**2**

**Data**

Contents

[**Attributes and Measurement** 2](#_Toc63336145)

[**True/False and Short Answer** 6](#_Toc63336146)

[**Types of Data Sets** 12](#_Toc63336147)

[**Aggregation and Sampling** 16](#_Toc63336148)

[**Discretization and Binarization** 21](#_Toc63336149)

[**Measures of Similarity and Dissimilarity** 22](#_Toc63336150)

# **Attributes and Measurement**

1. Classify the following attributes as binary, discrete, or continuous. Further, classify the attributes as qualitative (nominal or ordinal) or quantitative (interval or ratio). Some of the 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. Age as measured by whether age is less than 30 (value = 0) or greater than or equal to 30 (value = 1).

Binary, qualitative, ordinal.

1. Increase in profit over the previous year.

Continuous, quantitative, ratio.

1. Ranking of a candy bar by a person on a scale of 1 to 10.

**Alternative:**  Sweetness of a candy bar on a scale of 1 to 10.

Discrete, qualitative, ordinal (if no fractional ranking allowed, e.g., 8.5)

Continuous, qualitative, ordinal (if a fractional rating allowed, e.g., 8.5)

1. Ranking of a sports team based on the votes cast by 500 sportswriters. Each sports writer is asked which team they believe to be #1, and the ranking for a team is based on the number of #1 votes received.

Discrete, quantitative, ratio (these are counts). An alternative answer is discrete, qualitative, ordinal, if only the ranking matters and not the total number of votes.

1. Ranking of movies based on the rating given by the viewers. Each viewer can rate a movie on a scale of 1 to 10, and the ranks of movies are determined based on the average of all ratings received for a movie.

discrete, qualitative, ordinal

Movie ratings are ordinal to begin with, and converting the sums to ranks (i.e., 1, 2, 3, …) further muddles the difference information. For example, the difference in average score between movies ranked 10 and 11 could be different than the difference in average score between movies ranked 11 and 12.

1. Number of votes received by political candidates in a specific race.

discrete, quantitative, ratio

1. Unix time, which is the number of seconds that have elapsed since 00:00:00 Thursday, 1 January 1970, Coordinated Universal Time (UTC), minus leap seconds.

discrete, quantitative, interval. (It is interval since the origin of 1970 is arbitrary)

1. Marks scored on an exam by a student. The marks are bounded between 0 and 100.

continuous/discrete (both are OK), quantitative, ratio.

1. Brightness in lumens as measured by a light meter.

continuous, quantitative, ratio

1. Brightness as measured by people’s judgments.

discrete, qualitative, ordinal

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

discrete, qualitative, ordinal

1. Height above sea level.

continuous, quantitative, interval/ratio (depends on whether sea level is regarded as an arbitrary origin)

1. Number of patients in a hospital.

discrete, quantitative, ratio

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

discrete, qualitative, nominal (ISBN numbers do have order information, though)

1. Military rank.

discrete, qualitative, ordinal

1. Angles as measured in degrees between 0◦ and 360◦.

continuous, quantitative, ratio

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

discrete, qualitative, ordinal

1. Density of a substance in grams per cubic centimeter.

continuous, quantitative, ratio

1. 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

1. Speed of vehicle measured in miles per hour.

continuous, Quantitative, Ratio

1. Altitude of a region.

continuous, Quantitative, Ratio

1. Intensity of rain as indicated using the values: no rain, intermittent rain, constant rain.

discrete, Qualitative, Ordinal

1. Barcode number printed on each item in a supermarket.

discrete, Qualitative, Nominal

1. The size of different objects as measured by people’s judgments.

discrete, qualitative, ordinal

1. Distances between two astronomical objects as measured in light-years.

continuous, quantitative, ratio

1. Percentage of ones in an m-by-n binary matrix (only 0/1 entries).

continuous, quantitative, ratio

1. Seat numbers assigned to passengers on a flight.

discrete, qualitative, nominal/ordinal

1. Different flavors of ice cream.

discrete, qualitative, nominal

1. House numbers assigned for a given street.

discrete, qualitative, ordinal

1. Your calorie intake per day

continuous. quantitative, ratio

1. Shape of geometric objects commonly found in geometry classes.

discrete, qualitative, nominal

1. Routes in rock climbing

\*\*NOTE: Routes in rock climbing are rated using the Yosemite Decimal System (YDS).

Following is a description of the class 5 range routes in YDS (Reference: Mountaineering, Freedom of the Hills, Ed Peters):

○ 5.0 to 5.4 - There are two hands and two footholds for every move; the holds become progressively smaller as the number increases.

○ 5.5 to 5.6 - The two hand- and two footholds are there, obvious to the experienced, but not necessarily so to the beginner.

○ 5.7 -The move is missing one hand- or foothold.

○ 5.8 - The move is missing two holds of the four or missing only one but is very strenuous.

○ 5.9 - The move has only one reasonable hold, which may be for either a foot or a hand.

○ 5.10 - No hand- or footholds. The choices are to pretend a hold is there, pray a lot, or go home.

○ 5.11 - After a thorough inspection, you conclude this move is obviously impossible; however, occasionally, someone actually accomplishes it. Since there is nothing for a handhold, grab it with both hands.

○ 5.12 - The surface is as smooth as glass and vertical. No one has really ever made this move, although a few claim they have.

○ 5.13 - This is identical to 5.12, except it is located under an overhanging rock.”

discrete, qualitative, ordinal. Although 5.0 to 5.4 and 5.5 to 5.6 may seem to be continuous, 5.7, 5.8, 5.9, etc. should indicate the discrete nature of the variable.

1. ASCII codes. (http://en.wikipedia.org/wiki/ASCII).

discrete, qualitative, ordinal/nominal

1. Percentage of fat in your body.

Continuous, quantitative, ratio

1. The distance between any terrestrial body and the Earth.

Continuous, quantitative, ratio

1. Movie ratings on a scale of 1 to 10.

discrete, qualitative, ordinal

1. Degree to which people like various hobbies as measured by the fraction of their income they spend on them.

continuous, quantitative, ratio

1. State of a traffic light at an intersection where the light can be red, yellow, or green.

discrete, qualitative, nominal. (ordinal can be acceptable with justification.)

1. Amount of money in dollars received by a politician’s campaign.

continuous/discrete, quantitative, ratio.

1. Increase in profit of the current year over the profit obtained in the year 2000.

continuous, quantitative, interval.

1. The position of a runner on the track in terms of the distance they have run since the start of the race.

continuous, quantitative, ratio

1. The position of a runner on the track in terms of the angle the runner’s position makes with the center of the track. The starting line is at a position of 0 degrees, while the position halfway around the track is 180 degrees. When one lap is completed, that is, the runner passes the starting line again, the angle starts over at 0.

Continuous and quantitative but doesn’t fit either ratio or interval because the attribute is cyclical, and the true difference between two positions is not just their difference, e.g., consider the positions of 1 degree and 359 degrees.

1. The position of the runner among all other runners, i.e., first, second, third, etc. Discrete, qualitative, ordinal (with justification, discrete, quantitative, interval
2. The different teams to which the runners belong, Discrete, qualitative, nominal

# **True/False and Short Answer**

1. For each of the following questions, give a True or False answer and a one-sentence justification.
2. Noise is not a problem with count data. Explain.

False. Counts can be erroneous as can any other value.

1. For any two sets of real values, such as two vectors of size n > 0, the correlation is a value between -1 and 1. Explain.

False. If one of the sets of values is constant, the correlation will be undefined.

1. For reducing the size of a daily time series, it would be better to sample than aggregate since sampling is a simpler process. Explain.

False. Sampling would pick values at random and could miss entire periods of time, e.g., weeks or months, while aggregation can be done in a way that preserves the structure of the data, e.g., aggregating by weeks or months.

1. Noise and outliers are sometimes the same. Explain.

True. Noise can produce an anomalous value.

1. If an object is an outlier, then it is noise.

False. An outlier can result from noise, but that is not always the case.

1. A binary attribute with values 0 or 1 is also an asymmetric binary attribute.

False. An ordinary binary attribute converts to two asymmetric binary attributes.

1. If vectors of counts have a cosine measure of 1, the objects are identical. Explain.

False. The two vectors will have a cosine measure if both are non-zero, and one is a multiple of the other.

1. Discrete variables cannot be ratio.

False: Discrete variables can be ratio, e.g., counts.

1. Quantitative variables are continuous.

False: Counts are quantitative but are discrete.

1. Converting ordinal variables to asymmetric binary variables does not lose any information.

False, it loses order.

1. Converting a continuous variable to an ordinal categorical variable loses only magnitude information.

False. To do such a conversion, you will have to assign an interval of values for the continuous variable to the same ordinal value. This will lose order information to some extent.

1. A continuous attribute is always a ratio or an interval attribute.

False. Consider a ratio variable that is transformed by squaring or taking the square root.

