Data Mining Assignment 2

1) Read Chapter 1 (all) and Chapter 2 (only sections 2.1, 2.2 and 2.3).  
  
2) Redo In Class Exercises #1 and #2, but use different examples from those which we used in class.

     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, although predicting the profitability of a new customer                             might belong to data mining. But this is an accounting calculation       which is followed by the application of threshold.

c) Computing the total sales of a company.

    No, this is 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, this is a probability calculation because the dice are fair.         Incase if the dices were not fair and if we need to estimate the         probabilities of each outcome then we can consider this as data

                           mining problem. Also for this specific problem as the solutions

                           were developed by mathematicians a long time ago we can not

    consider it as a data mining problem.

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

  Yes, this is an example of an area in data mining which is known as

  predictive modelling. We can also use regression for this modelling,               although researchers in many fields have developed a wide variety

  of techniques for predicting time series.

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

    Yes, this would be an example of an area in data mining which is               known as anomaly detection. This can also be considered as a           classification problem if we have samples of both normal and         abnormal heart behaviour.

h) Monitoring seismic waves for earthquake activities.

    Yes, this case is an example of an area of data mining which is         known as classification.

i) Extracting the frequencies of a sound wave.

  No, this is signal processing.

3. For each of the following data sets, explain whether or not data privacy is an important issue.

(a) Census data collected from 1900–1950.

      No

(b) IP addresses and visit times of Web users who visit your Website.

                Yes

(c) Images from Earth-orbiting satellites.

    No

(d) Names and addresses of people from the telephone book.

      No

(e) Names and email addresses collected from the Web.

      No

     Chapter 2 Exercises :

1. In the initial example of Chapter 2, the statistician says, “Yes, fields 2 and

    3 are basically the same.” Can you tell from the three lines of sample data

    that is shown why she says that?

    Field 2  ≈ 7  for the values displayed, it can be dangerous to draw

    Field 3           conclusions from this kind of small samples as the two fields   contain the same information.

3. You are approached by the marketing director of a local company, who believes that he has devised a fool proof way to measure customer satisfaction' He explains his scheme as follows: "It's so simple that I can't believe that no one has thought of it before. I just keep track of the number of customer complaints for each product. I read in a data mining book that counts are ratio attributes, and so, my measure of product satisfaction must be a ratio attribute. But when I rated the products based on my new customer satisfaction measure and showed them to my boss, he told me that I had overlooked the obvious, and that my measure was worthless. I think that he was just mad because our best selling product had the worst satisfaction since it had the most complaints. Could you help me set him straight?"

    (a) Who is right, the marketing director or his boss? If you answered, his

          boss, what would you do to fix the measure of satisfaction?

The boss is right. It can be measured as,

Satisfaction(product) = number of complaints for the product

      total number of sales for the product

    (b) What can you say about the attribute type of the original product

satisfaction attribute?

We can not say anything about the attribute type of the original product because two products may have same level of customer satisfaction or something else but they may also have different numbers at some point.

4. A few months later, you are again approached by the same marketing director as in Exercise 3. This time, he has devised a better approach to measure the extent to which a customer prefers one product over other, similar products. He explains, "When we develop new products, we typically create several variations and evaluate which one customer prefers. Our standard procedure is to give our test subjects all of the product variations at one time and then ask them to rank the product variations in order of preference. However, our test subjects are very indecisive, especially when there are more than two products. As a result, testing takes forever. I suggested that we perform the comparisons in pairs and then use these comparisons to get the rankings. Thus, if we have three product variations, we have the customers compare variations I and 2, then 2 and 3, and finally 3 and

1. Our testing time with my new procedure is a third of what it was for the old procedure, but the employees conducting the tests complain that they cannot come up with a consistent ranking from the results. And my boss wants the latest

    (a) Is the marketing director in trouble? Will his approach work for

          generating an ordinal ranking of the product variations in terms of

          customer preference? Explain

Yes, the marketing director is in trouble. As the customer can also give inconsistent data.

eg : A customer can prefer 1 - 2, 2 - 3, 3 - 1.

    (b) Is there a way to fix the marketing director’s approach? More generally,

        what can you say about trying to create an ordinal measurement scale

        based on pairwise comparisons?

                One solution is to do only the first two comparisons for the three items.

    In general also it is difficult to create an ordinal measurement scale based     on pairwise comparison because of possible inconsistencies.

    (c) For the original product evaluation scheme, the overall rankings of each

        product variation are found by computing its average over all other

        test subjects. Comment on whether you think that this is a reasonable

        approach. What approaches might you take?

        There is an issue that the scale is not probably an interval or ratio scale.

        But for practical purposes an average will be a good one. Also one more       important point is that the extreme ratings might result in misleading

the overall rating, so the median or trimmed mean would be a better         choice.

5. Can you think of a situation in which identification numbers would be useful for prediction?

    For example take phone numbers as a id number and with the help of

    area codes we can give a good prediction of locations.

