Abelardo Riojas
ISC 4245C Homework
Professor Meyer-Baese Fall 2020

Chapter 1:

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.

        Data mining often deals with complex, highly dimensional, or distrubed data using techniques like clustering, feature extraction, and dimensionality reduction to deduce new insights or information. This is just a separation of a list by one variable.

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

        This would involve a higher amount of complexity suitable for data mining. Profitability isn't directly calculable as it involves many different factors. This makes it a higher dimension problem.

    c) Computing the total sales of a company.

        While it does involve a large data set, you aren't really transforming it in a way that uses data mining techniques.

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

        Student databases aren't extraordinarily large, and sorting algorithms, while suitable for extracting information from that kind of set, aren't data mining techniques. Also the data wasn't transformed.

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

        This is something that you do on pen and paper using probability theory, not a big data mining algorithm.

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

        This would require a large data set (check), multidimensional problem (records are a variety of variables), and probably some dimensionality reduction techniques would need to be used to see what actually affects stock price.

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

        Hearts can beat anywhere from 50 to 100 times per minute, so for a given patient pumping blood for several hours, the dataset would be large. The data would probably have to be analyzed in a way that uses data mining techniques to extract abnormal features from the heart rate.

    h) Monitoring seismic waves for earthquake activities.

        This would be a large data set as the period of seismic waves on a large scale can take some time. Feature extraction would have to be implemented to deduce patterns in the seismic activity to produce an accurate prediction.

   i) Extracting the frequencies of a sound wave.

      Sound waves are inherently multidimensional as they have a time, frequency and volume dimension associated with them. Extracting the frequencies from a sound wave would involve dimension reduction, which is a data mining technique.

2) 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.

   For an Internet search engine company called, say, Boogle. Data mining could help the company improve its search engine by implementing document clustering. You can use frequently mentioned words or phrases in a document to perform similarity measurements on other documents and cluster.
   Search engine companies also use a tremendous amount of user-generated data such as click-speed, link visit duration, and previous history to improve their searches. This would be a highly dimensional data set that would require data mining to make insights about.

Chapter 2:

1) Classify the data.
   a) Time in terms of AM or PM.

      Binary, qualitative, nominal.

   b) Brightness as measured by a light meter.

      Quantitative, continuous, ratio.

   c) Brightness as measured by people's judgments.

      Qualitative, ordinal.

   d) Angles as measured in degrees between 0◦ and 360◦.

      Quantitative, interval, continuous.

   e) Bronze, Silver, and Gold medals as awarded at the Olympics.

      Qualitative, ordinal.

   f) Height above sea level.

      Quantitative, continuous, ratio.

   g) Number of patients in a hospital.

      Quantitative, discrete, ratio.

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

      Quantitative, discrete, interval.

i) Ability to pass light in terms of the following values: opaque, translucent, transparent.

Qualitative, ordinal.

j) Military rank.

Qualitative, ordinal.

k) Distance from the center of campus.

Quantitative, ratio, continuous.

l) Density of a substance in grams per cubic centimeter.

Quantitative, interval, continuous.

m) Coat check number.

Quantitative, interval, discrete.

2) An educational psychologist wants to use association analysis to analyze test results. The test consists of 100 questions with four possible answers each.

a) How would you convert this data into a form suitable for association analysis?

Association analysis requires asymmetric data, where only a non-zero value is regarded as important. The data would have to be converted to a class where only the presence of a particular answer for a question matters.

b) In particular, what type of attributes would you have and how many of them are there?

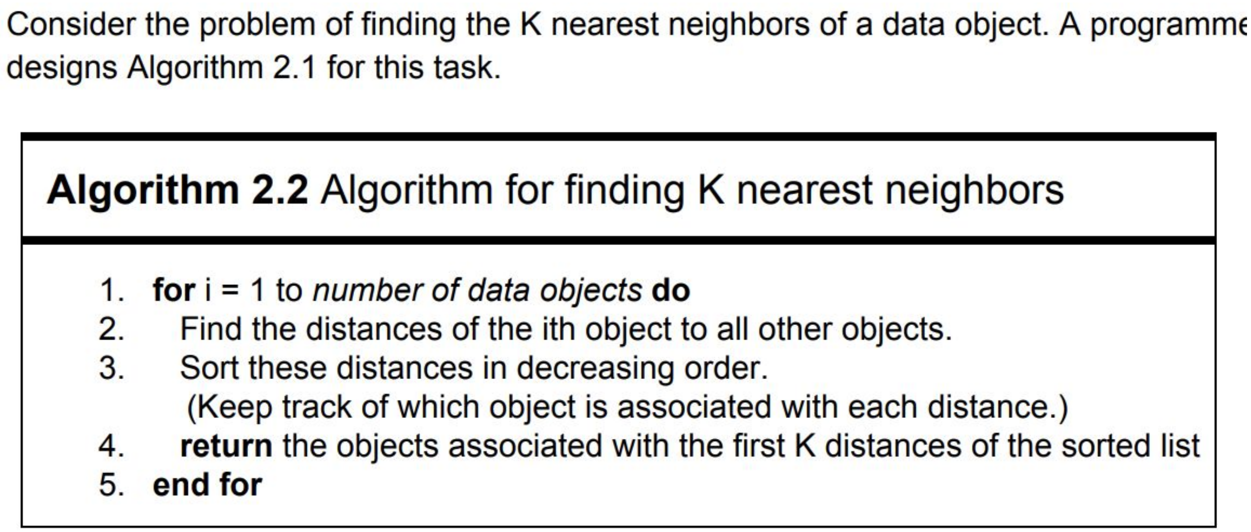
Four different attributes: A, B, C, and D answers.