1. A ratio attribute is also an interval attribute.

True. By definition, a ratio attribute also has the property of meaningful attributes.

1. The correlation between two vectors [5,4,3,4,5] and [3,3,3,3,3] is undefined.

True. Correlation is undefined whenever one of the vectors is constant since the standard deviation is 0.

1. All ratio type data is also interval data.

True. Ratio data has meaningful differences in addition to meaningful ratios.

1. Cosine is a better measure than Jaccard for comparing two sparse high-dimensional, real-valued data.

True. Both are suited for sparse data, but cosine also works for continuous data.

1. The only benefit of aggregation is to reduce the number of instances in a data set.

False. Aggregation can also reduce variability.

1. Stratified sampling can only be used with sampling without replacement.

False. Stratified sampling means that we take samples from different segments of a population according to the fraction with which they occur. This concept is independent of whether we sample with or without replacement.

1. Some noise points are indistinguishable from ordinary data. True or False?

True. For example, random noise can produce either normal or abnormal values or objects.

1. To generate a number of different random samples of size n from a data set of size n we can use either sampling with replacement or sampling without replacement.

False. Sampling without replacement will just generate the original data set each time.

1. It is a good idea to standardize an attribute (subtract the mean and divide by the standard deviation) when the attribute is constant except for small variations due to noise.

False. An almost constant attribute with a bit of random noise will show misleading variation after being standardized since the standard deviation will be small and will tend to magnify any deviation from the mean.

1. It is not a good idea to standardize an attribute (subtract the mean and divide by the standard deviation) when the attribute has outliers.

True. Outliers will distort the mean and variance.

1. There are cases in which Euclidean distance may not be symmetric, i.e., d(x,y) ≠ d(y,x).

False. Euclidean distance is always symmetric.

1. Let a1, a2, and a3 be three attributes. If correlation(a1, a2) = 0.5 and correlation of (a2, a3) = 0.5, then correlation (a1, a3) ≤ 0.5.

False. Correlation does not obey the triangle inequality.

1. The Mahalanobis distance is a good distance measure to use when the attributes are correlated.

True, unlike Euclidean distance, the Mahalanobis distance takes variable correlation into account.

1. Euclidean Distance is better than Mahalanobis distance when the attributes are correlated.

False. Mahalanobis distance is better since it takes into account the covariance between attributes when computing the distance between data instances.

1. The Hamming distance between two binary strings (i.e., a string of 0s and 1s), will never exceed the Euclidean distance.

False. If there is a difference of k bits, the Hamming distance is k while the Euclidean distance is sqrt(k).

1. The Hamming distance between two binary strings (i.e., a string of 0s and 1s) is equal to the Euclidean distance squared.

True.

1. The Jaccard coefficient for two binary strings (i.e., a string of 0s and 1s) is always greater than or equal to its cosine similarity.

False. It is easy to find counterexamples. x = [1, 1, 1, 1, 1], y = [1, 1, 1, 1, 0], Jaccard = 0.8, cosine = 0.8944

1. The cosine measure can be used for binary data.

True. However, a person may prefer the Jaccard measure, which is designed for sparse binary data.

1. The correlation of the vectors (1, 1, 1, 1) and (2, 2, 2, 2) is 1.

False. The correlation is undefined since the variance is 0 for a constant vector.

1. The correlation of the vectors (-1, -1, -1, -1) and (2, 2, 2, 2) is 1.

False. The correlation is undefined since the variance is 0 for a constant vector.

1. Mutual information can handle non-linear relationships and asymmetric binary variables.

False. Mutual information, as defined, doesn’t work for asymmetric binary variables since it uses both 0’s and 1’s.

1. Cosine Similarity is better than Correlation because it ignores 0-0 matches in non-binary vector data.

False. It depends on the use case. You might need to capture 0-0 matches for recognizing trends or patterns.

1. If two variables have 0 mean, the cosine and correlation measures will have the same value.

True. If the mean of a variable is 0, then the definition of correlation becomes the same as cosine.

1. Two variables that have a correlation of -1 are unrelated.

False. They are perfectly linearly related, but in a negative way, i.e., an increase in one variable corresponds to a decrease in the other variable.

1. The Manhattan distance between two binary vectors is not the same as the Hamming distance between them.

False. It is the same.

1. For binary vectors, mutual information is more similar to correlation than the Jaccard measure.

True. Both mutual information and correlation consider 0’s and 1’s to be equally important. Thus, they consider 1-1, 1-0, 0-1, and 0-0 matches. The Jaccard measure, like cosine, counts only 1-1 matches.

1. We cannot calculate the mutual information between the vectors (1, 0, 2, 0, 1, 2) and (2, 1, 2, 1, 0, 0).

False. Mutual information can be calculated between any two vectors as long as the distribution of the values can be computed. In this case, where we have 3 values, 0, 1, and 2, we can easily build a probability table and compute the entropy.

1. Multiple choice.
2. Which of the following similarity measures would be most appropriate for document data?

(i) Correlation

(ii) Cosine

(iii) Jaccard

(iv) (i) and (ii)

(v) (ii) and (ii)

(ii) is the correct answer. Typically, document data is represented as a set of counts, which is normalized in the standard way, i.e., TFIDF. (c) Jaccard is appropriate for sparse data, such as document data, but is only for binary data.

1. **Variation of above.** Which one of the following similarity measures is least appropriate for document data? Briefly explain.

i. Correlation

ii. Cosine

iii. Jaccard

Correlation is not appropriate because document data is very sparse, and hence, computing similarity based on 0-0 matches would make almost all documents very similar to one another. Jaccard and cosine ignore 0-0 matches, but correlation does not.

1. If the correlation of two attributes is -0.95, which of the following statements is true?

(i) They provide the opposite information.

(ii) They provide almost the same information.

(iii) They are completely unrelated.

(iv) None of these.

(ii) The magnitude of their correlation is almost 1, which means that one is almost a perfect predictor of the other. The fact that the sign is negative just means that when one attribute goes up, the other goes down. There is no such thing as opposite information.

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

One example: Student IDs are a good predictor of graduation date.

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

A feature shows temporal autocorrelation if its value is more similar at times that are closer to each other than at more distant times. It is more common for temporally close days to have similar temperatures than similar rainfall since the amount of rainfall can change abruptly from one day to another. Therefore, daily temperature shows more temporal autocorrelation than daily rainfall.

1. Complete the following table. Place a tick into a cell if the combination is possible, place a cross otherwise. For example, place a tick into the bottom right corner of the table if you believe that it is possible to create an attribute that is continuous and ratio.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Binary | Discrete | Continuous |
| Nominal |  |  |  |
| Ordinal |  |  |  |
| Interval |  |  |  |
| Ratio |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
|  | Binary | Discrete | Continuous |
| Nominal | X | X |  |
| Ordinal | X | X |  |
| Interval |  | X | X |
| Ratio |  | X | X |

The horizontal labels refer to the number of values an attribute has, while the vertical labels refer to the properties of the attributes. It is difficult to imagine categorical attributes ordered or not that have a potentially uncountable number of values. Indeed, it is hard to imagine actual categorical variables that have even a countable number of values, i.e., categorical variables have only a finite number of values. Those values could, however, be represented by integers or real numbers. A variable is an interval variable only if the differences between different values have the same meaning. Since there are only two values for binary attributes, this cannot be true except in a vacuous sense. Since a binary variable cannot be interval, it cannot be ratio since a ratio.

1. For the following questions, give an answer and a short explanation. (1-2 sentences)

**(**i**)** In general, converting categorical variables (ordinal or nominal variables) to asymmetric binary variables does not lose any information.

This statement is false for ordinal variables since it loses the order information that the values of ordinal variable have. This statement is true for nominal variables since a nominal variable does not have any ordering among values.

**(**ii**)** Suppose you are given a temperature time series the records the temperature at hourly intervals for a specific location. If a single value is missing, could you estimate what this missing value might be with reasonable accuracy? Why or why not?

Yes, this missing value could be estimated reliably because of temporal autocorrelation in the temperature time series.

1. For this question, assume each object is characterized by a set of continuous-valued attributes. Write a brief explanation (1-2 sentences) for your answer.

(a) If two objects have a cosine similarity of 1, must their attribute values be identical? Explain.

No. A cosine similarity of 1 simply implies that the two attribute vectors are parallel to each other. For example, when x = (1,2) and y = (2,4), then their cosine similarity is 1.

(b) If two objects have a correlation value of 1, must their attribute values be identical? Explain.

No. A correlation value of 1 simply implies that there is a linear relationship between the two attribute vectors. For example, if x = (l, 2,3) and y = (3, 5,7), then their correlation is 1.

(c) If two objects have a Euclidean distance of 0, must their attribute values be identical? Explain.

