CS422 Data Mining
HW 1
SHIQI LIU

1.1 Chapter 1 Exercises: 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.

No, this is a simple database query

b. Dividing the customers of a company according to their profitability.

no, this is a simple accounting calculation

c. Computing the total sales of a company.

No, simple accounting calculation

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

No, sorting is a simple work for database

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

No, because die is fair, we cannot create a model for this. However, tossing a pair of dice is a probability calculation.

F. Predicting the future stock price of a company using historical records.

Yes. We can create a model for predict the stock price with continuous value. This is an example of the area of data mining known as predictive modeling.

g. Monitoring the heart rate of a patient for abnormalities.

Yes. We can create a model of the normal behavior of heart rate and raise an alarm when unusual heart behavior occurred. This would involve the area of data mining known as anomaly detection.

h. Monitoring seismic waves for earthquake activities.

Yes. We can create a model of different types of seismic wave behavior 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. signal processing is not data mining.

1.2 Chapter 2 Exercises: 2,7,15,16,17,18,19

2.2Classify 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

1. Time in terms of AM or PM. Binary, qualitative, ordinal
2. Brightness as measured by a light meter. Continuous, quantitative, ratio
3. Brightness as measured by people's judgments. Discrete, qualitative, ordinal
4. Angles as measured in degrees between 0 and 360. Continuous, quantitative, ratio
5. Bronze, Silver, and Gold medals as awarded at the Olympics. Discrete, qualitative, ordinal
6. Height above sea level. Continuous, quantitative, ratio

7. Number of patients in a hospital. Discrete, quantitative, ratio
8. ISBN numbers for books. (Look up the format on the Web.) discrete, qualitative, nominal
9. Ability to pass light in terms of the following values: opaque, translucent, transparent. Discrete qualitative, ordinal
10. Military rank. Discrete, qualitative, ordinal
11. Distance from the center of campus continuous, quantitative, interval
12. Density of a substance in grams per cubic centimeter. Discrete, quantitative, ratio
13. 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) discrete, qualitative, nominal

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

Daily temperature.

Because rainfall can be very localized, for example, south of Chicago might rain but north of Chicago might not rain in a same day. However, there are similar temperature in close locations, that implies daily temperature usually remain the same for adjacent places.

2.15. You are given a set of m objects that is divided into k groups, where the i^{th} group is of size mi. If the goal is to obtain a sample of size n<m, what is the difference between the following two sampling schemes? (Assume sampling with replacement.)

a. We randomly select n×mi/m elements from each group.

b. We randomly select n elements from the data set, without regard for the group to which an object belongs.

The first scheme random selects by rate, it implies if the groups have same amount, the picked number will be constant. Therefore, you will get same number from each group.

For the second scheme, you will get a random number for each group therefore it implies the picked number of groups are varied.

Even though there are a random number of elements from every group, on average, in the second selection, there will be n*mi/m elements from every group.

2.16. Consider a document-term matrix, where tfij is the frequency of the ith word (term) in the jth document and m is the number of documents. Consider the variable transformation that is defined by

tfij′=tfij×log(m/dfi), (2.31)

where dfi is the number of documents in which the ith term appears, which is known as the **document frequency** of the term. This transformation is known as the **inverse document frequency** transformation.

1. What is the effect of this transformation if a term occurs in one document? In every document?

   Terms that occur in every document have 0 weight, while those that occur in one document have maximum weight, i.e., log m.

2. What might be the purpose of this transformation

   This normalization reflects the observation that terms that occur in every document do not have any power to distinguish one document from another, while those that are relatively rare do.

2.17. Assume that we apply a square root transformation to a ratio attribute *x* to obtain the new attribute x*. As part of your analysis, you identify an interval (*a*, *b*) in which x* has a linear relationship to another attribute *y*.

a. What is the corresponding interval (*A*, *B*) in terms of *x* ?

(a^2,b^2)

b. Give an equation that relates *y* to *x*.

$y=x^2$

2.18. This exercise compares and contrasts some similarity and distance measures.

a. For binary data, the L1 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

Hamming distance = # of different bits = 3

Jaccard Similarity = # of 1-1 matches /(number of bits - number 0-0 matches) = 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.)

