + \frac{1}{9} \times 0 = 0.8889$$

$Information\ Gain = 0.9909 - 0.8889 = 0.102$

CASE 7:



This split has no information gain as all instances fall on one side of the tree.

(d) Based on information gain, a_1 would give the best split as it's information gain is the highest.

(e) The classification error before splitting is:

$$E_{original} = 1 - max(\frac{4}{9}, \frac{5}{9}) = \frac{4}{9}$$

Now, we split over a_1:

Table 6: Splitting Over a_1

Class	True	False
+	3	1
-	1	4

$$a_{1_T} = 1 - max(\frac{3}{4}, \frac{1}{4}) = \frac{1}{4}$$
$$a_{1_F} = 1 - max(\frac{1}{5}, \frac{3}{5}) = \frac{1}{5}$$
$$total = \frac{4}{9} \cdot \frac{1}{4} + \frac{5}{9} \cdot \frac{1}{5} = \frac{2}{9}$$

Now, we split over a_2:

Table 7: Splitting Over a_2

Class	True	False
+	2	2
-	3	2

$$a_{1_T} = 1 - max(\frac{2}{5}, \frac{3}{5}) = \frac{2}{5}$$
$$a_{1_F} = 1 - max(\frac{2}{4}, \frac{2}{4}) = \frac{2}{4}$$
$$total = \frac{5}{9} \cdot \frac{2}{5} + \frac{4}{9} \cdot \frac{2}{4} = \frac{4}{9}$$

a_1 would give a better split as its classification error is lower.

(f) Gini Index for attribute a_1:

$$Gini = \frac{4}{9}[1 - (\frac{3}{4})^2 - (\frac{1}{4})^2] + \frac{5}{9}[1 - (\frac{1}{5})^2 - (\frac{4}{5})^2] = 0.3444$$

Gini Index for attribute a_2:

$$Gini = \frac{5}{9}[1 - (\frac{2}{5})^2 - (\frac{3}{5})^2] + \frac{4}{9}[1 - (\frac{2}{4})^2 - (\frac{2}{4})^2] = 0.4899$$

As a_1 has lower gini impurity, it would produce a better split.