Yes. Consider a pair of objects with attribute vectors x and y. Suppose their Euclidean distance is d(x, y) = 0, which is true if and only if x\_i = y\_i for all i.

# **Types of Data Sets**

1. A group of biologists conducts a field study to evaluate the occurrence of different types of birds at a number of locations. Their data is collected in a table where

- the rows correspond to locations,   
- the columns correspond to different species of birds, and   
- the ijth entry is the number of birds of the jth species at the ith location.   
The following table indicates the representation but is intended only for illustration.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Location/Species** | Blue Jay | Crow | **…** | Robin | Sparrow |
| Alabama | 0 | 30 |  | 50 | 0 |
| Arkansas | 0 | 15 | 0 | 0 | 50 |
| **…** |  |  |  |  |  |
| Wisconsin | 20 | 10 |  | 0 | 0 |

* 1. To which type of data described in Chapter 2 (Section 2.1.2) is this data most similar?

Document data represented as a document-term matrix.

* 1. Describe the attribute type.

The attributes are discrete, quantitative (ratio). Assuming that the number of birds types is very large and only a few of these bird types visit a specific state, the attribute can also be considered as: asymmetric discrete, quantitative (ratio).

* 1. What proximity measure would you use if you want to find areas that were similar in terms of the percentage of birds of each species? Explain.  
       
     The cosine is appropriate since it is invariant to scaling but not to translation. As an example, consider data with only 4 types of birds and three locations (A, B, C) with bird type frequencies: A: (5, 3, 2, 0); B:(50, 30, 20, 0); C: (10,8,7,5). Locations A and B are highly similar (i.e., percentage of birds of each type is identical for these two locations, and this is reflected in the fact that cosine of the corresponding frequency vectors is 1. Location A and C do not have similar frequency of different bird types, and they will not be considered highly similar by cosine similarity. In contrast, correlation (which is invariant to both scaling and translation) will consider A to be as similar to B as it is to C (correlation amongst all three pairs is 1).
  2. Suppose that you only care about the presence or absence of a bird species at a location. How would you change the data representation, and to which type of data set in Chapter 2 would this correspond?  
       
     Binarize the data so that all non-zero entries become 1. This would correspond to binary transaction (market basket) data.
  3. For which similarity measure and distribution of birds at a location would the similarities of two locations be the same for both data representations?

If the distributions were such that non-zero entries of each location were the same, then cosine similarity would be the same for count and binary data since the cosine measure is invariant to the length of a vector. Correlation would also work in the same case but would not be a preferred measure of similarity for this kind of data (see the answer to c). Let’s take an example to illustrate this.

State1 = 4,0,4,0,4 => 1,0,1,0,1

State2 = 7,7,7,0,0 => 1,1,1,0,0

Cosine(State1, State2) = 0.67 for both representations

Correlation(State1, State2) = 0.167 for both representations

1. You are given data from a study of 1,000 children (500 males, 500 females) in five different cities. For each child, the data consists of measurements of height and weight taken when the children were between the ages of 7 and 18, as well as information about the children’s birthdate, sex, and city. Assume that, for the duration of the study, each child stayed with the same physician and remained in the same city. Finally, all children in the study had their weight and height recorded at least every two years, and the dates of any measurement are always available. (Dates are recorded in terms of the month and year.) However, height and weight measurements were also taken whenever a child came in for a visit, and thus, the number of height and weight measurements can vary from one child to another.

a. What issues do you encounter when you try to represent this data as record data? (Regard the children as the objects.)

There are a variable number of height and weight measurements from one child to another. Thus, this data cannot be represented as standard record data, which assumes a fixed number of attributes.

b. Could the weight (or height) measurements taken by themselves be considered as a traditional time series data set of the same type as a time series of average monthly temperature measured at a specific weather station? Why or why not?

No. Within a set of weight or height measurements for a child, measurements do not necessarily occur at regularly spaced intervals. Also, for any two children, the time measurements do not necessarily occur at the same time.

1. 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.

1. You are involved in a study of 1,000 children (500 males, 500 females) in five different cities. For each child, the data consists of measurements of height and weight taken when they were between the ages of 7 and 18, as well as information about their current age, sex, and the city. Assume that only children who stayed with the same physician and in the same city were involved in the study. Finally, to participate in the study, a child must have had their weight and height recorded at least every two years and that the dates of any measurement are always available. However, height and weight measurements are also taken whenever the child comes in for a visit, and thus, the number of height and weight measurements can vary from one child to another. Dates are recorded in terms of month and year.

Comment on the characteristics and the type of data set that we have based on the discussion in Chapter 2. In particular, answer the following questions. Consider each child as a data object.

a. Why is it difficult to represent this data as record data?

There are a variable number of height and weight measurements from one child to another. Thus, this data cannot be put in as standard record data.

b. Could the weight (or height) measurements taken by themselves be considered as traditional time series data set of the same type as a time series of average monthly temperature measured at a specific weather station? Why or why not?

No. Within a set of weight or height measurements for a child, measurements do not necessarily occur at regularly spaced intervals. Also, for any two children, the time measurements do not necessarily occur at the same time.

c. Could the weight (or height) measurements taken by themselves be considered as a sequence data set of the type considered in traditional association analysis? Why or why not?

No. The values of the measurements are continuous, not binary.

1. A team of researchers is given a large data set consisting of a time series of average monthly temperatures at a million points on the surface of the globe over 20 years. The columns correspond to months, and each row is a time series. The research team analyzes this data using correlation, L1 distance, and Euclidean distance to compare the time series and find which time series are similar to one another. Specifically, they compute the correlation/distance of each time series with every other time series. However, after visualizing some time series in the data set, the team discovers that the columns (the months) in each year of the original data are out of order, i.e., instead of (Jan, Feb, Mar, Apr, May, Jun, Jul, Aug, Sep, Oct, Nov Dec) the order is (Jan, Jul, Feb, Aug, Mar, Sep, Apr, Oct, May, Nov, Jun, Dec). Rows are in proper order.

a) What property of temperature time series data would make it obvious that the months were not in proper order when the time series were plotted?

Temporal autocorrelation. Average monthly temperatures should change gradually over the course of a year. More specifically, time series should have a roughly sinusoidal shape over the course of the year, except perhaps in areas, such as the equator where there are no seasons. If the months are out of order, this will not be the case.

b) The team member who performed the correlation/distance calculations submits a job with the properly ordered months to recalculate the correlations/distances, but after doing so, realizes that this was unnecessary. Why?

None of the three measures – correlation, L1 distance, and Euclidean distance – take the order of the columns into account. Said another way, the formulas for those three measures are invariant to the order of the columns.

# **Aggregation and Sampling**

1. Briefly explain which of the following are true statements about sampling.

a. A sample of 10 objects cannot be representative.

b. The representativeness of a sample depends only on the size of the sample.

c. A sample that is of size 2 \* n is more representative than a sample of size n.

d. None of these.

d. None of these are completely true. A sample of 10 is representative if the data set is small. The representativeness of a sample also depends on the size of the data, the technique used for sampling, and the characteristic that needs to be preserved. Although, in general, larger samples are more representative, a sample of size 2\*n can be less representative than one of size n, just by random chance.

1. The population for a clinical study has the following distribution among 3 distinct groups: 500 of Group A, 1000 of Group B and 500 of Group C. What is a good way of sampling this population to ensure that the distribution of various subpopulations is maintained if only 100 samples have to be chosen? Give the distribution of the various sub-populations in the original sample.

Stratified sampling. The final distribution will be:

500 / 2000 = 25 from Group A

1000 / 2000 = 50 from Group B

500 / 2000 = 25 from Group C

1. Data Reduction

Data reduction – sampling, dimensionality reduction, or selecting a subset of features – is necessary or useful for a wide variety of reasons but can be problematic if the information necessary to the analysis is lost in the process. The following questions explore several issues at a conceptual level.

a) Assume the property of interest is the rate at which a particular event occurs, i.e.,

rate = number of times a particular type of event occurs / total number of all events.

i. If the event occurs at a rate of 0.001, i.e., 0.1% of the time, then what problems, if any, would you encounter in trying to estimate the rate from a single sample of size 100?

The event is unlikely to occur in a single sample, so the estimate would typically and mistakenly be given as 0. If the event did occur once, which would happen infrequently, then estimate would mistakenly be given as 0.01, which is also in error. More generally, small samples do not provide good estimates of rare events.

ii. If the event occurs at a rate of 0.50, i.e., 50% of the time, then what problems, if any, would you encounter in trying to estimate the rate from a single sample of size 100?

Assume the event occurs k times. Since you have only a single sample, your estimate of the rate is k/100. Most of the time, k will be close to 50, so our estimate of the rate will typically be relatively accurate, especially in comparison to the previous situation. The message is that small rates are harder to estimate than larger ones. Many intermediate level general statistics books have a more formal analysis of this result.

