CH1

1. 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.
Answer: No, this does not involve finding any patterns. It is a simple task to separate customers into two groups.

(b) Dividing the customers of a company according to their profitability.
Answer: No, as only finding such customers isn't going to increase the profitability of the company.

(c) Computing the total sales of a company.
Answer: No, it is a simple mathematics calculation.

(d) Sorting a student database based on student identification numbers.
Answer: No, it is a database query rather than mining task.

(e) Predicting the outcomes of tossing a (fair) pair of dice.
Answer: No, as it doesn't include prediction based on any previous outcomes.

(f) Predicting the future stock price of a company using historical records.
Answer: Yes, historical records can be used to mine for useful data.

(g) Monitoring the heart rate of a patient for abnormalities.
Answer: Yes, using the continuous values any sudden changes can be noticed.

(h) Monitoring seismic waves for earthquake activities.
Answer: Yes, any unusual activity can be tracked and can be helpful to determine quake prone areas.

(i) Extracting the frequencies of a sound wave.
Answer: No, as only extracting frequency is not a data mining task.

CH2

2. 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.
Answer: Binary, Qualitative (Ordinal).

b) Brightness as measured by a light meter.
Answer: Continuous, Quantitative (Ratio)

c) Brightness as measured by people's judgments.
Answer: Discrete, Qualitative (Ordinal).

d) Angles as measured in degrees between 0 and 360.
Answer: Continuous, Quantitative (Ratio).

e) Bronze, Silver, and Gold medals as awarded at the Olympics.
Answer: Discrete, Qualitative (Ordinal)

f) Height above sea level.
Answer: Continuous, Quantitative (Ratio).

g) Number of patients in a hospital.
Answer: Discrete, Quantitative (Ratio).

h) ISBN numbers for books. (Look up the format on the Web.)
Answer: Discrete, Qualitative (Nominal).

i) Ability to pass light in terms of the following values: opaque, translucent, transparent.
Answer: Discrete, Qualitative (Ordinal)

j) Military rank.
Answer: Discrete, Qualitative (Ordinal).

k) Distance from the centre of campus.
Answer: Continuous, Quantitative (Interval).

l) Density of a substance in grams per cubic centimetre.
Answer: Discrete, 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.)

Answer: Discrete, Qualitative (Nominal).

7) Which of the following quantities is likely to show more temporal autocorrelation: daily rainfall or daily temperature? Why?

As temporal autocorrelation deals with the location of the objects in close proximity, daily temperature measures might be quite similar than the daily rainfall. Daily temperature tends to vary less than rainfall.

18) This exercise compares and contrasts some similarity and distance measures.
(a) For binary data, the Ll distance corresponds to the Hamming distance; that is, the number of bits that are different between two binary vectors. The Jaccard similarity is a measure of the similarity between two binary vectors. Compute the Hamming distance and the Jaccard similarity between the following two binary vectors.
x: 0101010001
y : 0100011000
Answer:
Hamming distance is the difference in bits between two vectors. Therefore, the hamming distance for the above problem is 3.
Jaccard's similarity can be calculated by using the given formula –
J= Number of matching presences/Number of attributes not involved in 00 matches
J= f11/f01+f10+f11

Where,
f01 =1, number of attributes where x=0,y=1
f10 =2, number of attributes where x=1,y=0
f00 =5, number of attributes where x=0,y=0
f11 =2, number of attributes where x=1,y=1

J=2/2+1+2=2/5=0.4

(b) Which approach, Jaccard or Hamming distance, is more similar to the Simple Matching Coefficient, and which approach is more similar to the cosine measure? Explain. (Note: The Hamming measure is a distance, while the other three measures are similarities, but don't let this confuse you.)

Answer: Hamming distance is more similar to SMC.

Simple Matching Coefficient is defined as SMC = Number of matching attribute values/ Number of values.

Where, matching values corresponds to bits equal in both the vectors which is nothing but hamming distance.