6. 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 rule analysis works with binary attributes, so you have to convert original data into binary form as follows

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Q1 = A | Q1 = B | Q1 = C | Q1 = D | ...... | Q100 = A | Q100 = B | Q100 = C | Q100 = D |
| 1 | 0 | 0 | 0 | …... | 1 | 0 | 0 | 0 |
| 0 | 0 | 1 | 0 | …... | 0 | 1 | 0 | 0 |

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

       are there?

400 asymmetric binary attributes.

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

    If locations that are closer to each other are more similar with respect to the

    values of that feature than locations that are farther away than the spatial

    auto-correlation feature will be shown. So, physically close areas will have

    similar temperatures, hence they show more autocorrelation when compared

    to rainfall because rainfall may change from one location to another.

8. Discuss why a document-term matrix is an example of a data set that has asymmetric discrete or asymmetric continuous features.

    The ijth entry of a document-term matrix is the number of times that term j

    occurs in document i. Most of the documents contain only a small fraction of all

    the possible terms due to which zero entries are not very meaningful either in

       describing or comparing documents.So a document-term matrix will have

    asymmetric discrete features.If we apply a TFIDF normalization to terms and

    normalize the documents to have an L2 norm of 1 then this creates a

    term-document matrix with continuous features. Then also the features are still

    asymmetric because these transformations do not create any non-zero entries

    for any entries that were previously 0. So, zero entries are still not of any

    meaning.

9. Many sciences rely on observation instead of (or in addition to) designed experiments. Compare the data quality issues involved in observational science with those of experimental science and data mining.

    Observational sciences have the issue of not being able to completely control

    the quality of the data that they obtain.

      For example, until earth orbiting satellites became available measurements of sea surface temperature relied on measurements from ships. Similarly, weather measurements are often taken from stations located in towns or cities. So, it is necessary to work with the data available rather than data from a carefully designed experiment. Hence data analysis for observational

Science resembles data mining.

10. Discuss the difference between the precision of a measurement and the terms single and double precision, as they are used in computer science, typically to represent floating-point numbers that require 32 and 64 bits, respectively.

      The precision of floating point numbers is the maximum precision. Also

      clearly precision is often expressed in terms of the number of significant digits

     used to represent a value. Thus, a single precision number can only represent

     values with up to 32 bits ≈ 9 decimal digits of precision. Also, the precision of

     a value represented using 32 bits (64 bits) is far less than 32 bits (64 bits).

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

1. Text files can be easily modified and also can be inspected by typing the file or viewing it with a text editor.
2. Text files are more portable than binary files, both across systems and programs.

     12. Distinguish between noise and outliers. Be sure to consider the following questions.

(a) Is noise ever interesting or desirable? Outliers?

     No, but based on definition yes.

(b) Can noise objects be outliers?

      Yes, because random distortion of the data is usually responsible for

      outliers.

(c) Are noise objects always outliers?

     No, random distortion can result in an object or value much like a

     normal one.

(d) Are outliers always noise objects?

     No, generally outliers will only represent a class of objects that are

     different from normal objects.

(e) Can noise make a typical value into an unusual one, or vice versa?

      Yes

13. Consider the problem of finding the K nearest neighbours of a data object. A programmer designs Algorithm 2.2 for this task.

Algorithm 2.1 Algorithm for finding K nearest neighbors.

1: for i = 1 to number of data objects do

2: Find the distances of the ith object to all other objects.

3: Sort these distances in decreasing order. (Keep track of which object is associated    with each distance.)

4: return the objects associated with the first K distances of the sorted list

5: end for

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.

    Some of the problems are,

1. The order of duplicate objects on a nearest neighbor list will depend on details of the algorithm and the order of objects in the data set.
2. If there are enough duplicates, the nearest neighbor list may consist only of duplicates.
3. An object may not be its own nearest neighbor.

b) How would you fix this problem?

     There are several approaches,

    One approach is to keep only one object for each group of duplicate objects.

    In this case, each neighbor can represent either a single object or a group of

    duplicate objects.

14. The following attributes are measured for members of a herd of Asian elephants: weight, height, tusk length, trunk length, and ear area. Based on these measurements, what sort of similarity measure from Section 2.4 would you use to compare or group these elephants? Justify your answer and explain any special circumstances.

These attributes are all numerical but can have a widely varying range of values depending on the scale used to measure them. Also the attributes are not asymmetric and the magnitude of an attribute matters.These two facts later eliminate the cosine and correlation measure.Euclidean distance, applied after standardizing the attributes to have a mean of 0 and a standard deviation of 1, would be appropriate.

15. You are given a set of m objects that is divided into K groups, where the ith 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 is guaranteed to get the same number of objects from each

    group while for the second scheme will vary. More specifically, the second

    scheme only guarantees that on average the number of objects from each

    group will be n ∗ mi/m.

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

tf’ ij = tfij ∗ log m/dfi ,

where dfi is the number of documents in which the i th term appears and is known as the document frequency of the term. This transformation is known as the inverse document frequency transformation.

  (a) What is the effect of this transformation if a term occurs in one document? In

    every document?