Another point to keep in mind is that sampling does not guarantee that the chosen sample reflects the exact distribution of the whole population. Therefore, even if the event has a probability of 50%, the sample might not necessarily reflect that. The deviation varies based on the nature of the event and the sampling method.

b) You are given a data set of 10,000,000 time series, each of which records the temperature of the Earth at a particular location on the surface of the Earth daily for 10 years. The locations are arranged in a regular grid that covers the surface of the Earth. (Details of the exact nature of the grid are unimportant. The important fact is that each point has neighbors to the left and right, up and down.) Note that temperature displays considerable spatial autocorrelation, i.e., the temperature at a given location and time is similar to that of nearby locations and times. The size of the data needs to be reduced so that you can apply your favorite data analysis algorithm. Both aggregation and sampling could be used to reduce the amount of data.

i. If you use aggregation, would you aggregate over location or time or both?

Either or both could be used depending on the algorithm. If the algorithm you want to use is sensitive (performs poorly or has high compute time) to high dimensionality, then aggregating over time could be useful. If the algorithm is sensitive to having lots of objects, then aggregating over space could be helpful. Additional considerations that might drive the aggregation decision are a desire to analyze at a coarser scale and the need to use averaging to reduce the noise level in the data.

ii. How would you use the spatial and temporal autocorrelation of temperature to guide you in aggregating the data?

Because climate data has strong temporal and spatial autocorrelation, aggregating over either to a moderate degree would not lose too much information. For instance, aggregating days into weeks or months might still retain a fair amount of information. However, aggregating months to years would eliminate a great deal of potentially useful detail.

iii. If you use sampling, would you sample over location or time or both?

Again, sampling over both could be useful for the same reasons as given in the answer to (ii).

iv. Would you prefer aggregation or sampling or both? (You can argue any of these as long as you support your answer.)

Either could be used, but aggregation is probably best for many applications for this data.

The most important reason is that the autocorrelation provides a simple way of guiding the aggregation, i.e., aggregate points that are close to each other in space and time. This autocorrelation also means that a more systematic approach than random sampling should be used to ensure a representative sample. Also, comparing time series of different locations would require that the same time periods are sampled for each time series. Also, an average might be a better representative of the original points being aggregated than any sampled time series and would have the advantage of noise reduction and potentially better representation on average. Finally, aggregation in space and time are easily and systematically accomplished by aggregating locations or time periods to a coarser unit of space or time.

1. Consider a data set where the objects are images from a weather satellite, and each image consists of one million pixels. (Assume that each pixel consists of a real value representing the brightness. Also, assume that the images are snapshots of different areas and do not represent images of the same area at successive intervals in time.) The data can be represented as record data, where each image is a record (object), and each pixel is an attribute.

a. Is there any spatial autocorrelation? Explain.  
  
Yes, the images coming from neighboring locations should have spatial auto-correlation, since the weather does not change drastically among the images.

b. An image often has missing values for scattered pixels. (A pixel is missing, but those around it are not.) Which of the three techniques described in the book for handling missing values (p. 46-47) would be the most appropriate for this situation, and why?  
  
Spatial autocorrelation could be used to estimate the value of a missing pixel. For example, the missing pixel might be estimated as an average of the immediately adjacent pixels. This is similar to the approach that estimates a missing value from the values of the same attribute from similar records. However, in this case, the values used for the estimation come from the same record. Thus, it is more like estimating missing values for a time series.

c. Some images are missing large blocks of pixels. Which of the three techniques for handling missing values would be the most appropriate for this situation and why?  
  
Estimation would work only for pixels on the boundary of the missing block, but not for most of the pixels in the block. It would be better to either discard the image, or if the analysis method allows, work only with the pixels that are present.

d. Would any of the following proximity measures – correlation, cosine, or Euclidean - be suitable for computing similarity/distance among images? Explain.   
  
Correlation is most appropriate, although an argument could be made for cosine or Euclidean with appropriate justification. Correlation automatically corrects for differences in the average brightness level of images and the variance of brightness across the pixels in an image, while Euclidean distance and cosine do not unless the data is normalized by removing the mean and dividing by the standard deviation. If images are normalized by subtracting the average brightness, cosine would be equivalent to cosine. Euclidean would be useful for detecting images that were very similar pixel for pixel, such as areas of the ocean or land areas with very similar vegetation or cloud cover. Note that real images similarity calculations would likely be much more sophisticated, for example, to take into account that parts of two images might be similar. Still, two images might not match very well pixel to pixel. Usually, features are extracted in some way, and then images are compared based on these features.

1. Sampling

a. You are given a set of *m* objects that are divided into *K* groups, where the *i*th group is of size *m*i. 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* \* *m*i / *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 (stratified sampling) provides a better representative sample of the data in that the sample has the same proportion of members from each group as in the original data set. An n / m fraction of each group will be randomly picked to create the final sample. On the other hand, for the second scheme, the number of objects from each group will vary. More specifically, the second scheme only guarantees that, on average, the number of objects from each group will be n ∗ mi / m. The first scheme ensures that there is representation from each group and in the right proportion. The second scheme may underrepresent and miss out on some groups altogether if there is a large variation in the sizes of the groups.

b. Mapping fire activity is an important problem for supporting climate and carbon cycle studies as well as forest management. Automated approaches to address this problem use reflectance data collected from satellites orbiting the Earth to learn a classifier that can differentiate burnt areas on the ground from the unburned ones. The algorithms output a probability of fire activity for each pixel on the day the satellite

imagery was taken.

The classification algorithm needs to be trained to recognize the signatures of burned pixels (and how they are different from non-burned ones). Typically, there is a huge imbalance in the number of fire pixels (<0.1% of the total area) and non-fire pixels in any given region. Keeping this in mind, answer the following

(i) What would happen if one were to randomly sample some locations (pixels) for training the classifier ignoring the presence/absence of fire at the location during Sampling?

If we draw a random sample ignoring the class membership of the pixels, the high skew of non-fire pixels will result in most of the training instances being picked from the non-fire class. Due to the lack of sufficient representation in the training set, the classifier wouldn’t be able to learn the right signature of the fire pixels that distinguishes them from the non-fire pixels.

(ii) How would stratified sampling help here?

The input to the classifier must have sufficient representation from both fire and non-fire classes. Since the number of non-fire instances is disproportionately large, picking a random sample would give us little to no representation from the fire class. Instead, we can stratify the instances into two groups – fire and non-fire – and draw a sufficient number of samples from each of them. In this way, the classifier would learn the correct signatures of fire and non-fire instances. Note that a “sufficient number” means enough to have a representative sample from each group. This may mean that we have proportionately more instances from the fire class, which is rare. Sometimes, all of the rare class is used, except for a sample reserved for testing, and then matched with an equal number of instances from the non-rare class. However, one may sometimes wish to sample proportionally more instances from the non-fire class to account for the large variance that might exist.

# **Discretization and Binarization**

1. For each attribute distribution below, state which of the discretization approaches (equal-width versus equal frequency) is more natural. Assume that you are to use 3 bins:

a. Income = (50K, 55K, 55K, 65K, 70K, 80K, 100K, 120K, 950K).

Equal frequency is better because the extreme value, 950K, would distort, equal width.

b. Age = (24, 25, 26, 26, 27, 29, 50, 51, 52, 75, 76, 78).

Equal width is better because each interval has width (78 – 24)/3 = 18, which gives intervals [24,42], (42, 60], and (60,78].

1. Equal width discretization is better than equal depth (equal frequency) discretization for data sets that contain outliers.

False. The bins defined by the equal frequency approach are not affected much by outliers since outliers are small in number and don’t affect the percentiles on which equal frequency bins are based.

1. Equal frequency discretization is always better than equal width.

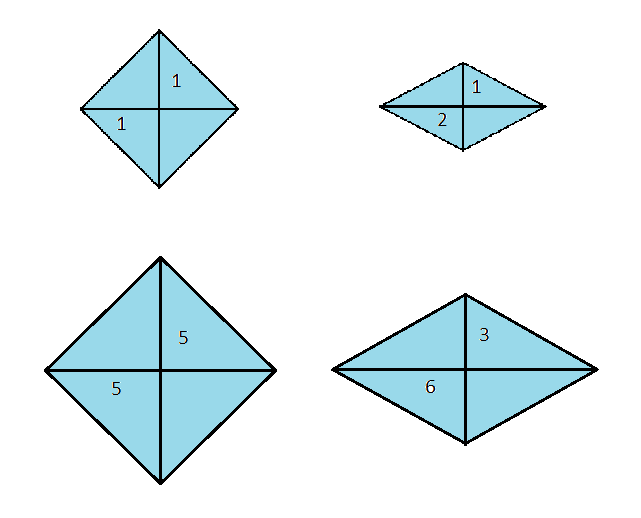
False. Equal frequency discretization is usually, but not always, better than equal width discretization. In certain applications, the data is structured so that discretization by equal width is of interest for the task at hand. For example, consider a variable whose value is the day of the year associated with a particular record, e.g., the record of a sale. The company may be interested in a number of sales by the seasons of the year: winter, spring, summer, fall. In this case, it would be more appropriate to discretize the data by using equal width discretization approach (length of the season).