(c) Suppose that you are comparing how similar two organisms of different species are in terms of the number of genes they share. Describe which measure, Hamming or Jaccard, you think would be more appropriate for comparing the genetic makeup of two organisms. Explain. (Assume that each animal is represented as a binary vector, where each attribute is 1 if a particular gene is present in the organism and 0 otherwise.)

Answer: Jaccard's is mostly use to handle objects of asymmetric binary attributes. As, two species differ in genes it would be appropriate to use Jaccard to compare them.

(d) If you wanted to compare the genetic makeup of two organisms of the same species, e.g., two human beings, would you use the Hamming distance, the Jaccard coefficient, or a different measure of similarity or distance? Explain. (Note that two human beings share > 99.9% of the same genes.)

Answer: It would be easy to compare values using the hamming distance as only few genes would differ and difference can be evaluated.

19) For the following vectors, x and y, calculate the indicated similarity or distance measures.

(a) x : (1, 1, 1, 1), y : (2,2,2,2) cosine, correlation, Euclidean

Answer:

Cosine(x,y) = d1.d2/||d1||.||d2||

= 2+2+2+2/$\sqrt{4}+\sqrt{16}$=8/8=1

Correlation (x,y)=co-variance(x,y)/$S_x.S_y$

Covariance =1/n-1$\sum$(x-$\bar{x}$)(y-$\bar{y}$)

$\bar{x}$= (1+1+1+1)/4=1

$\bar{y}$= (2+2+2+2)/4=2

$S_{x=}$ 1

$S_{y=}$ 4

Co-var(x,y)=0

Therefore, Correlation(x,y)=0

Euclidean(x,y)=√1+1+1+1=2

(b) x : (0, 1,0, 1), y : (1,0, 1,0) cosine, correlation, Euclidean, Jaccard

Answer:

Cosine(x,y) = d1.d2/||d1||.||d2||

= 0+0+0+0/√2+√2=0

Correlation (x,y)= Correlation (x,y)=co-variance(x,y)/$S_x.S_y$

Covariance =1/n-1∑(x-$\overline{x}$)(y-$\overline{y}$)

$\overline{x}$= (0+1+0+1)/4=0.5

$\overline{y}$= (0+1+0+1)/4=0.5

Co-var(x,y)=-1

Therefore, Correlation(x,y)=0

Euclidean(x,y)=√1+1+1+1=2

Jaccard(x,y)= f11/f01+f10+f11

= 0/2+2

=0

(c) x: (0,- 1,0, 1) , y: (1,0,- 1,0) cosine,correlation,Eucl idean

Answer:

Cosine(x,y) = d1.d2/||d1||.||d2||

= 2+2+2+2/√4+√16=8/8=1

Correlation (x,y)=0

Euclidean(x,y)=√1+1+1+1=2

(d) x : (1,1 ,0,1 ,0,1) , y : (1,1 ,1 ,0,0,1) cosine,correlation ,Jaccard

Answer:

Cosine(x,y) = d1.d2/||d1||.||d2||

= 1+1+0+0+0+0+1/√4+√4=3/4=0.75

Correlation (x,y)=Covariance =1/n-1∑(x-x)(y-y)

$\overline{x}$= (1+1+1+1)/6=2/3

$\overline{y}$= (1+1+1+1)/6=2/3

Co-var(x,y)=0.75

Euclidean(x,y)=√1+1+1+1=2

Jaccard(x,y)= f11/f01+f10+f11

= 3/1+1+3

=0.6

(e) x : (2, -1,0,2,0, -3) , y : (-1, 1,- 1,0,0, -1) cosine,correlation

Answer:

Cosine(x,y) = d1.d2/||d1||.||d2||

= -2-1+3/√18+√4=0

Correlation (x,y)=0

CH3

8) Describe how a box plot can give information about whether the value of an attribute is symmetrically distributed. What can you say about the symmetry of the distributions of the attributes shown in Figure 3.11?

