BANARAS HINDU UNIVERSITY

VARANASI, (U.P.)

NAME – Vijay Gwala

class – MCA 4th Sem.

roll No. – 18419MCA053

Enroll. No. – 409244

class Roll No. – 47

Email ID – vijaygwala97@gmail.Com

whatsupp No. – 7000054532

1) 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.

Ans: Datamining is the process of discovering interesting Knowledge from large amount of data stored either in databases, data warehouses or other information repositeries.

1) clustering: is the process of grouping a set of physical or abstract objects into classes of similer objects. the objects are grouped based on the principles of increasing intraclass similarity and decreasing iterclass similarity. In the Context of Search engine, clustering Can help to display the result that not only contain the keyword specified in the "Search" box but also related results.

eg. on entering keyword 'data mining tutorials' in the Search box, the search engine not display only 1 resource which contain keyword although it display all the releted resources contain particular keyword.

2) classification: it is a process of finding a set of releted functions that describe and distinguish data classes or concepts related to objects.

eg.: where the list of research papers associated with keyword could be provided by the search engines. This is done by either classification rules or decision tree.

3) Association Rule mining: This method of data mining is used to discover patterns within the input and the data base Creating a strong link that associates the two variables This type of data mining can take a link of words i.e. a sentence or short phrase, and Compare it to previous searchs that have been performed in the past.

eg.:- If one person was in search for the famous words like 'data structure' then when the next user types just a portion of a phase then search engine convey that there are 'data structure & algorithms' in the same point of intrest.

4) Anomaly detection:- Anomaly detection is when an input Variable is very dissimilar from other variables (or events) contained in database. this is a helpful tool to insure that only pertinent information is included in search results.

eg:- Anomaly detection can be useful in the first pharmacology example to insure that the only information relayed to the user is about related drugs rather than drugs associated with treating correlated symptoms.

2) 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.

→ No, this is a simple database query.

(b) Dividing the Customers of a Company according to their profitability.

→ No, this is an accounting Calculation, followed by the application of threshold. However, predicting the profitability of a new Customer would be data mining.

(c) Computing the total Sells of Company.

→ No, this is a simple accounting.

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

→ No, this is a simple database query.

(e) predicting the outcomes of tossing a (fair) pair of dice.

→ No, since the die is fair, this is a probability Calculation. if the die were not fair, and we needed to estimate the probabilities of each outcome from the data then this is more like the problems considered by datamining. However in this specific Case solution to this problem were develped by mathematicians a long time ago. and thus we wouldn't considered to be data mining.

(f) predicting the feature stock prices of a Company using historical records.

→ Yes, we would attempt to create a model that can predict the Continuous value of stock price. this is an example of the area of data mining known as predictive modeling. we Could use regression for this modeling.

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

→ yes, we would build a model of the normal behaviour of heart rate and raise an alarm when an unusual heart behaviour occured. This would involve the area of data mining known as anomaly detection.

(h) monitering seismic waves for earthquake activities.

→ yes, In this case, we would build a model of different types of seismic wave behaviour associated with earthquake activities and raise an alarm when one of these different types of seismic activity was observed. this is an example of the area of data mining known as classification.

(i) Extracting the frequencies of a sound wave.

→ No, this is signal processing.

3) 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.

Cosine:

x · y = 1*2 + 1*2 + 1*2 + 1*2 = 8

||x|| = sqrt(1*1 + 1*1 + 1*1 + 1*1) = sqrt(4) = 2

||y|| = sqrt(2*2 + 2*2 + 2*2 + 2*2) = sqrt(16) = 4

Cos(x,y) = (x * y) / (||x|| * ||y||) = (8)/(2*4)

cos(x,y) = 1

Correlation:

Corr(x,y) = [Covariance (x,y)] / [Standard deviation(x) * Standard deviation(y)]

Mean of x = (1+1+1+1)/4 = 1

mean of y = (2+2+2+2)/4 = 2

Covariance (x,y) = 1/(4-1) [(1-1) (2-2) + (1-1) (2-2) + (1-1)(2-2) + (1-1) (2-2)] = 0

Standard deviation(x) = sqrt[((1/(4-1)) * {(1-1)^2 + (1-1)^2 + (1-1)^2
+ (1-1)^2 }] = sqrt[(1/3) * 0] = 0

Standard deviation(y) = sqrt[((1/(4-1)) * { (2-2)^2 + (2-2)^2 + (2-2)^2
+ (2-2)^2 }] = sqrt[(1/3) * 0] = 0

Corr (x,y) = 0/0 = undefined

Euclidean:

d(x,y) = sqrt((1-2)^2 + (1-2)^2 + (1-2)^2 + (1-2)^2)

Euclidean distance = 2

Scanned with CamScanner

(b) $x = (0,1,0,1)$, $Y = (1,0,1,0)$ Cosine, Correlation, Euclidean, Jaccard

Cosine :