# **Measures of Similarity and Dissimilarity**

1. Given the four boxes shown in the following figure, answer the following questions. In the diagram, numbers indicate the lengths and widths, and you can consider each box to be a vector of two real numbers, length and width. For example, the top left box would be (2,1), while the bottom right box would be (3,3). Restrict your choices of similarity/distance measure to Euclidean distance and correlation. Briefly explain your choice.
2. Which proximity measure would you use to group the boxes based on their shapes (not size)?  
     
   Cosine. You would have a vector of 2 components, and you would want to be sure that the sides are proportional. For two points of two components, correlation will assign either 1 or -1 since two such points determine a straight line.
3. Which proximity measure would you use to group the boxes based on their size?

Euclidean, since this would take the magnitude of each size into account. If two boxes have sides of the same length, then they would be identical. Notice that this does not detect the situation where one box is a rotated copy of another.

1. Given the four rhombuses shown in the following figure, answer the following questions. In the diagram, numbers indicate the lengths of the diagonals of the rhombuses, and you can consider each box to be a vector of two real numbers, diagonal1 (horizontal) and diagonal2 (vertical). For example, the top left box would be (1,1), while the bottom right box would be (6,3). Restrict your choices of similarity/distance measure to Euclidean distance and Cosine similarity. Briefly explain your choice.



* 1. Which proximity measure would you use to group the boxes based on their shapes (not size)?

We are comparing two rhombuses. The closer the ratio of the diagonals of the two rectangles, the more similar they are. In other words, if two rhombuses are similar and if the length of diagonal1 (diagonal2) of the first rectangle is k times the length of diagnonal1 (diagonal2) of the second.

* 1. Which proximity measure would you use to group the boxes based on their size?  
       
     Euclidean distance, since the distance between larger rhombuses will be smaller, that the distance between a smaller and a larger box. This is because Euclidean distance takes the scale of attributes (diagonal1 and diagonal2) into account. If the diagonals of the rhombuses are similar, their area will be similar as well.

1. Consider time-series coming from sensor readings at 10,000 locations. Each time-series is of length 100,000. It is known that small perturbations occur in the sensor reading all the time, but significant deviations occur only in a handful of time steps (~100). Our job is to identify the top-k closest locations based on the similarity (or dissimilarity) of their sensor readings. One way to do this is to compute the pairwise similarity between time-series across every pair of locations.
2. What is the problem with using Euclidean distance for pairwise similarity in this case?  
   Since the number of dimensions is so large, the small random variations that are actually insignificant will add up to overshadow the significant differences between the two time series, if any. This will result in all pairs of locations having almost the same similarity value, and the resulting top-k computation would be meaningless.
3. How could binarization of the data help here?  
   By selecting the appropriate threshold for binarization, the small random variations would be set to 0, and the significant deviations would be set to 1. Once this is done, one could use an asymmetric similarity measure defined on binary data like the Jaccard coefficient.
4. Let a1, a2, and a3 be three attributes. If correlation(a1, a2) = 0.5 and correlation(a2, a3) = 0.5, then what can you conclude about correlation(a1, a3)?

(A variation is to ask if correlation(a1, a3) > 0.5.)

The desired answer is “not much.” In particular, correlation(a1, a3) could be greater or less than 0.5. The easiest way to analyze this problem is to assume the attributes (vectors) have been normalized by subtracting their mean. Then, correlation is equivalent to the cosine measure. Thus, a cosine measure of 0.5 corresponds to two vectors that have an angle of 60 degrees between them. Three vectors can be arranged in many ways so that two pairs – (a1, a2) and (a2, a3) – have an angle of 60 degrees between them. The resulting angle between (a1, a3) could be as large as 120 or as small as 0. This corresponds to correlations between -0.5 and 1.

1. Distance measures.

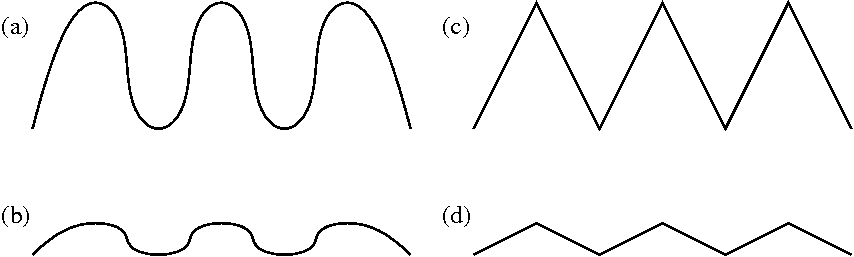
a. Given document data, where the documents are encoded in terms of the frequencies of the constituent words, what issues will be encountered in using the following similarity measures:

i. Correlation

ii. Jaccard

Since document data is sparse (most attributes values are zero for most of the documents), the consideration of matches between these zeros, as included in correlation, is not ideal. On the other hand, Jaccard is for binary attributes, and discretization/binarization of the attributes for using it will lead to a loss of information.

b. Consider the following four examples of time series data.



i. Which of the pairs will be closest according to the Euclidean distance measure? Briefly explain.

(a,c) and (b,d) since Euclidean distance is very dependent on magnitude.

ii. Which of the pairs will be most similar according to the correlation measure? Briefly Explain.

Actually, any pair is highly similar according to correlation, but it would be fine if the answer is given as (a,b) and (c,d).

1. 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 widely varying ranges of values, depending on the scale used to measure them. Furthermore, the attributes are not asymmetric, and the magnitude of an attribute matters. These latter two facts 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. The Mahalanobis distance would be even more suitable since it also accounts for the correlation between attributes.

1. 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 = number of different bits = 3

Jaccard Similarity = number 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.)

The Hamming distance is similar to the SMC. In fact, SMC = (1 - 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; since 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.)

Two human beings share >99.9% of the same genes. If we want to compare the genetic makeup of two human beings, we should focus on their differences. Thus, the Hamming distance is more appropriate in this situation.

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

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

(f) x = (1 0 0 0 0), y = (0 0 0 0 1) Jaccard, Cosine, Euclidean, Correlation

Jaccard Similarity = 0; Cosine = 0; Euclidean = 1.414; Correlation = -0.25;

(g) x = (0 1 1 1 1), y = (1 0 0 0 0) Jaccard, Cosine, Euclidean, Correlation

Jaccard Similarity = 0; Cosine = 0; Euclidean = 2.236; Correlation = -1;

(h) x = (0 1 0 1 0 1), y = (1 0 1 0 1 0) Jaccard, Cosine, Euclidean, Correlation

Jaccard Similarity = 0; Cosine = 0; Euclidean = 2.445; Correlation = -1;

(i) x = (1 1 0 0 0), y = (0 0 0 1 1) Jaccard, Cosine, Euclidean, Correlation

Jaccard Similarity = 0; Cosine = 0; Euclidean = 2; Correlation = -0.6667;

(j) x = (0 1 0 1 1), y = (1 0 1 0 0) Jaccard, Cosine, Euclidean, Correlation

Jaccard Similarity = 0; Cosine = 0; Euclidean = 2.2.2361; Correlation = -1;

(k) x = (0 1 2 4 5 3), y = (5 6 7 9 10 8) Jaccard, Cosine, Euclidean, Correlation

Jaccard Similarity is undefined; Cosine = 0.9304; Euclidean = 12.2474; Correlation = 1;

(l) x = (0, 1, 0, 0, 0), y = (0, 1, 0, 0, 1) Jaccard, Cosine, Euclidean, Correlation

Jaccard = 0.5, Cosine = 0.707; Euclidean = 1; Correlation = 0.612;

(m) x = (0, 1, 1, 2, 2), y = (0, 2, 2, 4, 4) Cosine, Euclidean, Correlation

Jaccard Similarity is undefined; Cosine = 1; Euclidean = 3.16; Correlation = 1;

(n) x = (-1, -1, -1, -1), y = (1, 1, 1, 1): Cosine, Euclidean, Correlation

Jaccard Similarity is undefined; Cosine = -1; Euclidean = 4; Correlation is undefined;

(o) x = (1, 0, 0, 0, 0), y = (0, 0, 0, 0, 1) - Jaccard, Cosine, Euclidean, Correlation

Jaccard Similarity = 0; Cosine = 0; Euclidean = 1.414; Correlation = -0.25;

(p) x = (0, 1, 1, 1, 1), y = (1, 0, 0, 0, 0) - Jaccard, Cosine, Euclidean, Correlation

Jaccard Similarity = 0; Cosine = 0; Euclidean = 2.236; Correlation = -1;

(q) x = (0, 1, 0, 1, 0, 1), y = (1, 0, 1, 0, 1, 0) - Jaccard, Cosine, Euclidean, Correlation

Jaccard Similarity = 0; Cosine = 0; Euclidean = 2.445; Correlation = -1;

(r) x = (1, 1, 0, 0, 0), y = (0, 0, 0, 1, 1) - Jaccard, Cosine, Euclidean, Correlation, Mutual Information

0, 0, 2, -0.6667, 0.42

(s) x = (0, 1, 2, 4, 5, 3), y = (5, 5, 5, 5, 5, 5) - Cosine, Euclidean, Correlation

0.8257, 7.4162, undefined

(t) x = (0, 1, 2, 4, 5, 3), y = (5, 5, 5, 5, 5, 5.0001) - Cosine, Euclidean, Correlation

0.8257, 7.4162, 0.1309

(u) x = (1, 1, 1, 1, 1), y = (1, 1, 1, 1, 10) Cosine, Euclidean, Correlation

Cosine = 0.6139, Euclidean = 9, Correlation is undefined.