     Terms that occur in every document have zero weight while those which occur

     in only one document will have maximum weight, i.e., log m.

(b) What might be the purpose of this transformation?

     This normalization reflects the observation that terms which occur in every

     document do not have any power to distinguish one document from another

     while those that are relatively rare do.

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?

     (a2, b2)

(b) Give an equation that relates y to x.

      In this interval y = x2

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 = 3

    there are only 3 binary numbers different between the x and y.

   Jaccard coefficient : J= (number of matching presences) / (number of attributes not involved in 00 matches)

    J = (2) / (1 + 2 + 2)

    J = 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.)

   The Hamming distance is similar to the SMC.

            SMC = Hamming distance / number of bits.

   The Jaccard measure is similar to the cosine measure because both ignore 0-0

   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 is more appropriate for comparing the genetic makeup of two

  organisms as we want to see how many genes these two organisms share.

(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.)

    As two human beings share >99.9% of the same genes. If we want to

    compare the genetic makeup of them we should focus on their differences. So

    the hamming distance is more appropriate in this situation

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) = 1, corr (x, y) = 0/0 (undefined), Euclidean (x, y) = 2

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

    cos (x, y) = 0, corr (x, y) = −1, Euclidean (x, y) = 2, Jaccard (x, y) = 0

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

     cos (x, y) = 0, corr (x, y) = 0, Euclidean (x, y) = 2

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

     cos (x, y) = 0.75, corr (x, y) = 0.25, Jaccard (x, y) = 0.6

(e) x = (2, -7,0,2,0, -3), y = ( -1, 1, -1,0,0, -1) cosine, correlation

     cos (x, y) = 0, corr (x, y) = 0

20. Here, we further explore the cosine and correlation measures.

     (a) What is the range of values that are possible for the cosine measure?

[−1, 1]. Many times the data will have only positive entries and in such

case the range is [0, 1].

      (b) If two objects have a cosine measure of 1, are they identical? Explain.

Not necessarily. All we know is that the values of their attributes differ by a

constant factor.

     (c) What is the relationship of the cosine measure to correlation, if any? (Hint:

Look at statistical measures such as mean and standard deviation in

cases where cosine and correlation are the same and different.)

   For vectors x and y that have a mean of 0 corr(x, y) = cos(x, y).

    (d) Figure 2.1(a) shows the relationship of the cosine measure to Euclidean

        distance for 100,000 randomly generated points that have been normalized

        to have an L2 length of 1. What general observation can you make about

                    the relationship between Euclidean distance and cosine similarity when

          vectors have an L2 norm of 1?

      Since all the 100,000 points fall on the curve, there is a functional relationship

      between Euclidean distance and cosine similarity for normalized data. More

      specifically, there is an inverse relationship between cosine similarity and

      Euclidean distance. For example, if two data points are identical, their cosine

      similarity is one and their Euclidean distance is zero, but if two data points

      have a high Euclidean distance, their cosine value is close to zero. Note that

      all the sample data points were from the positive quadrant, i.e., had only

      positive values. This means that all cosine (and correlation) values will be

      positive.

    (e) Figure 2.1(b) shows the relationship of correlation to Euclidean distance for

100,000 randomly generated points that have been standardized to have

a mean of 0 and a standard deviation of 1. What general observation can

you make about the relationship between Euclidean distance and

correlation when the vectors have been standardized to have a mean of 0

and a standard deviation of 1?

     Same as previous answer, but with correlation substituted for cosine.

      (f) Derive the mathematical relationship between cosine similarity and

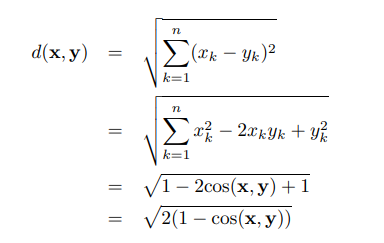
Euclidean distance when each data object has an L2 length of 1.

    Let x and y be two vectors where each vector has an L2 length of 1. For

such vectors, the variance is just n times the sum of its squared attribute

values and the correlation between the two vectors is their dot product

divided by n.



    (g) Derive the mathematical relationship between correlation and

Euclidean distance when each data point has been been standardized by

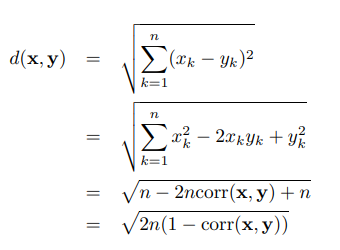
subtracting its mean and dividing by its standard deviation.

    Let x and y be two vectors where each vector has an a mean of 0 and a

standard deviation of 1. For such vectors, the variance is just n times the

sum of its squared attribute values and the correlation between the two

vectors is their dot product divided by n.

