**Homework 1**

**Question 1 (4 points)**

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*

1. Age as measured by whether the age in years is ≤ 30 (value = 0) or greater than 30 (value = 1).
   * Binary, qualitative, ordinal because 0 means less years than 1 but not necessarily by how much less.
2. Speed of a vehicle measured in mph.
   * Continuous, quantitative, ratio because speed is formulaic and can be measured to an infinitesimally small degree. It is a ratio because 0.0 has a means 0 speed.
3. Intensity of rain as indicated using the values: no rain, intermittent rain, incessant rain.
   * Discrete, qualitative, ordinal because intermittent rain means less than incessant rain but not necessarily by how much less.
4. Different flavors of ice cream.
   * Discrete, qualitative, nominal because the flavors of ice cream have no computational value.

**Question 2 (4 points)**

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

1. House numbers assigned for a given street

* + Discrete, qualitative, ordinal because a house with the number 400 and another house with the number 500 does not mean that they are 100 houses apart, instead it means that the 500 house comes after the 400 house moving away from the 100 house.

2. Your calorie intake per day

* + Continuous, quantitative, ratio because 0 calorie intake has meaning and your actual calorie intake can be measured to an infinitesimally small degree.

3. Shape of a geometric objects commonly found in geometry classes.

* + Discrete, qualitative, nominal because the amount of common geometric objects is finite and the objects have no computational value.

4. 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 because the amount of common geometric objects is finite and the objects have no computational value.

**Question 3 (12 points)**

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 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.*

1. 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?
   1. With a small sample size of only 100, the event cannot be calculated accurately and will either appear once, showing a higher percentage rate or none at all showing a lower percentage rate than 0.1%. A larger sample size is needed to draw correct conclusions.
2. 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?
   1. I would not expect any problems estimating this rate and would expect the rate to be accurate.

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 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.

1. If you use aggregation, would you aggregate over location or time or both?
   1. I would aggregate over location because the data set is so large, and each individual temperature does not tell as much information as groups of locations would tell.
2. How would you use the spatial and temporal autocorrelation of temperature to guide you in aggregating the data?
   1. I would use the spatial and temporal autocorrelation of temperature to group the data and find to what degree each observation at their respective locations were similar.
3. If you use sampling, would you sample over location or time or both?
   1. I would sample over both location and time because both are extremely important in understanding the temperature of Earth. Because Earth is sphere, different areas experience different seasons and different times, and this information must be taken into account, otherwise the data is not very useful.
4. Would you prefer aggregation or sampling or both? (You can argue any of these as long as you support your answer.)
   1. I would support aggregation because there is already a large data set given, this data needs to be grouped and simplified to be useful via aggregation techniques.

**Question 4 (10 points)**

For the following vectors, **x** and **y**, calculate the indicated similarity or distance measure.

1. **x** = (1,1,1,1), **y** = (2,2,2,2) cosine, correlation, Euclidean
   1. **cosine =** ||x|| = 2 ||y|| = 4 and x • y = 8 therefore cos() = 8/8 = 1 and = 0
   2. **Pearson correlation =** 4\*(8)-(4)(8)/= invalid divide by 0
   3. **Euclidean =**
2. **x** = (0,1,0, 1), **y** = (1,0, 1,0) cosine, correlation, Euclidean, Jaccard
   1. **cosine =** ||x|| = ||y|| = and x • y = 0 therefore cos() = 0 and =
   2. **Pearson correlation =** -4/ = -4/4 = -1
   3. **Euclidean =**
   4. **Jaccard =** (0)/(2+2+0) = 0/4 = 0
3. **x** = (0,-1,0, 1), **y** = (1,0,-1,0) cosine, correlation, Euclidean
   1. **cosine =** ||x|| = ||y|| = and x • y = 0 therefore cos() = 0 and =
   2. **Pearson correlation =** 0/ = 0/16 = 0
   3. **Euclidean =**
4. **x** = (1,1,0,1,0,1), **y** = (1,1,1,0,0,1) cosine, correlation, Jaccard
   1. **cosine =** ||x|| = 2 ||y|| = 2 and x • y = 3 therefore cos() = 3/4 and = 0.7227342
   2. **Pearson correlation =** 6\*(3)-(4)(4)/= 2/8 = 0.25
   3. **Jaccard =** (3)/(1+1+3) = 3/5 = 0.6
5. **x** = (2,-1,0,2,0,-3), **y** = (-1, 1,-1,0,0, -1) cosine, correlation
   1. **cosine =** ||x|| = 3 ||y|| = 2 and x • y = 0 therefore cos() = 0 and =
   2. **Pearson correlation =** 6\*(0)-(0)(2)/= 0

