CSC522 - HW1

[Rutvij Mehta](https://moodle-courses1617.wolfware.ncsu.edu/user/view.php?id=115037&course=3662)(rmehta4), Sagar Gupta(sgupta31) and Tanay Pande(tpande)

Q1. Classify the following attributes as binary, discrete, or continuous. Also classify them as qualitative (nominal or ordinal) or quantitative (interval or ratio). Some case may have more than one interpretation, so briefly indicate your reasoning if you think there may be some ambiguity.

**Soln.**

|  |  |  |
| --- | --- | --- |
| **Attribute** | **Binary /Discrete/ Continuous** | **Reasoning** |
| Time in terms of AM and PM | Discrete | Time has finite values from 0 to 12 |
| Brightness as measured by a light meter | Continuous | Measuring of light can be decimal values and have long range |
| Brightness as measured by people's judgments | Binary | By taking dark as `1` and bright as `0` |
| Temperature as measured in Fahrenheit | Continuous | Can have decimal values and can rise till large values (in Billions) |
| Bronze, Silver, and Gold medals awarded at the Olympics | Discrete | Can have 3 values |
| Height above sea level | Continuous | Can have decimal values and can rise till large values |
| Number of patients in a hospital | Discrete | Have numerical values |
| ISBN numbers for books | Discrete | Fixed range with numerical values |
| Ability to pass light in the terms of the following values | Discrete | Have only 3 values (Opaque-0, Translucent-1, Transparent-2) |
| Military rank | Discrete | Have numerical values |
| Latitude/Longitude values | Continuous | Have a fixed range of values |
| Volume of a substance in cubic centimeter | Continuous | Can have decimal values and no limit |
| Coat check number | Discrete | Have numerical values |

|  |  |  |
| --- | --- | --- |
| **Attribute** | **Qualitative /Quantitative** | **Reasoning** |
| Time in terms of AM and PM | Quantitative (Interval) | There is no `0` time and time can be greater or less than |
| Brightness as measured by a light meter | Quantitative (Ratio) | Measurements can have 0 values and can be nominal, ordinal, interval and ratio |
| Brightness as measured by people's judgments | Qualitative (Nominal) | Have only two values 0 or 1 as we have taken dark and bright |
| Temperature as measured in Fahrenheit | Quantitative (Interval) | There is no `0` temperature in Fahrenheit and we can’t judge half or double hot/cold |
| Bronze, Silver, and Gold medals awarded at the Olympics | Quantitative (Ratio) | If we take number of medals awarded for each category than it can be nominal, ordinal, interval and ratio |
| Height above sea level | Quantitative (Ratio) | Measurements can have 0 values and can be nominal, ordinal, interval and ratio |
| Number of patients in a hospital | Quantitative (Ratio) | Number of patients can be compared by greater/less, twice/half |
| ISBN numbers for books | Quantitative (Interval) | There is no `0` value as they are unique identification values |
| Ability to pass light in the terms of the following values | Qualitative (Ordinal) | As can only be greater or less than 1 |
| Military rank | Qualitative (Ordinal) | As can only be greater or less than lieutenant |
| Latitude/Longitude values | Quantitative (Interval) | Does not have 0 latitude values |
| Volume of a substance in cubic centimeter | Quantitative (Ratio) | Measurements can have 0 values and can be nominal, ordinal, interval and ratio |
| Coat check number | Quantitative (Interval) | There is no `0` value as they are unique identification values |

Q2.

1. Please specify what kind of data transformation needs to be done under the following situations, and give a brief description about your answer:
2. Contrast stretching (often called normalization) is a simple image enhancement technique that attempts to improve the contrast in an image. Suppose our images can be represented in a binary format of 16 bits. But the actual pixel values are distributed between 290 and 3000. Transform (or stretch) these pixel values into full possible range of the binary values.

**Soln.**

Since our Initial Pixel values are n a range of 290 to 3000, we have to perform Normalization for Contrast Stretching them to enhance it to 0 to 65535 range.