1. 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 has only positive entries, and in that 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 two vectors, x and y that have a mean of 0, corr(x, y) = cos(x, y).

1. Consider a database of chemical compounds, where each compound (molecule) is represented as a graph.



**Figure 1:** Graph database.

Each graph contains a set of nodes and undirected edges. A node corresponds to an atom while an edge corresponds to a bond between atoms. Consider the following similarity measure for two chemical compounds, c1 and c2:

*Approach 1 Node-based similarity*

where is a type of atom (carbon, hydrogen, oxygen, etc.), is the number of atoms in the compound, min(*a*,*b*) is a function that returns the minimum value between *a* and *b*, and is a function that returns the number of atoms of type in compound.

*Approach 2 Link-based similarity*

where is a bond between atoms of type and , is the number of bonds in the compound , min(*a*,*b*) is a function that returns the minimum value between *a* and *b*, and is a function that returns the number of bonds in the compound .

1. Suppose according to approach 1. Can the graphs for and be different? State your reasons clearly.

Yes, the graphs can be different since, for this similarity measure to be 1, only the number of each type of atoms in the compounds need to be the same. Consider, for example, the isomers of a compound propanol C3H80 (shown in the figure below).



Figure: Isomers of propanol (Note: Intersecting lines in the above figure correspond to the atom ‘C’), C3H80 [Source: Wikipedia]

While the oxygen atom is attached to the end carbon in one of the isomers, it is attached to the center carbon in another. All isomers have the same chemical formula and same kind of chemical bonds (hence making the node-based similarity 1), but because the atoms are arranged differently, they represent different graphs.

2. Suppose according to approach 2. Can the graphs for and be different? State your reasons clearly.

Yes, the graphs can be different since the only requirement is that they have the same number of bonds between each type of atom pairs. Consider the same example as above, isomers of propanol C3H80. As all isomers have the same kind of chemical bonds, their link-based similarity is 1. However, as the atoms are arranged differently, they represent different graphs.

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

where is the number of documents in which the term appears and 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., .

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.

1. Justify your answers for the following:

a. Is the Jaccard coefficient for two binary strings (i.e., a string of 0s and 1s) always greater than or equal to their cosine similarity?

No. Compare the definition of the cosine measure and the extended Jaccard measure, of which the Jaccard measure is a special case. The cosine and Jaccard measures have the same numerator, but the denominator of the Jaccard coefficient is equal to or larger to the denominator of the cosine measure.

b. The cosine measure can range between [-1,1]. Give an example of a type of data for which the cosine measure will always be non-negative.

For document or transaction data, where all the objects have positive values for their attributes, the cosine measure will be positive.

1. For every item *i* in a grocery store, a set s*i* is used to represent the IDs of transactions in which *i* is purchased. Assume that the data set to be analyzed contains hundreds of thousands of such transactions.

a. To analyze the proximity between any two of these sets *si* and *sj*, which measure, Jaccard or Hamming, would be more appropriate and why?

Jaccard. It captures how frequently both items are purchased together. Jaccard captures proximity based only on the items that are in one or both transactions, whereas Hamming captures proximity based on all items including those that are not in either transaction. Hence two items ({1,0,0}, {0,0,1}) that have no transactions in common can have as much hamming distance as another pair of items ({0,1,1}, {1,1,0}) that have 1/3rd of the transactions that appear in one or both items.

b. Suppose whatever measure not used in part a) was used to analyze the proximity between any two of these sets *si* and *sj* for items *i* and *j*? What would that tell us.

The Hamming measure would tell us the number of transactions in which only one of the two items is bought.

1. The similarity between two undirected graphs G1 and G2 that have the same n vertices can be defined using:

Where indicates the degree of a vertex in graph and indicates the number of edges in *G*.

If , are the two graphs equivalent? Provide an example to justify your answer.

No. For , consider two edges *ab* and *cd*, where *a*, *b*, *c* and *d* are the vertices. Let be a new

graph obtained from by replacing *ab* and *cd* with edges *ad* and *cb*. Now, , since

the degrees of *a*, *b*, *c* and *d* are the same in and , but they are not equivalent.

1. Given a graph, where V is the set of nodes in G, E is the set of edges in G, and W is the set of positive weights assigned to edges in G. For a pair of nodes u and v, consider two graph-based similarity measures:

a. Jaccard similarity (Sim1): , where and are the set of neighbors of *u* and *v,* respectively.

b. Sim2: , where *x* is the length of the shortest path between *u* and *v*, in terms of the total weights of the edges in the path.

Discuss the advantages and disadvantages of each measure.

1*. Advantage:* A conceptually simple similarity measure based on the common neighbors of the nodes.

2. *Advantage:* Efficient to compute: only an operation is needed for computing both the numerator and the denominator.

3. *Disadvantage:* Will produce 0 similarity when two nodes do not have any common neighbors, which can happen for sparse graphs. However, the two nodes may be connected by a path longer than 2, in which case the 0 similarity is not completely accurate.

4. *Disadvantage:* Does not take edge weights into account.

1. *Advantage:* Can measure the similarity when two nodes do not have any common neighbors since it takes the structure of the whole graph into account.

2. *Advantage:* Can take edge weights into account.

3. *Disadvantage:* Inefficient to compute for a pair of nodes. Dijkstra’s algorithm for finding a single pair shortest path has a complexity of , which can be quite high for large graphs.

4. *Disadvantage:* Due to the exponential nature of this measure, the difference between similarity values, say e-6 and e-12, may not be significant, although the lengths of the shortest paths (6 and 12 respectively) are quite different.

1. Similarity measures for a specific application

An instructor gives his class a true/false test and records the answers for each student as a binary vector. The instructor wants to see which students answered the questions in a similar way.

a) Which of the following proximity measures would be appropriate for comparing student tests: the Jaccard measure, the simple matching coefficient (SMC) or the Hamming distance? Briefly explain.

Either the simple matching coefficient or the Hamming distance will work since they both count 1-1 and 0-0 matches. Hamming gives the number of mismatches, while SMC gives the percentage of matches as a fraction between 0 and 1.

b) If the instructor gives a multiple choice test, how could you define a proximity measure to compare student tests?

Each question can be evaluated with respect to whether the answers match or not. Using this result, we can compute a similarity as computed by SMC or a distance as computed by the Hamming distance.

c) Suppose the test consists of binary, categorical, and numerical answers. Would the measure you described for (b) still be appropriate? Briefly explain

Yes, as long as you evaluate the similarity for numerical answers according to match / don’t match.

1. Best similarity measure for documents

Label each of the following similarity measures as good or bad for finding the similarity in document-term data. Provide a one-line justification for each answer you provide.

* + 1. Correlation
    2. Cosine
    3. Euclidean

Because document data is very sparse, computing similarity-based 0-0 matches would make almost all documents very similar. Cosine ignores 0-0 matches, but correlation does not. Also, Euclidean distance does not perform standardization and hence will not do a good job if the scales of the two documents are different, but their trend is the same. Thus, correlation is not appropriate, cosine is appropriate, and Euclidean is not appropriate.

1. Appropriate similarity measure

Decide which of the similarity measures listed in Chapter 2 would be most appropriate for the following situations and why.

1. Suppose two of your friends are numismatists (collecting coins from different countries as a hobby). You also have coins from various countries. You want to decide which friend has the most similar collection to you. Hint: You can represent each collection as a vector of length 196 of the official independent countries of the world, where the corresponding entry denotes the number of coins collected for that country.

Here, we can represent the collection of coins as a vector of length 196, where each entry of that vector represents the number of coins collected for that country. However, it will be a sparse representation where most of the entries will be zero, and thus, we need to discard the zeros during the similarity measure. Any kind of asymmetric measures like Jaccard, extended Jaccard or cosine measure will work for this case. However, Jaccard works only for binary vectors and thus, will not be able to take the frequencies of coins for each country into account. So, extended Jaccard and Cosine will be the most appropriate for this case.