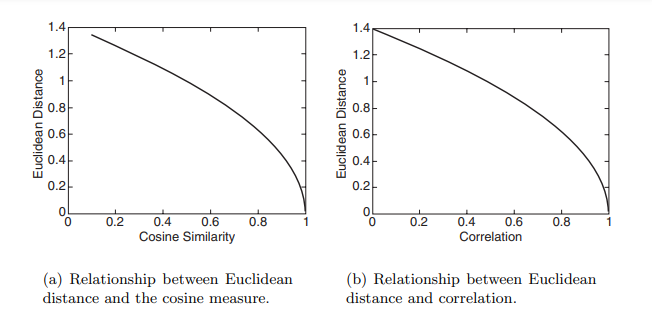


 21. Show that the set difference metric given by

          d(A, B) = size(A − B) + size(B − A)

satisfies the metric axioms given on page 70. A and B are sets and A − B

is the set difference.



1(a). Because the size of a set is greater than or equal to 0, d(x, y) ≥ 0.

1(b). if A = B, then A − B = B − A = empty set and thus d(x, y)=0

2. d(A, B) = size(A−B)+size(B−A) = size(B−A)+size(A−B) = d(B,A)

3. First, note that d(A, B) = size(A) + size(B) − 2size(A ∩ B).

  ∴ d(A, B)+d(B,C) = size(A)+size(C)+2size(B)−2size(A∩B)−2size(B∩ C)

   Since size(A ∩ B) ≤ size(B) and size(B ∩ C) ≤ size(B),

  d(A, B) + d(B,C) ≥ size(A) + size(C)+2size(B) − 2size(B) = size(A) +

   size(C) ≥ size(A) + size(C) − 2size(A ∩ C) = d(A, C)

∴ d(A, C) ≤ d(A, B) + d(B,C)

22. Discuss how you might map correlation values from the interval [−1,1] to the

      interval [0,1]. Note that the type of transformation that you use might depend

      on the application that you have in mind. Thus, consider two applications:

      clustering time series and predicting the behavior of one time series given

      another.

        For time series clustering, time series with relatively high positive correlation

      should be put together. For this purpose, the following transformation would

      be appropriate:

https://lh4.googleusercontent.com/qMuAG9jlLxSjecBREnQ5Ru5XjNlOTFjdc9Zzr9jj--0IkZlDZK_DY9O7TQ8V_Z-lilapkmmAnpoGNDwnNc9TazB69H3s-5lZ1l5bqyBNdjvNOCWkKru5_Qjf3GdGWg

     For predicting the behavior of one time series from another, it is necessary to

     consider strong negative, as well as strong positive, correlation. In this case,

     the following transformation, sim = |corr| might be appropriate. Note that this

     assumes that you only want to predict magnitude, not direction.

23. Given a similarity measure with values in the interval [0,1] describe two ways

      to transform this similarity value into a dissimilarity value in the interval [0,∞].

https://lh6.googleusercontent.com/jjL3wMqcG5zFsFFZeHjJlSPbgggDT97t6cWY4GfWpZK2MZQ2sowQwczi6_KX1b8ef_ExrTfOUlLAVKhrt8-h_lTnaP18EOzsdyHJDZhVz7VzinaJw7N1WNaxulUE7Q

24. Proximity is typically defined between a pair of objects.

1. Define two ways in which you might define the proximity among a group

of objects.

Two examples are the following:

      (i) based on pairwise proximity, i.e., minimum pairwise similarity or

    maximum pairwise dissimilarity, or

(ii) for points in Euclidean space compute a centroid and then compute the

     sum or average of the distances of the points to the centroid.

     (b) How might you define the distance between two sets of points in Euclidean

space?

One approach is to compute the distance between the centroids of the two

sets of points.

      (c) How might you define the proximity between two sets of data objects?

(Make no assumption about the data objects, except that a proximity

measure is defined between any pair of objects.)

  One approach is to compute the average pairwise proximity of objects in

one group of objects with those objects in the other group. Other

approaches are to take the minimum or maximum proximity.

25. You are given a set of points S in Euclidean space, as well as the distance of

      each point in S to a point x. (It does not matter if x ∈ S.)

      (a) If the goal is to find all points within a specified distance ε of point y, y= x,

explain how you could use the triangle inequality and the already

calculated distances to x to potentially reduce the number of distance

calculations necessary? Hint: The triangle inequality, d(x, z) ≤ d(x, y) + d(y,

x), can be rewritten as d(x, y) ≥ d(x, z) − d(y, z).

        Unfortunately, there is a typo and a lack of clarity in the hint. The hint

should be phrased as follows:

Another application of the triangle inequality starting with d(x, z) ≤ d(x, y) +

d(y, z), shows that d(y, z) ≥ d(x, z) − d(x, y). If the lower bound of d(y, z)

obtained from either of these inequalities is greater than , then d(y, z) does

not need to be calculated. Also, if the upper bound of d(y, z) obtained from

the inequality d(y, z) ≤ d(y, x)+d(x, z) is less than or equal to , then d(x, z)

does not need to be calculated.

      (b) In general, how would the distance between x and y affect the number of

distance calculations?

    If x = y then no calculations are necessary. As x becomes farther away,

typically more distance calculations are needed.