**Question 5 (10 points)**

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

1. 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?
   1. I would use Dynamic time warping because it would match similar temperatures between cities at similar times.
2. What similarity measure would you use if you were just interested in comparing trends in the time series and not absolute temperature differences?
   1. I would use Correlation because it captures the up and down movements in the temperatures.
3. 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.
   1. I would use Jaccard coefficient because I am only concerned with the 1’s not the 0’s

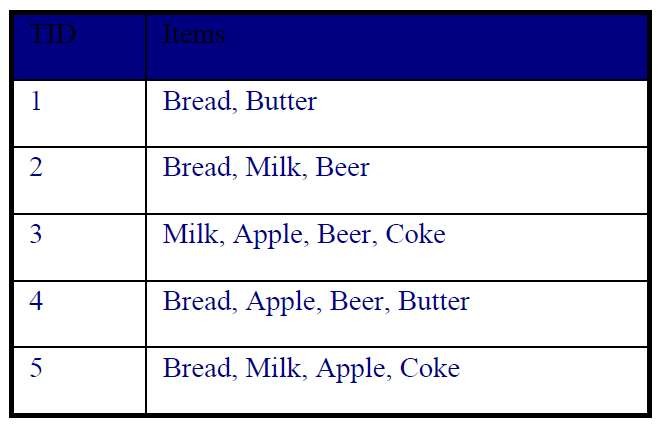
**Question 6 (5 points)**

You are approached by the marketing director of a local company, who believes that he has devised a foolproof 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 bestselling product had the worst satisfaction since it had the most complaints. Could you help me set him straight?"

1. Who is right, the marketing director or his boss? If you answered, his boss, what would you do to fix the measure of satisfaction?
   1. His boss because the number of customer complaints can only be as large as the number of times the item is sold, therefore the bestselling products will likely have the most complaints. To fix the measure, the marketing director should divide the number of complaints by the number of sold products to get a percentage of how many people send complaints after receiving the product. This would be a more comparable measure.
2. What can you say about the attribute type of the original product satisfaction attribute?
   1. The original product satisfaction attribute is discrete, quantitative, and ratio because it’s 0 complaints value has a true meaning and is comparable to other product satisfactions even though the marketing director’s conclusions were false.

**Question 7 (10 points)**

Consider the market basket transactions shown in the table below:



a) What is the maximum number of association rules that can be extracted from this data

(including rules that have zero support)?

6 unique items: Bread, Butter, Milk, Beer, Apple, Coke = 3^6-2^(6+1)+1 = 602

b) What is the maximum size of frequent itemsets that can be extracted (assuming minsup > 0)?

The maximum size itemset is 4 {Apple, Coke, Milk, Bread} because its support is 1/5 which is > minsup.

c) Write an expression for the maximum number of size3 itemsets that can be derived from this data set.

The maximum number of size 3 itemsets = (n+r-1)!/r!(n-1)! Where n is the number of unique objects and r is size of the itemset.

d) What is the support of {Bread}, {Milk}, {Bread, Milk}?

{Bread} = 4/5 = 0.8

{Milk} = 3/5 = 0.6

{Bread,Milk} = 2/5 = 0.4

e) What is the confidence of the rule {Bread} >{Milk} and {Milk}>{ Bread}?