1. Suppose you measure the precipitation level in Minnesota for each zip code every day. The similarity is to be computed between the precipitation levels in Minnesota today and the same day of last month.

Variation: Suppose you measure the precipitation level in 10 widely distributed major cities across the country for each zip code every day. The similarity is to be computed between the precipitation levels in each city today and the same day of last month.

Here, we can represent the precipitation of each day as a vector of length *n*, where *n* is the number of zip codes in Minnesota. There will be a correlation among some of the variables, i.e., those representing nearby zip codes. So, the Mahalanobis distance would be more appropriate.

Answer to variation: Here, we can represent the precipitation of each day as a vector of length *n*, where *n* is the number of locations. There will be little if any correlation among the 10 locations since they are widely distributed, and precipitation is relatively localized. So, simple correlation would suffice. Mahalanobis distance could be used but is not necessary in this case and is much more expensive to compute.

1. A nutritionist wants to measure the similarity between you and your friend based on the following attributes: your height (in meters), weight (in pounds), your dietary requirement (in calories), and your daily activity level (low, medium, high, extreme). Note that the feature set includes continuous and discrete features.

Here, each attribute is of a different type. For example, height and weight are quantitative, and activity level in ordinal variable. So, we have to compute similarity for each attribute separately and then combine them using some weights as described in the book.

1. Consider the following distance measure:

d(*x***,** *y*) = **|** max(*x*) **−**  max(*y*) **|**

where *x* and *y* are real-valued vectors.

(a) Let *x* = [3**,** 4**,** 2**, -**5**,** 1] and *y* = [8**, -**6**,** 9**, -**3**,** 0]. Compute their distance using the above measure.

max(*x*) = 4**,** max(*y*) = 9 **⇒** d (*x***,** *y*) = **|** 4 **−**  9**|**  = 5

(b) State whether the above measure has the following properties.

i. Positive definiteness.

No. Although d (*x*, *y*) ≥ 0 for all *x* and *y*, d(*x*, *y*) could be 0 when *x* ≠ *y* . For example,

*x* = [3, 4, 2, -5, 1], *y* = [2, -6, 4, -3, 0] ⇒ d(*x*, *y*) = 0

ii. Symmetry.

Yes. | max(*x*) **−**  max(*y*)**|**  = **|** max(*y*) **−**  max(*x*)**| ⇒** d (*x***,** *y* ) = d(*y***,** *x*)

iii. Triangle Inequality.

We want to show d(*x*,*z*) ≤ d(*x*,*z*) + d(*y*,*z*), where *x*, *y*, and *z* are real-valued vectors.

d(*x*,*z*) = **|** max(*x*) **−**  max(*z*) **|**

**= |** ( max(*x*) **−**  max(y) ) – ( max(*z*) – max(*y*) ) **|**

≤ **|** ( max(*x*) **−**  max(y) ) | + | ( max(*z*) – max(*y*) ) **|**

≤ d(*x*,*z*) + d(*z*,*y*)

= d(*x*,*z*) + d(y,*z*)

(c) Is **d** (*x***,** *y*) a metric?

No. It lacks the positive definiteness property.

1. The collection of books in a library can be represented by a vector whose length is equal to the number of distinct books available at any library and whose elements indicate the number of copies of that particular book the library owns. You want to compare the similarity of two libraries based on the collection of their books.

Which similarity measure is more suited for this task between cosine and correlation. Briefly justify your answer.

The nature of the data vector is sparse and asymmetric, i.e., only presence should be counted to the similarity measure. So, either extended Jaccard or cosine will be appropriate.

What is one strength and one weakness of Euclidean distance for this task? Briefly justify your answer.

strength - Euclidean can capture the exact similarity in the number of books when both libraries have the same books

weakness - it cannot capture the difference between two scenarios: one where both libraries have a different collection of books; two where both libraries have the same collection of books but in different numbers.

1. Which similarity measure, Jaccard or Simple Matching Coefficient, would you use for a data set of responses to True/False test questions by a set of students if you want to capture the similarity between answers of two students. Briefly justify.

The Simple Matching Coefficient is more appropriate because both True and False answers are equally important and need to be considered by the similarity measure.

1. Consider the following distance measure: d (x, y ) = | min(x − y)|

where *x* and *y* are real-valued vectors and min(*z*) is a function that returns

the minimum value in *z*.

(a) Let *x* = [3**,**  4**,**  2**, -**5] and *y* = [8**, -**6**,**  9**, -**3]. Compute their distance

d (*x***,** *y*) using the above measure.

d (*x***,** *y*) = **|** min(*x* **−** *y*)**|**  = **|** min( **-**5**,**  10**,−** 7**, -**2)**|**  = 7.

(b) State whether the above measure has the following properties. Show

your proofs clearly.

i. Positive definiteness.

No. Although d (*x***,** *y*) **≥**  0 for all *x* and y , d (*x***,** *y*) could be 0

when x ≠ y . For example, if x = [1 0] and y = [1 1], d(x,y)=0.

ii. Symmetry.

No. For example, suppose *x* = [1**,**  2], *y* = [ **-**1**,**  3], then

d (*x***,** *y*) = **|** min(*x* **−** *y*)**|**  = **|** min( 2**, -**1)**|**  = 1

but

d (*y*,*x*) = **|** min(*y* **−** *x*)**|**  = **|** min( -2**,** 1)**|**  = 2

Therefore d (*x***,** *y*) is not necessarily equal to **d** (y**,** x ).

iii. Triangle Inequality.

No. For example, suppose *x* = [3**,**  8], *y* = [3**,**  7], *z* = [1**,**  7], then

d (*x***,** *z*) = **|** min(*x* **−** *z* )**|**  = **|** min(2**,**  1)**|**  = 1

d (x**,** y ) = **|** min(*x* **−** *y* )**|**  = **|** min(0**,**  1)**|**  = 0

d (*y***,** *z*) = **|** min(*y* **−** *z* )**|**  = **|** min(2**,**  0)**|**  = 0

Therefore **d** (x**,** z ) **> d** (x**,** y ) + **d** (y**,** z ).

1. Which similarity or distance measure is most effective for each of the domains given below? Write a brief explanation (1-2 sentences) for your choice:
   1. Which measure, Jaccard or Simple Matching Coefficient, is most appropriate for comparing a data set of responses to True/False test questions by a set of students. For a pair of students, the measure should capture the similarity between the answers of the two students.

SMC. When comparing two students, we need to look at 1-1 matches, 0-0 matches, and 1-0 and 0-1 mismatches.

* 1. Which measure, Jaccard or Simple Matching Coefficient, is most appropriate for comparing a data set of languages spoken by a set of students. For a pair of students, the measure should capture the extent to which they speak the same languages.

Jaccard. Because we want the similarity to be based on the common languages that students speak, not those that they don’t speak.

* 1. Which measure, Jaccard or Simple Matching Coefficient, is most appropriate for comparing a data set of courses taken by a set of students. For a pair of students, the measure should capture the extent to which they take the same courses.

Jaccard, because we want the similarity to be based on the common courses that students take, not those that they didn’t take.

* 1. Which measure, Jaccard or Simple Matching Coefficient, is most appropriate for comparing a data set of the locations visited by tourists at an amusement park. Assume the location information is stored as binary yes/no attributes (yes means a location was visited by the tourist and no means a location has not been visited).

Jaccard. Here places visited by the tourists should play a more significant role in computing similarity than places they did not visit.

* 1. Which measure, Euclidean distance or correlation coefficient, is most appropriate to compare two flows in network traffic. For each flow, we record information about the number of packets transmitted, number of bytes transferred, number of acknowledgments sent, and duration of the session (In other words, assume that the data is represented in record data format, and each flow is recorded as a row in the table).

Euclidean distance (after standardizing each attribute). The correlation coefficient is not meaningful here because it is not meaningful to correlate two flows, which has different attribute values (i.e., correlating the attributes are meaningful but correlating the flows are not).

* 1. Which measure, Euclidean distance or cosine similarity, is most appropriate to compare the coordinates of a moving object in a 2-dimensional space. For example, using GPS data, the object may be located at (31. 4° West,12.4° North) at time t 1 and (29.4° West,12.5 ° North) at another time f2. Note: we may use +/- to indicate East/West or North/South directions when computing the similarity or distance measures.

Euclidean distance. This is because cosine measures the angle of the two locations. Thus, if two locations lie along the same line through the origin, their cosine similarity will be 0 even though they are located far away from each other.

* 1. Which measure, Euclidean distance or cosine similarity, is most appropriate to compare the similarity of items bought by customers at a grocery store. ﻿Assume that the data is represented in document data format, with each customer represented as row, and each item is a column. The presence of items for a given customer will be indicated by 0/1 binary vector (1 if the customer bought the item, otherwise 0).