$x \cdot y = 0*1 + 1*0 + 0*1 + 1*0 = 0$

$||x|| = sqrt(0*0 + 1*1 + 0*0 + 1*1) = sqrt(2)$

$||y|| = sqrt(1*1 + 0*0 + 1*1 + 0*0) = sqrt(2)$

$cos(x,y) = (x \cdot y)/(||x|| * ||y||) = (0)/(sqrt(2) * sqrt(2))$

$cos(x,y) = 0$

Correlation :

Corr(x,y) = [Covariance (x,y)] / [standard deviation (x) * standard deviation(y)]

Mean of X = (0+1+0+1)/4 = 1/2 = 0.5

Mean of Y = (1+0+1+0)/4 = 1/2 = 0.5

Covariance (x,y) = 1/(4-1) * [(0-1/2)(1-1/2) + (1-1/2)(0-1/2) + (0-1/2)(1-1/2)
+ (1-1/2)(0-1/2)]

Covariance (x,y) = (1/3) * [(-1/4) + (-1/4) + (-1/4) + (-1/4)]

Covariance (x,y) = -1/3

standard_deviation (x) = sqrt[((1/(4-1)) * {(1-1/2)^2 + (0-1/2)^2
+ (1-1/2)^2 + (0-1/2)^2}] = sqrt[(1/3) * 1]

= 0.57735

Similarly standard deviation (y) = 0.57735

Corr(x,y) = (-1/3) / (0.57735 * 0.57735)

Corr(x,y) = -1

Euclidean :

d(x,y) = sqrt((0-1)^2 + (1-0)^2 + (0-1)^2 + (1-0)^2)

euclidean distance = 2

Jaccard :

J = (number of matching presence) / (number of attributes not involved in 00 matches)

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

f_{01} = 2 the number of attributes where X was 0 and Y was 1

f_{10} = 2 the number of attributes where X was 1 and Y was 0

f_{00} = 0 the number of attributes where X was 0 and y was 0

f_{11} = 0 the number of attributes where x was 1 and y was 1

$J = (0)/(2+2+0)$

$J = 0$

(c) $x = (0,-1, 0,1)$, $y = (1,0,-1,0)$ Cosine , Correlation, Euclidean.

Cosine:

$x \cdot y = 0*1 + (-1)*0 + 0*(-1) + 1*0 = 0$

$||x|| = sqrt(0*0 + (-1)*(-1) + 0*0 + 1*1) = sqrt(2)$

$||y|| = sqrt(1*1 + 0*0 + (-1)*(-1) + 0*0) = sqrt(2)$

$Cos(x,y) = (x \cdot y) / (||x|| * ||y||) = (0)/(sqrt(2) * sqrt(2))$

$Cos(x,y) = 0$

Correlation:

corr(x,y) = [Covariance (x,y)] / [standard deviation (x) * standard deviation (y)]

Mean of x = (0+ (-1) + 0+1) / 4 = 0

Mean of y = (1+0+(-1)+0)/4 = 0

covariance (x,y) = 1/(4-1) * [(0-0) (1-0) + (-1-0) (0-0) + (0-0)(-1-0) + (1-0) (0-0)] = (1/3) * 0 = 0

corr(x,y) = 0

Euclidean:

d(x,y) = sqrt ((0-1)^2 + (-1-0)^2 + (0+1) ^2 + (1-0)^2)

Euclidean distance = 2

(d) x = (1,1,0,1,0,1) , y = (1,1,1,0,0,1) Cosine, Correlation, jaccard

Cosine:

x · y = 1*1 + 1*1 + 0*1 + 0*0 + 1*1 = 3

||x|| = Sqrt (1.*1 + 1*1 + 0*0 + 1*1 + 0*0 + 1*1) = 2

||y|| = sqrt (1*1 + 1*1 + 1*1 + 0*0 + 0*0 + 1*1) = 2

Cos (x,y) = (x · y)/ (||x|| * || y||) = (3) / (2*2)

cos (x,y) = 3/4 = 0.75

Correlation:

Corr (x,y) = [Covariance (x,y)] / [standard deviation (x) * s. deviation (y)]

Mean of x = (1+1+0+1+0+1) / 6 = 4/6

Mean of y = (1+1+1+0+0+1) / 6 = 4/6

Covariance (x,y) = 1 / (6-1) * [(1 - 4/6) (1-4/6) + (1-4/6) (1-4/6)
+ (0 - 4/6) (1-4/6) + (1-4/6) (0,-4/6) + (0-4/6)
(0-4/6) + (1-4/6)] = (1/5) (1/3) = 1/15