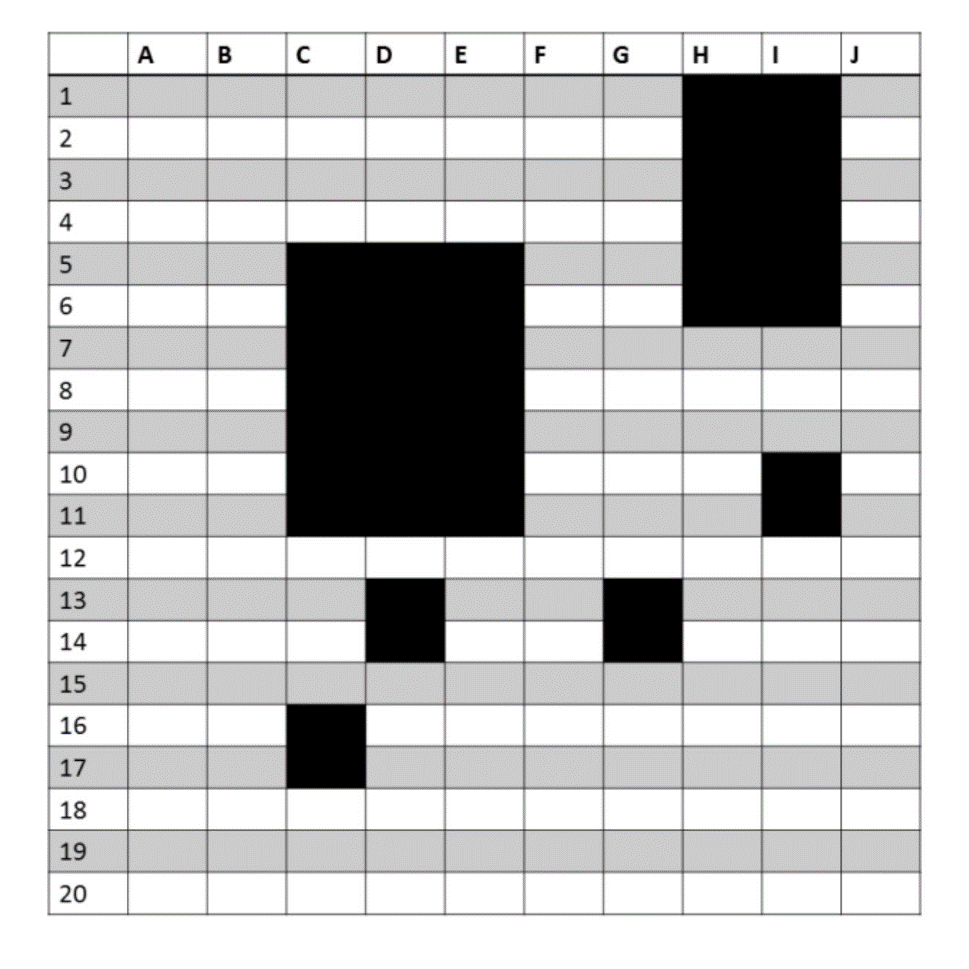
{Bread} >{Milk} = 2/4 = 0.5

{Milk}>{ Bread} = 2/3 = 0.67

f) Comment on the situation where the rules {a} −> {b} and {b} −> {a} have the same confidence and the situation where the rules {a} −> {b} and {b} −> {a} have the different confidence.

These rules have the same confidence if they appear the same number of times throughout the data set. If they do not appear the same number of times, then they will have different confidence values.

**Question 8 (10 points)**

The figure below depicts a transaction matrix with 10 items and 20 transactions. Dark cells indicate the presence of items and white (or grey) cells indicate the absence of items. We apply the Apriori algorithm to extract frequent itemsets with minsup=20% (i.e. itemsets must be contained in at least 4 transactions). Answer the following questions:

1. List all the maximal itemsets in the dataset.

{DG}{HI}

1. List all the frequent itemsets in the dataset.

{C}{D}{E}{H}{I}{CD}{CE}{CI}{DE}{DI}{HI}{CDE}{CDI}{CEI}{DEI}

1. List all the closed frequent itemsets in the dataset.

{C}{D}{E}{H}{I}{CD}{CE}{CI}{DE}{DI}{HI}{CDE}

**Question 9 (16 points)**

For each of the following measures, determine whether it is monotone, anti-monotone, or non-monotone (i.e., neither monotone nor anti-monotone).