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

Daily temperature since the daily rising and setting of the sun would produce a common pattern for each one-day period while daily rainfall is more random and harder to predict.

4) Give at least two advantages to working with data stored in text files instead of in a binary format.

Text files can be read by humans which makes understanding the data’s features and classes much easier. Binary text files have to be read by a computer first before the data can be looked at by a human. Binary files also cannot directly be read into an IDE like text files can.



Consider the problem of finding the K nearest neighbors of a data object. A programme designs Algorithm 2.1 for this task.

Algorithm 2.2 Algorithm for finding K nearest neighbors

```
for i = 1 to number of data objects do
    Find the distances of the ith object to all other objects.
    Sort these distances in decreasing order.
     (Keep track of which object is associated with each distance.)
    return the objects associated with the first K distances of the sorted list
end for
```

5)

a) Describe the potential problems with this algorithm if there are duplicate objects in the data set. Assume the distance function will only return a distance of 0 for objects that are the same.

If there is a duplicate object in the data set, then the algorithm will only return the duplicate object in each cluster since it returns the objects associated with the smallest distance in the shortest list. Ergo, the shortest distance would be 0, and only the duplicate at the cluster point would be in the cluster.

b) How would you fix this problem?

Instead of returning the objects associated with the smallest distance, I would return the objects by the smallest distance to each of the cluster points, find the centers, and then calculate the closest clusters again. Do this until the cluster points don't move by a certain tolerance.

6) 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: **x • y =** (1*2)+(1*2)+(1*2)+(1*2) = 8 /

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

8/(2*4) = 1

Vectors are multiples of each other (dependent) so cosine = 1.

Correlation: r(**x**,**y**) = Sum((xi-xbar) * (yi-ybar)) / stdev(**X**)*stdev(**Y**)

xbar = 1, ybar = 2 by inspection

Std.deviation for x and y = 0 by inspection

= 0*0/(0*0) = NaN = No correlation

Euclidean:
sqrt(sum(xi-yi)^2) = 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)
Cosine: **x • y =** (0*1)+(1*0)+(0*1)+(1*0) = 0 /
||**x**|| = sqrt(0+1+0+1) = sqrt(2) * ||**y**|| = sqrt(1+0+1+0) = sqrt(2)
= 0 (orthogonal vectors)
Correlation:
xbar = .5, ybar = .5
std(x) = .5774, std(y) = .5774
cov(x,y) = -.33333
= -1

Euclidean:
sqrt(sum(xi-yi)^2) = sqrt((0-1)^2 + (1-0)^2 + (0-1)^2 + (1-0)^2)) = 2

Jaccard = 1 - J(x,y) = 1 - (x intersect y/ |x| + |y| - |x intersect y|)
|x| = 2 |y| = 2
|x intersect y| =2
= 2/(2+2 -2) = 1
1-1= 0

c) x = (0, −1, 0, 1), y = (1, 0, −1, 0)
Cosine: **x • y =** (0*1)+(-1*0)+(0*-1)+(1*0) = 0 /
||**x**|| = sqrt(0+1+0+1) = sqrt(2) * ||**y**|| = sqrt(1+0+1+0) = sqrt(2)
= 0 (orthogonal again)

Correlation:
xbar = 0, ybar = 0
std(x) = 0.8165, std(y) = 0.8165
cov(x,y) = 0
= 0

Euclidean:
sqrt(sum(xi-yi)^2) = sqrt((0-1)^2 + (-1-0)^2 + (0+1)^2 + (1-0)^2)) = 2

d) x = (1, 1, 0, 1, 0, 1), y = (1, 1, 1, 0, 0, 1)
Cosine: **x • y =** (1*1)+(1*1)+(0*1)+(1*0) (0*0)+(1*1) = 3 /
||**x**|| = sqrt(1+1+0+1+0+1) = 2 * ||**y**|| = sqrt(1+1+1+0+0+1) = 2
= .75

Correlation:

xbar = 2/3, ybar = 2/3

std(x) = 0.5164, std(y) = 0.5164

cov(x,y) = 1/15

= ¼

Jaccard:

Both vectors share 100% of their unique values, by inspection the Jaccard distance is 0.

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

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

The vectors are orthogonal (dot product = 0) so the cosine is 0.

Correlation:

xbar = 0, ybar = -1/3

std(x) = 1.89736, std(y) = 0.8165

cov(x,y) = 0

= 0