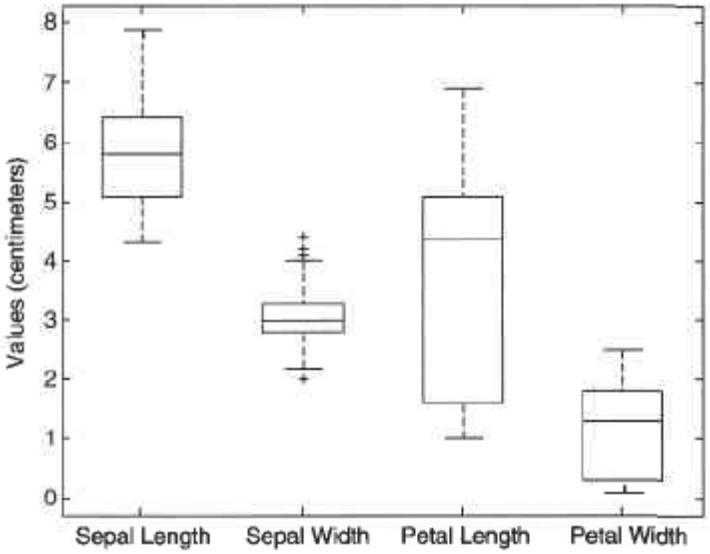

Figure 3.11. Box plot for Iris attributes.

Answer: According to the definition of symmetrically distributed data, if the median line is in the middle of the box then the data shows symmetry. As per the above diagram, sepal width and length are much more symmetrically distributed than the overall distribution of petal length and width. Petal length and width are skewed in nature.

CH4

2) Consider the training examples shown in Table 4.7 for a binary classification problem.

Table 4.7. Data set for Exercise 2.

Customer ID	Gender	Car Type	Shirt Size	Class
1	M	Family	Small	C0
2	M	Sports	Medium	C0
3	M	Sports	Medium	C0
4	M	Sports	Large	C0
5	M	Sports	Extra Large	C0
6	M	Sports	Extra Large	C0
7	F	Sports	Small	C0
8	F	Sports	Small	C0
9	F	Sports	Medium	C0
10	F	Luxury	Large	C0
11	M	Family	Large	C1
12	M	Family	Extra Large	C1
13	M	Family	Medium	C1
14	M	Luxury	Extra Large	C1
15	F	Luxury	Small	C1
16	F	Luxury	Small	C1
17	F	Luxury	Medium	C1
18	F	Luxury	Medium	C1
19	F	Luxury	Medium	C1
20	F	Luxury	Large	C1

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

Answer: $GINI(t) = 1 - \sum_j [p(j \mid t)]^2$

$= 1 - [(10/20)^2 + (10/20)^2] = 1 - [1/4 + 1/4]$

$= 1/2$

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

Answer: $GINI(t) = 1 - \sum_j [p(j \mid t)]^2$

$= 1 - [(0/1)^2 + (1/1)^2] = 1 - [0 + 1]$

$= 0$

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

Answer: Female -

$GINI(t) = 1 - \sum_j [p(j \mid t)]^2$

$= 1 - [(6/20)^2 + (4/20)^2] = 1 - 0.52$

=0.48

Male - $GINI(t) = 1 - \sum_j [p(j \mid t)]^2$

$= 1 - [(6/20)^2 + (4/20)^2] = 1 - 0.52$

$= 0.48$

Weighted Average = [(10/20)*Female]+[(10/20)*Male]=0.48

(d) Compute the Gini index for the Car Type attribute using multi-way split.

Answer: Family -

$GINI(t) = 1 - \sum_j [p(j \mid t)]^2$

$= 1 - [(1/4)^2 + (3/4)^2] = 1 - 0.625$

$= 0.375$

Luxury - $GINI(t) = 1 - \sum_j [p(j \mid t)]^2$

$= 1 - [(1/8)^2 + (7/8)^2] = 1 - 0.7812$

$= 0.218$

Sports - $GINI(t) = 1 - \sum_j [p(j \mid t)]^2$

$= 1 - [(8/8)^2 + (0/8)^2] = 1 - 1$

$= 0$

Weighted Average =[(4/20)*Family]+[(8/20)*Luxury]+[(8/20)*Sports]=0.163

(e) Compute the Gini index for the Shirt Size attribute using multi-way split.