      (c) Suppose that you can find a small subset of points S , from the original

data set, such that every point in the data set is within a specified distance

ε of at least one of the points in S , and that you also have the pairwise

distance matrix for S . Describe a technique that uses this information to

compute, with a minimum of distance calculations, the set of all points

within a distance of β of a specified point from the data set.

  Let x and y be the two points and let x∗ and y∗ be the points in S that are

closest to the two points, respectively. If d(x∗, y∗)+2 ≤ β, then we can

safely conclude d(x, y) ≤ β. Likewise, if d(x∗, y∗)−2 ≥ β, then we can safely

conclude d(x, y) ≥ β. These formulas are derived by considering the cases

where x and y are as far from x∗ and y∗ as possible and as far or close to

each other as possible.

26. Show that 1 minus the Jaccard similarity is a distance measure between two data objects, x and y, that satisfies the metric axioms given on page 70. Specifically, d(x, y)=1 − J(x, y).

1(a). Because J(x, y) ≤ 1, d(x, y) ≥ 0.

1(b). Because J(x, x) = 1, d(x, x)=0

2. Because J(x, y) = J(y, x), d(x, y) = d(y, x)

3. (Proof due to Jeffrey Ullman)

    minhash(x) is the index of first nonzero entry of x prob(minhash(x) = k)

    is the probability the minhash(x) = k when x is randomly permuted.

  Note that prob(minhash(x) = minhash(y)) = J(x, y) (minhash lemma)

   Therefore, d(x, y) = 1−prob(minhash(x) = minhash(y))

          = prob(minhash(x) ≠ minhash(y))

   We have to show that,

     prob(minhash(x) ≠ minhash(z)) ≤ prob(minhash(x)≠ minhash(y)) +

prob(minhash(y) ≠ minhash(z)

27. Show that the distance measure defined as the angle between two data vectors, x and y, satisfies the metric axioms given on page 70. Specifically, d(x, y) = arccos(cos(x, y)).

Note that angles are in the range 0 to 180◦.

1(a). Because 0 ≤ cos(x, y) ≤ 1, d(x, y) ≥ 0.

1(b). Because cos(x, x)=1, d(x, x) = arccos(1) = 0

2. Because cos(x, y) = cos(y, x), d(x, y) = d(y, x)

3. If the three vectors lie in a plane then it is obvious that the angle between x

    and z must be less than or equal to the sum of the angles between x and y

    and y and z. If y is the projection of y into the plane defined by x and z, then

    note that the angles between x and y and y and z are greater than those

    between x and y and y and z.

28. Explain why computing the proximity between two attributes is often simpler than computing the similarity between two objects.

     In general, an object can be a record whose fields (attributes) are of

     different types. To compute the overall similarity of two objects in this

     case, we need to decide how to compute the similarity for each attribute

     and then combine these similarities. This can be done straightforwardly by

     using Equations 2.15 or 2.16, but is still somewhat ad hoc, at least compared

     to proximity measures such as the Euclidean distance or correlation, which

     are mathematically well founded. In contrast, the values of an attribute are all

    of the same type, and thus, if another attribute is of the same type, then the  computation of similarity is conceptually and computationally straightforward.

3) Do Chapter 2 textbook [problem #2](http://www.cob.sjsu.edu/mease_d/bus297D/ch2textbookquestion.doc) on page 89.

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.

      Binary, qualitative - nominal

(b) Brightness as measured by a light meter.

      Continuous, quantitative - ratio

(c) Brightness as measured by people's judgments.

     Continuous, qualitative - ordinal

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

     Continuous, quantitative - ratio

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

      Discrete, qualitative - ordinal

(f) Height above sea level.

     Continuous, quantitative - interval/ratio

(g) Number of patients in a hospital.

     Discrete, quantitative - ratio

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

     Discrete, qualitative - nominal

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

    Discrete, qualitative - ordinal

(j) Military rank.

    Discrete, qualitative - ordinal

(k) Distance from the center of campus.