Hamming is more similar to SMC as SMC=1-(Hamming Distance/# of bits)

Jaccard is similar to cosine measure as they both ignore 0-0(f_{00}) matches.

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.)

Jaccard Measure is more apt in comparing which genes the two organisms share as Jaccard accounts for the genes which both organisms have present in 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.)

To measure how similar two human beings are, Hamming Distance is the most apt as it points out the non-matching human genes. The lesser the Hamming Distance, the more similar the human beings.

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

cos(x,y) = cos((1,1,1,1),(2,2,2,2))=(**x*y**)/(|x|*|y|)=8/8=1

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) = √ [((1/(4-1))) * {(1-1)2+(1-1)2+(1-1)2+(1-1)2}] = √(1/3) * 0 = 0 Standard Deviation(y) = √ [((1/(4-1))) * {(2-2)2+(2-2)2+(2-2)2+(2-2)2}] = √(1/3) * 0 = 0

Corr(x,y) = covariance(x,y)/[standard deviation(x)*standard deviation(y)]=0/0=undefined

Euclidian Distance = √[(2-1)2+(2-1)2+(2-1)2+(2-1)2] = √4 = 2

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

Cos(x,y) = cos((0,1,0,1),(1,0,1,0))= (x.y)/(|x|.|y|) = (0*1+1*0+0*1+1*0)/(√2*√2) = 0/2 = 0

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

Mean of x = (0+1+0+1)/4 = 0.5

Mean of y = (1+0+1+0)/4 = 0.5

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

Standard Deviation(x) = √ [((1/(4-1)))*{(1-.5)2+(0-.5)2+(1-.5)2+(0-.5)2}]=√[(1/3)*1] =.5773 Standard Deviation(y) = √ [((1/(4-1)))*{(0-.5)2+(1-.5)2+(0-.5)2+(1-.5)2}]=√[(1/3)*1] = .5773

Corr(x,y) = (-1/3)/(0.5773)2 = (-1/3)/(1/3) = -1

Euclidean Distance = √[(0-1)2+(1-0)2+(0-1)2+(1-0)2]=√4=2

Jaccard (x,y) = f11/(f10+f01+f11) = 0/(2+2+0)=0/4=0

c. x=(0, −1, 0, 1), y=(1, 0, −1, 0) cosine, correlation, Euclidean

Cos(x,y) = cos((0,-1,0,1),(1,0,-1,0))=(x.y)/(|x|.|y|)=(0*1+-1*0+0*-1+1*0)/(√2*√2)=0/2=0
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 Distance = √[(0-1)2+(-1-0)2+(0-(-1))2+(1-0)2]=√4=2
d. x=(1, 1, 0, 1, 0, 1), y=(1, 1, 1, 0, 0, 1) cosine, correlation, Jaccard

Cos(x,y) = cos((1,1,0,1,0,1),(1,1,1,0,0,1))= (x.y)/(|x|.|y|)=(1*1+1*1+0*1+1*0+0*0+1*1)/(√4*√4) = 3/4

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

Mean of x = (1+1+0+1+0+1)/6 = 4/6 = 0.66666 Mean of y = (1+1+1+0+0+1)/6 = 4/6 = 0.66666

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)(1-4/6)] = (1/5)(1/3) = 1/15

Standard Deviation(x) = √[((1/(6-1))))*{(1-4/6)2+(1-4/6)2+(0-4/6)2+(1-4/6)2+(0-4/6)2+(1- 4/6)2}] = √[(1/5)*(4/3)]=.51639

Standard Deviation(y) = √[((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}] = √[(1/5)*(4/3)]=.51639

Corr(x,y)=(1/15)/(.51639)2=0.25

Jaccard (x,y) = f11/(f10+f01+f11) = 3/(1+1+3)=3/5=0.6

e. x=(2, −1, 0, 2, 0, −3), y=(−1, 1, −1, 0, 0, −1) cosine, correlation

Cos(x,y) = cos((2,-1,0,2,0,-3),(-1,1,-1,0,0,-1))= (x.y)/(|x|.|y|) = (2*-1+-1*1+0*-1+2*0+0*0+-3*-1)/(√18*√4) = 0/2√18=0

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 = -2/6 = -1/3

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

Corr(x,y) = 0