Answer: Small -

$GINI(t) = 1 - \sum_j [p(j \mid t)]^2$

$= 1 - [(3/5)^2 + (2/5)^2] = 1 - 0.52$

$= 0.48$

Medium - $GINI(t) = 1 - \sum_j [p(j \mid t)]^2$

$= 1 - [(3/7)^2 + (4/7)^2] = 1 - 0.51$

$= 0.49$

Large - $GINI(t) = 1 - \sum_j [p(j \mid t)]^2$

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

$= 0.5$

Extra Large - $GINI(t) = 1 - \sum_j [p(j \mid t)]^2$

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

$= 0.5$

Weighted Average = [(5/20)*Small]+[(7/20)*Medium]+[(4/20)*Large] +[(4/20)*Extra Large]=0.4915

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

Answer: When comparing Gender, Car Type, and Shirt Size using the Gini Index, Car Type would be the better attribute. The Gini Index takes into consideration the distribution of the sample with zero reflecting the most

distributed sample set. Out of the three listed attributes, Car Type has the lowest Gini Index.

(g) Explain why Customer ID should not be used as the attribute test condition even though it has the lowest Gini.
Answer: Customer ID should not be used as the attribute test condition because each attribute is unique.

3) Consider the training examples shown in Table 4.8 for a binary classification problem.

Table 4.8. Data set for Exercise 3.

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) What is the entropy of this collection of training examples with respect to the positive class?
Answer:
$Entropy(t)=-\sum_j p(j \mid t)\log p(j \mid t)$
$=-[(4/9)*\log_2(4/9)+ (5/9)*\log_2(5/9)]$
$=0.99107$
(b) What are the information gains of a1 and a2 relative to these training examples?
Answer:
For attribute a1, the corresponding counts and probabilities are:

a_1	+	-
T	3	1
F	1	4

The entropy for a_1 is:
$4/9\ [-(3/4)\log_2(3/4)-(1/4)\log_2(1/4)] + 5/9[-(1/5)\log_2(1/5)-(4/5)\log_2(4/5)]$
$=0.7616$.

Therefore, the information gain for a_1 is $0.9911-0.7616 = 0.2294$. For attribute a1, the corresponding counts and probabilities are:

a_2	+	-
T	2	3
F	2	2

The entropy for a_2 is:
$5/9\ [-(2/5)\log_2(2/5)-(3/5)\log_2(3/5)] + 4/9[-(2/4)\log_2(2/4)-(2/4)\log_2(2/4)]$
$=0.9839$
Therefore, the information gain for a_2 is $0.9911-0.9839 = 0.0072$.

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

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

a_3	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 5.0	- -	5.5	0.9839	0.0072
6.0	+	6.5	0.9728	0.0183
7.0 7.0	+ -	7.5	0.8889	0.1022

(d) What is the best split (among *a*1, *a*2, and *a*3) according to the information gain?
Answer:

According to information gain, $a1$ produces the best split due to its higher gain in comparison to a1 and a2.

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

Answer:

For attribute $a1$: error rate = 2/9=0.22.
For attribute $a2$: error rate = 4/9=0.44.
According to the classification error rate, the best split is a1 due to a lower classification error in comparison to a2. Also we know that classification error shows the accuracy of the sample set and thus higher the classification error the more error the sample set contains.

(f) What is the best split (between $a1$ and $a2$) according to the Gini index?

Answer:

For attribute $a1$, the gini index is
$4/9*[1-(3/4)^2 – (1/4)^2] + 5/9*[1 − (1/5)^2 − (4/5)^2] = 0.3444.$

For attribute $a2$, the gini index is
$5/9*[1 − (2/5)^2 − (3/5)^2] + 4/9[1 − (2/4)^2 − (2/4)^2] = 0.4889.$

Since the gini index for $a1$ is smaller, it produces the better split.