Standard deviation (x) = Sqrt [(1 / (6-1)) * { (1-4/6) ^2 +
(1-4/6) ^2 + (1-4/6)^2 + (0-4/6) ^2
+ (0-4/6)^2 + (1-4/6)^2 }]

= sqrt [(1/5) * (4/3)] = 0.5164

Standard deviation (y) = 0.5164

corr (x,y) = (1/15) / (0.5164 * 0.5164)

corr (x,y) = 0.25

Jaccard :

J = (number of matching presence)/ (number of attributes not involved in 00 matches)

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

$f_{01} = 1$ the number of attributes where x was 0 and y was 1

$f_{10} = 1$ the number of attributes where x was 1 and y was 0

$f_{00} = 1$ the number of attributes where x was 0 and y was 0

$f_{11} = 3$ the number of attributes where x was 1 and y was 1

J = (3) / (1+1+3)

J = 3/5 = 0.6

(e) x = (2, -1, 0, 2, 0, -3), y = (-1, 1, -1, 0, 0, -1) Cosine, Correlation.

Cosine :

x • y = 2 * (-1) + (-1) * 1 + 0 * (-1) + 2 * 0 + 0 * 0 + (-3) * (-1) = 0

||x|| = sqrt(2*2 + (-1) * (-1) + 0 * 0 + 2 * 2 + 0 * 0 + (-3) * (-3) = sqrt(18)

||y|| = sqrt((-1) * (-1) + 1 * 1 + (-1) * (-1) + 0 * 0 + 0 * 0 + (-1) * (-1)) = 2

Cos(x, y) = (x • y)/ (||x|| * || y||) = (0) / (sqrt(18) * 2)

Cos(x, y) = 0

Correlation :

Corr(x, y) = [Covariance (x, y)] / [Standard deviation (x) * Standard deviation (y)]

mean of x = (2 + (-1) + 0 + 2 + 0 + (-3))/6 = 0

mean of y = ((-1) + 1 + (-1) + 0 + 0 + (-1)) / 6 = -1/6

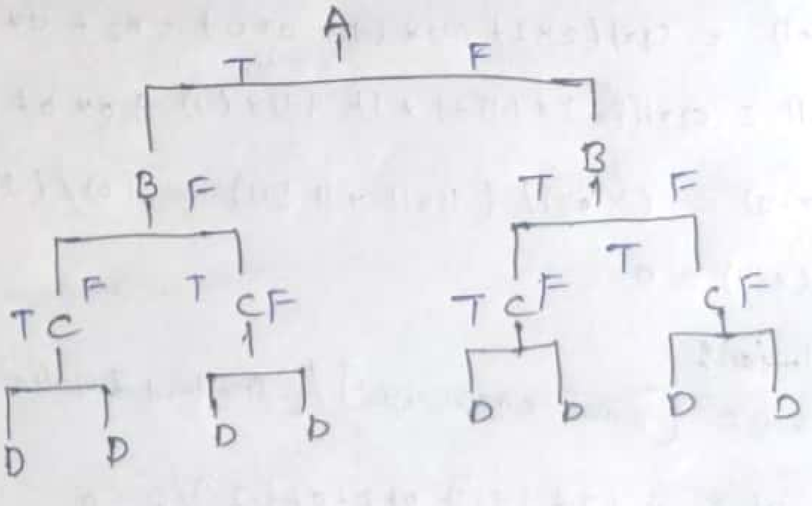
Covariance (x, y) = 1/(6-1) * [(2-0)(-1+1/6) + (-1-0)(1+1/6) + (0-0)
(1+1/6) + (2-0)(0+1/6) + (0-0)(0+1/6)
+ (-3-0)(-1+1/6)] = (1/5) * 0 = 0

Corr (x, y) = 0

4) Draw the full decision tree for the parity function of four Boolean attributes A, B, C. and D

Parity function:

A	B	C	D	class
F	F	F	F	T
F	F	F	T	F
F	F	T	F	F
F	F	T	T	T
F	T	F	F	F
F	T	F	T	T
F	T	T	F	T
F	T	T	T	F
T	F	F	F	F
T	F	F	T	T
T	F	T	F	T
T	F	T	T	T
T	T	F	F	F
T	T	F	T	T
T	T	T	F	F
T	T	T	T	F
				T



The decision tree has a stem for both possible values T, F and it cannot be reduced any further.

5) Suppose that for a data set.

- there are m points and k clusters
- half of the points and clusters are in "more dense" regions,
- half the points and clusters are in "less dense" regions, and
- the two regions are well-separated from each other.

For the given dataset which of the following should occure in order to mini-mize the squared error when finding k clusters:

(a) Centroids should be equally distributed b/w more dense and less dense regions.

(b) more Centroids should be allocated to the less dense region.

(c) More Centroids should be allocated to the denser region.

Ans:- The correct answer of this question is (c)

bcoz less dense regions require more centroids if the square error is to be minimized.