     Continuous, quantitative - ratio

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

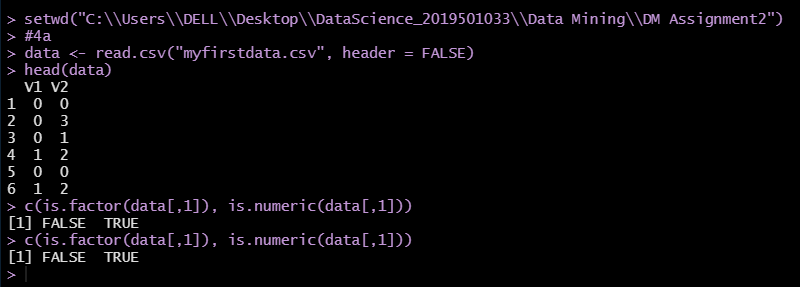
     Continuous, quantitative - ratio

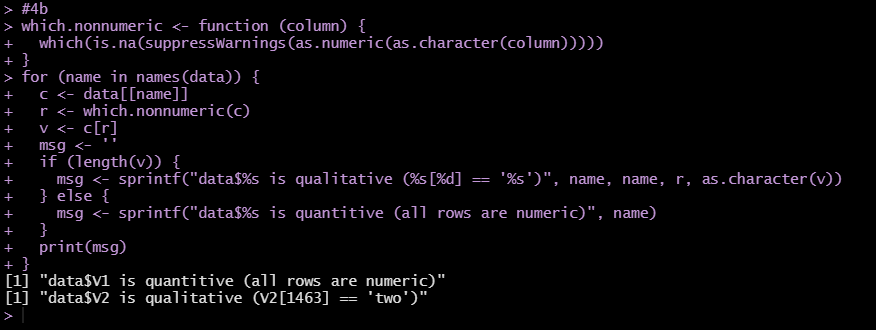
(m) Coat check number. (When you attend an event, you can often give your coat to someone who, in turn, gives you a number that you can use to claim your coat when you leave.)

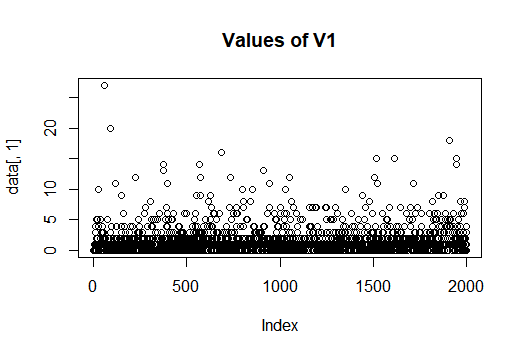
      Discrete, qualitative - nominal

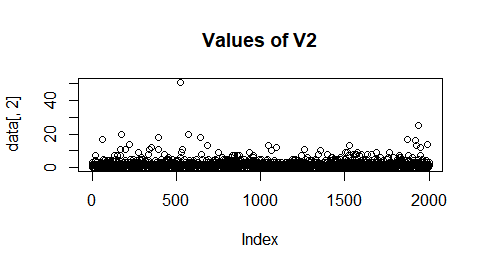
4) This question uses the data at <http://www.cob.sjsu.edu/mease_d/bus297D/myfirstdata.csv>. Download it to your computer.  
  
a) Read in the data in R using data←read.csv("myfirstdata.csv",header=FALSE).

Note, you first need to specify your working directory using the setwd() command. Determine whether each of the two attributes (columns) is treated as qualitative (categorical) or quantitative (numeric) using R. Explain how you can tell using R.

  
  
b) What is the specific problem that causes one of these two attributes to be read in as qualitative (categorical) when it seems it should be quantitative (numeric)?

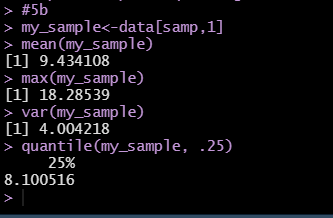
  
  
c) Use the command plot() in R to make a plot for each column by entering plot(data[,1]) and plot(data[,2]). Because one variable is read in as quantitative (numeric) and the other as qualitative (categorical) these two plots are showing completely different things by default. Explain exactly what is being plotted in each of the two cases. Include these two plots in your homework.

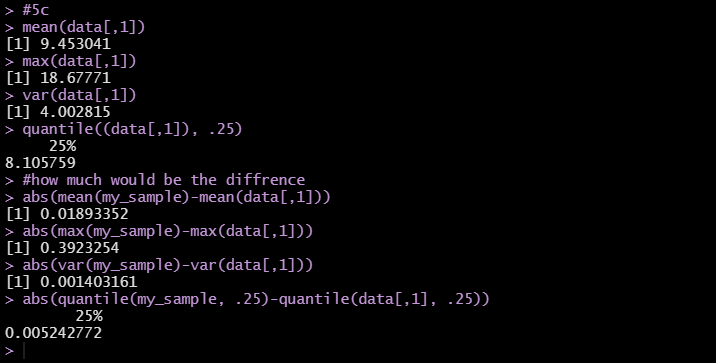


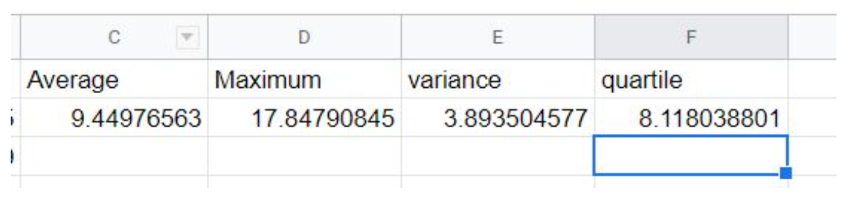
  
  
d) Read the data into Excel. Excel should have no problem opening the file directly since it is .csv. Create a new column that is equal to the second column plus 10. What is the result for the problem observations (rows) you identified in part b? What specific outcome does Excel display?

There is a non-numeric value present in line 1463. Due to which we get an undefined     error. The reason for that error is, because we are trying to add a string (i.e., "two") which is in the second column with int (i.e., 10) which is not possible.  
  