Example: Support, is anti-monotone because whenever .

1. A characteristic rule is a rule of the form , where the rule antecedent contains only a single item. An itemset of size k can produce up to k characteristic rules. Let be the minimum confidence of all characteristic rules generated from a given itemset:

Is monotone, anti-monotone, or non-monotone?

is anti–monotone.

1. A discriminant rule is a rule of the form , where the rule consequent contains only a single item. An itemset of size k can produce up to k discriminant rules. Let be the minimum confidence of all discriminant rules generated from a given itemset:

Is monotone, anti-monotone, or non-monotone?

is non-monotone.

1. Repeat the analysis in parts (a) and (b) by replacing the min function with a max function.

is non–monotone and is still non-monotone.

**Question 10 (10 points)**

Table 6.26 shows a 2 x 2 x 2 contingency table for the binary variables A and B at different values of the control variable C.

Table6.26. A Contingency Table.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | | A | |
| 1 | 0 |
| C = 0 | B | 1 | 0 | 15 |
| 0 | 15 | 30 |
| C = 1 | B | 1 | 5 | 0 |
| 0 | 0 | 15 |

(a) Compute the coefficient for A and B when C = 0, C = 1, and C = 0 or 1. Note that

When C = 0, coefficient = -1/3

When C = 1, coefficient = 1

When C = 0 or 1, coefficient = 0

(b) What conclusions can you draw from the above result?

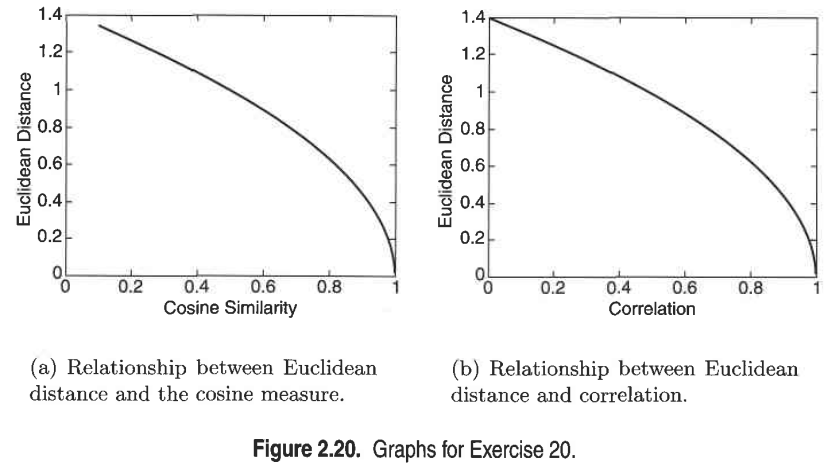
When C = 1, there is a perfect positive correlation between the points, and when C = 0 or 1 there is no correlation.

**Question 11 (10 points, CS 591 only)**

**NOT APPLICABLE**

Here, we further explore the cosine and correlation measures.

1. What is the range of values that are possible for the cosine measure?
2. If two objects have a cosine measure of 1, are they identical? Explain.
3. 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.)
4. Figure 2.20(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?
5. Figure 2.20(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?
6. Derive the mathematical relationship between cosine similarity and Euclidean distance when each data object has an L2 length of 1.
7. Derive the mathematical relationship between correlation and Euclidean distance when each data point has been standardized by subtracting its mean and dividing by its standard deviation.



**Question 12 (5 points)**

Given the network below.



1. Please compute the *closeness centrality* of node A and node D respectively. Which one has higher closeness centrality?
2. Suppose the graph is a road network with nodes corresponding to locations, and edges corresponding to travel costs between locations. Which node between A and D would you use as a service center to minimize total travel costs to other locations?