Cosine similarity because the presence of an item in the transaction plays a more important role in determining similarity than the absence of the item.

1. There is a group of n students. Two n-dimensional vectors A and B record the heights and weights of the students respectively. Consider the variable transformation that is defined by

A' = (A - mean(A)) / std(A) (1)

B' = (B - mean(B)) / std(B) (2)

where std stands for the standard deviation of a vector.

a. What will be the effect of this transformation?

The new vectors will have a mean of zero and a standard deviation of 1. This is known as the z-score transformation

* 1. How will the correlation between A and B change after this transformation?

It will not change. To compute the correlation, we subtract the mean from each observation and divide by the standard deviation. (See definition of correlation.) Thus, correlation is essentially performing a z-score transformation on the data before doing the final computation, which is an inner product of A’ and B.’ Note also that applying a z-score transformation to data that already has a mean of 0 and a standard deviation of 1 does not change the data.

1. Consider the following proximity measure defined on two points (x1,y1) and (x2,y2) in Cartesian space. Is it a distance metric? Explain.

((x1-x2)^3 + (y1–y2)^3)^(1/3)

Not a metric. Symmetry is violated. d((x1,y1), (x2,y2)) = -d((x2,y2), (x1,y1))

1. Under what conditions is a proximity measure called a metric? Which of the following measures are distance metrics? Provide a short proof if it is, or a counterexample if you think it is not.

To be called a metric, a measure should satisfy the conditions of positive definiteness, symmetry, and the triangle inequality.

a. The proximity measure between two integers *x* and *y*: |x2 – y2|

This is not a metric because it takes the value 0 not only if *x* =*y* but also if *x* =-*y*. This violates the positivity property.

b. The proximity measure between two vectors (*x*1,*y*1) and (*x*2,*y*2): (*x*1 × *x*2)2 + (*y*1 – *y*2)2

This is not a metric because it violates the triangle inequality. Consider the three vectors (-2,0), (0,0), (2,0). Distance from (-2,0) to (2,0) is bigger than the distance from (-2,0) to (0,0) PLUS distance from (0,0) to (2,0).

It also violates the positivity property that two points should have 0 distance only if they are identical.

c. Hamming distance between two binary strings of length *n*.

The Hamming distance, by definition, is positive and zero only if the two strings are identical. It is symmetric because the number of 0-1 flips needed to change one binary string to the other is the same as those needed to convert the other string into this one. For triangle inequality, H(*x*,*y*) = minimum number of flips from *x* to *y* ≤ minimum number of flips from *x* to *z* + minimum number of flips from *z* to *y* =   
H(*x*, *z*) + H(*z*, *y*), where H(X,Y) = minimum number of flips from x to y <= minimum number of flips from X to Z + minimum number of flips from Z to Y = H(X,Z) + H(Z,Y).

1. Suppose you were given temperature measurements at several locations across the country over a period of time. (A time series for each city.)

a. What similarity measure (defined on two time series) would you use if you want to know which cities have a similar temperature profile as your current city?

A measure like Euclidean distance that takes the magnitude of absolute differences into account.

b. What similarity measure would you use if you were just interested in comparing the shape of the time series and not absolute temperature differences?

Correlation since we are not interested in absolute magnitudes.

c. In some studies, it might be important to pay attention to just anomalies in the data. Say we transformed the real-valued temperature time series at each location into a binary time series that is 1 if the temperature is anomalous and 0 otherwise. Define a similarity measure that quantifies how similar two locations are with respect to anomalous events.

An asymmetric measure defined on binary data like the Jaccard coefficient.

1. Proximity measure for real data

Suppose that we want to measure the proximity between the climatic conditions of two cities in terms of 3 variables: temperature, air pressure, and humidity, all of which are measured on quite different scales. We want to use this vector of length 3 to compute proximity between locations.

(**Alternative version:** Suppose that we want to compare the metabolic profiles between people in terms of 4 variables: weight, height, body fat% and muscle mass - all of which are measured on quite different scales. We want to use this vector of length 4 to compute the similarity between metabolic profiles between people.)

a. What is the main difficulty with using Euclidean distance in this case?

The attributes have different scales, and this would distort the distance measure.

b) How would you transform the data to remedy the problem in a, that is, to make the data more suitable for Euclidean distance?

We can standardize each attribute using z-normalization so that their ranges are comparable.

c) Temperature, air pressure, and humidity are correlated. What measure discussed in Chapter 2 would handle both the issue in a, as well as the correlation? Briefly explain your answer.

Mahalanobis distance is invariant to the different scales since it scales the difference in each attribute by its respective variance. It can also handle correlation among the variables since it uses a covariance matrix.

1. In order to compute the similarity between two documents:

i. Each document can be represented as a vector of binary features, where each feature is a word of interest in the document. In this vector, a 1 indicates the presence of the word, and a 0 indicates its absence.

ii. Each document can be represented as a vector of term frequencies, which represent the frequency with which each term occurs in the document. (Details of exactly how this is computed are unimportant to this problem.)

Consider the following example in which there are two documents, D1 and D2, containing four words.

Using the representation described in (i), the two documents are denoted as follows:

D1 = (1, 0, 0, 1), D2 = (1, 0, 1, 0).

Using the representation described in (ii), the two documents are denoted as follows:

D1 = (0.5, 0, 0, 0.5), D2 = (0.5, 0, 0.5, 0). The cosine similarity in this case is 0.5.

a) What is the Jaccard similarity using the representation described in approach (i)?

0.33.

b) What is the cosine similarity using the representation described in approach (i)?

0.5

c) Provide an example of a pair of documents for which Jaccard similarity will be larger than cosine similarity.

Cosine can take into account the frequencies of the terms while Jaccard has to work with binary vectors.

In the given example, the term frequencies are assumed to be the same. If you change the frequencies of the words, you would see a different cosine similarity. The Jaccard similarity would remain the same because the word occurrences are not changed, as only the frequencies are changed. The following is one such choice of frequencies when the cosine similarity is smaller than the Jaccard similarity:

D1 = (0.25, 0, 0, 0.75), D2 = (0.25, 0, 0.75, 0)

1. Both the L1 distance and correlation are widely used to compare two time series. The most appropriate measure depends on the specific problem or domain requirements. Decide the proper measure for each of the following scenarios in climate data analysis:

i. We want to compare the temperatures of two cities for 100 days with respect to their levels.

L1 distance because the comparison is focusing on the change of level (difference of absolute value).

ii. We want to compare the temperatures of two cities for 100 days with respect to the patterns (up and down that occur in temperature during the 100 days.

Correlation because the comparison is focusing on the shape/pattern, not the magnitude of the differences.

1. Similarity measures

Which one of the following proximity measures would be least appropriate to calculate the proximity of two transactions in a market basket (transaction) data set? Briefly explain.

a) Hamming distance

b) Cosine

c) Jaccard

The Hamming distance. Because transaction data is very sparse, computing similarity-based 0-0 matches would make almost all transactions very similar. Jaccard and cosine ignore 0-0 matches, but the Hamming distance does not.

1. Which of the two similarity measures, Jaccard or Correlation, would you use for the following data? Briefly explain. (1-2 sentences)

(i) Data set of responses to True/False test questions by a set of students.

Correlation. True/False answers are equally important.

(ii) Data set of courses taken by a set of students. For a pair of students, the measure should capture the extent to which they take the same courses.

Jaccard, because we want the comparison to be based on the extent to which they take the same courses, and not on a comparison of courses they didn’t take.

1. A doctor wants to determine if two subjects in their study have similar proportions in terms of four body measurements: height, arm length, waist, and leg length.

Assume that each person is represented as a vector of four continuous features, where each feature is a length in centimeters. Consider the following two tasks that the doctor wants to perform.

1. **Task:** Find pairs of patients that are as close as possible in each of the four body measurements. The doctor will provide a threshold for closeness.

Which of the following proximity measures would work for this task: Manhattan, Euclidean, or correlation? (You can choose none, one, or more than one.) Briefly explain.

Correlation is not good because it could be high (near 1) even if the absolute differences of the pairs measurements are high, e.g., one set of measures is 25% larger than the other. Either Manhattan or Euclidean could work since they are based on absolute differences. Given the limited description of the problem, there is no reason to prefer one to another.

1. **Task:** Find pairs of patients whose four body measurements are in the same proportion (ratio).

Which of the following proximity measures would work for this task: Manhattan, Euclidean, correlation, or Mahalanobis distance? (You can choose none, one, or more than one.) Briefly explain.

For this type of comparison, neither Manhattan nor Euclidean distance would be appropriate since they are based on absolute differences. Correlation would be best. Mahalanobis distance would not be appropriate since it is a distance-based measure, just one that takes the different scales and correlation of attributes into account.