  
5) This question uses the data at <http://www.cob.sjsu.edu/mease_d/bus297D/twomillion.csv>. Download it to your computer.  
  
a) Read the data into R using data<-read.csv("twomillion.csv",header=FALSE). Note, you first need to specify your working directory using the setwd() command. Extract a simple random sample with replacement of 10,000 observations (rows). Show your R commands for doing this.

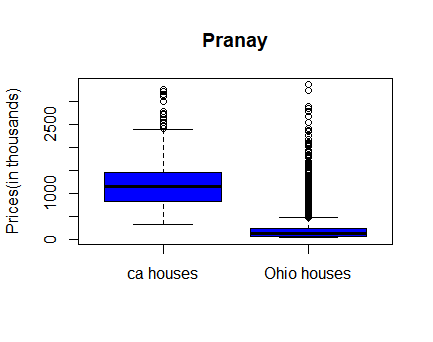
C:\Users\DELL\Desktop\DM\DM Assignment2\5a.PNG  
  
b) For your sample, use the functions mean(), max(), var() and quantile(,.25) to compute the mean, maximum, variance and 1st quartile respectively. Show your R code and the resulting values.

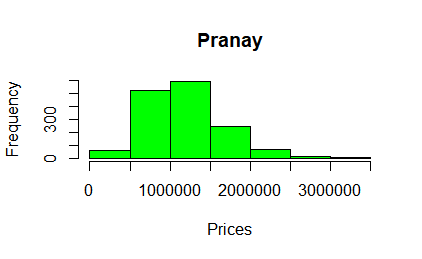
  
c) Compute the same quantities in part b on the entire data set and show your answers. How much do they differ from your answers in part b?

  
  
d) Save your sample from R to a csv file using the command write.csv(). Then open this file with Excel and compute the mean, maximum, variance and 1st quartile. Provide the values and name the Excel functions you used to compute these.

  
  
e) Exactly what happens if you try to open the full data set with Excel?  
 All the data of 1048576 rows will be displayed.  
6) Read Chapter 3 (only sections 3.1, 3.2 and 3.3).  
  
7) This question uses a sample of 1500 California house prices at <http://www-stat.wharton.upenn.edu/~dmease/CA_house_prices.csv> and a sample of 10,000 Ohio house prices at <http://www-stat.wharton.upenn.edu/~dmease/OH_house_prices.csv>. Download both data sets to your computer. Note that the house prices are in thousands of dollars.

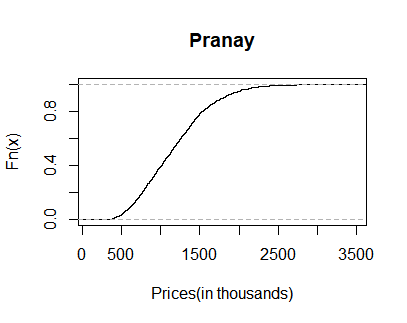
C:\Users\DELL\Desktop\DM\DM Assignment2\7.PNG  
  
a) Use R to produce a single graph displaying a boxplot for each set (as in ICE #16). Include the R commands and the plot. Put your name in the title of the plot (for example, main="Britney Spears' Boxplots").

C:\Users\DELL\Desktop\DM\DM Assignment2\7ar.PNG   
  
b) Use R to produce a frequency histogram for only the California house prices. Use intervals of width $500,000 beginning at 0 and ending at $3.5 million. Include the R commands and the plot. Put your name in the title of the plot.

C:\Users\DELL\Desktop\DM\DM Assignment2\7br.PNG   
  
  
c) Use R to plot the ECDF of the California houses and Ohio houses on the same graph (as in ICE #11). Include a legend. Include the R commands and the plot. Put your name in the title of the plot.

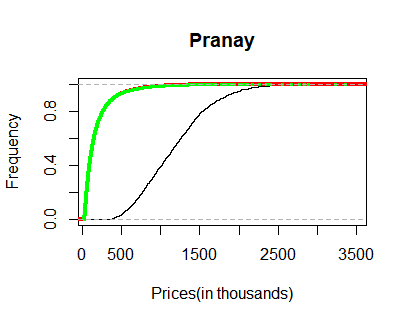
> #7c

> plot(ecdf(ca\_data[,1]),verticals = TRUE,do.p=FALSE,main    ="Veda",xlab="Prices(in thousands)",ylabs="Frequency")



> plot(ecdf(ca\_data[,1]), verticals=TRUE,do.p = FALSE, main = "Veda",xlab="Prices(in thousands)",ylab="Frequency")

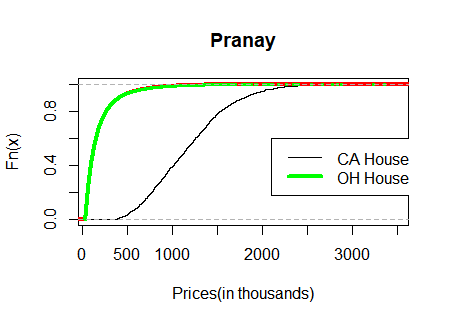
> lines(ecdf(oh\_data[,1]),verticals= TRUE,do.p = FALSE,col.h="red",col.v="green",lwd=4)



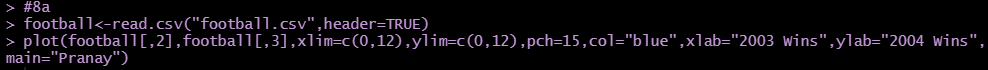
> plot(ecdf(ca\_data[,1]),verticals = TRUE,do.p=FALSE,main ="Veda",xlab="Prices(in thousands)",ylabs="Frequency")

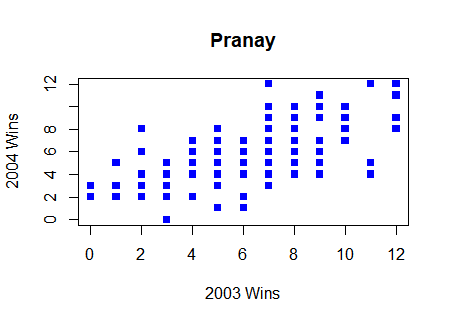
> lines(ecdf(oh\_data[,1]),verticals = TRUE,do.p=FALSE,col.h="red",col.v="green",lwd=4)

> legend(2100,.6,c("CA Houses","OH Houses"),col=c("black","green"),lwd=c(1,4))



8) This question uses the data at <http://www-stat.wharton.upenn.edu/~dmease/football.csv>. Download it to your computer. This data set gives the total number of wins for each of the 117 Division 1A college football teams for the 2003 and 2004 seasons.   
  
a) Use plot() in R to make a scatter plot for this data with 2003 wins on the x-axis and 2004 wins on the y-axis. Use the range 0 to 12 for both the x-axis and y-axis. Include the R commands and the plot. Put your name in the title of the plot.



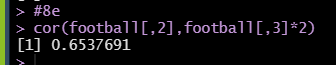
  
  
b) Why are there fewer than 117 points visible on your graph in part a? Describe the solution we discussed in class to deal with this problem (but don't actually do it).

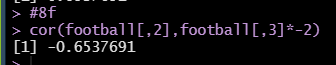
As we can see from the graph some of the data is not visible as they are plotted on the same axes, due to which they are plotted on top of each other. We can overcome this problem by adding a small amount of noise to the data plotted on the graph.

c) Compute the correlation in R using the function cor().

C:\Users\DELL\Desktop\DM\DM Assignment2\8c.PNG  
  
d) How does the value in part c change if you add 10 to all the values for 2004?

C:\Users\DELL\Desktop\DM\DM Assignment2\8d.PNG  
  
e) How does the value in part c change if you multiply all the 2004 values by 2?

  
  
f) How does the value in part c change if you multiply all the 2004 values by -2?

  
  
9) This question uses the sample of 10,000 Ohio house prices at <http://www-stat.wharton.upenn.edu/~dmease/OH_house_prices.csv>. Download the data set to your computer. Note that the house prices are in thousands of dollars.

C:\Users\DELL\Desktop\DM\DM Assignment2\9.PNG

a) What is the median value? Is it larger or smaller than the mean?

C:\Users\DELL\Desktop\DM\DM Assignment2\9a.PNG

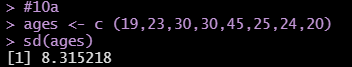
b) What does your answer to part a suggest about the shape of the distribution (right-skewed or left-skewed)?

C:\Users\DELL\Desktop\DM\DM Assignment2\9b.PNG  
  
c) How does the median change if you add 10 (thousand dollars) to all the values?

C:\Users\DELL\Desktop\DM\DM Assignment2\9c.PNG

d) How does the median change if you multiply all the values by 2?

C:\Users\DELL\Desktop\DM\DM Assignment2\9d.PNG  
  
10) This question uses the following people's ages: 19,23,30,30,45,25,24,20. Store them in R using the syntax ages<-c(19,23,30,30,45,25,24,20).  
  
a) Compute the standard deviation in R using the sd() function.

  
  
b) Compute the same value by hand and show all the steps.

Numbers : 19, 23, 30, 45, 25, 24, 20

     Mean : (19 + 23 + 30 + 30 + 45 + 25 + 24 + 20) / 8 = 216 / 8 = 27

     List of deviations : -8, -4, 3, 3, 18, -2, -3, -7

     Sum of deviations : 64 + 16 + 9 + 9 + 324 + 4 + 9 + 49 = 484

     Squares of deviations : 64, 16, 9, 9, 324, 4, 9, 49

    Divided by one less than the number of items in the list : 484 / 7 = 69.14285

     Square root of this number : sqrt (69.14285) = approx. 8.31521  
  
  
c) Using R, how does the value in part a change if you add 10 to all the values?

C:\Users\DELL\Desktop\DM\DM Assignment2\10c.PNG  
  
d) Using R, how does the value in part a change if you multiply all the values by 100?