Let’s take I as initial dataset, where,

Max = 3000 (Maximum pixel value in the set)

Min = 290 (Minimum pixel value in the set)

Now, to normalize the data,

where, *I= Any random value which we want to normalize*

*NewMax = Maximum intensity value of new Intensity range*

*NewMin = Minimum intensity value of new Intensity range*

*Max = Maximum intensity pixel value of given range*

*Min = Minimum intensity pixel value of given range*

We have,

NewMax = 65,535

NewMin = 0

Max = 3000

Min = 290

Therefor,

For example,

(1) I = 3000

*I’ = 65535*



(2) I = 300

*I’ ≈ 242*

(3) I = 1000

*I’ ≈ 17,170*

In the above problem we learned that Contrast Stretching Normalization is a transformation technique used in Image Processing, in which we enhance or diminish the the pixel intensity of our greyscale image from the given range to desired range. In our problem we have 290 to 3000 as our given pixel range and 0 to 65535 as desired pixel range (As we want to represent in a 16-Bit binary number. So, it’s range is 0-65535). So, we are enhancing the quality of our image by representing it in a 16 bpp image in such a way that every bit of our display is filled by a pixel and making the image clearer.

**(ii)** The domain of variable x is (-∞, ∞), please use a proper function

to map x to z ϵ (0; 1) (Open set). Make sure that z1, z2, z1 > z2 ↔ (if and

only if) x1, x2, x1 > x2

**Soln.**

We are given,

x = (-∞, ∞)

and we have to transform our data values into the range of (0,1)

To perform our transformation, we will use Min-Max Scaling to normalize our data.

In Min-Max-Scaling,



where, xi = value in the range we want to normalize

xmin = minimum value in the range

xmax = maximum value in the range

Let’s take an example,

x = (-100,100)

here,

xmin = -100

xmax = 100

But xi can’t be -100 or 100 as it is an open set.

(1) Take x1=50



z1= 0.75

(2) Take x2=-90



z2 = 0.05

Here, x1>x2 and we get z1>z2. Therefore, condition satisfied and we will get unique values for every value of x.

Q3

1. Describe the advantage and disadvantage of using sampling to reduce the number of data points.

**Soln.**

|  |  |  |
| --- | --- | --- |
| Sr No. | Advantages | Disadvantages |
| 1 | Smaller data size allows the use of a more expensive algorithm. | There can be a bias or error in collecting data |
| 2 | Low cost for data collection | Difficult to select sample that truly represents the data set. |
| 3 | Having correctly collected a sample, the accuracy of data is high. | Data collected from population may not be accurate. |
| 4 | Using a sample data set takes less time. | Using sampling requires knowledge of statistical analysis and probability. Lack of this knowledge may result in inaccurate. |

1. What is simple random sampling (without replacement)? Explain when we should use simple random sampling and when we shouldn't use simple random sampling. Give an example for each situation.

**Soln.**

Simple random sampling without replacement is the process of randomly selecting an item from data population for the sample and removing it from the population.

Example: We have a stock of certain item and we want to check the quality of the batch. For this, we will use simple random sampling, as it is very difficult for our stock to check each and every item.

Simple random sampling should be used when the data population size is very large, as it will help in faster process.

Eg. Using population census data to find average age where the population size is in billions and it is very difficult to calculate using each value. So, we will use simple random sampling and take fewer values, like a few hundred thousand, for calculations.

Simple random sampling shouldn’t be used when the population is varied as it can fail when to create a proper sample set representing a population.

Eg. When building classification models for rare classes, it is critical that the rare classes be adequately represented in the sample. Using population census data to find state wide population density

1. Describe what is noise and what is outlier? Which one is acceptable?

**Soln.**

Outliers are data objects whose characters are significantly different from rest of the data.

Noise is the modification of original value. It may involve the distortion of a value or the addition of spurious data.

Outliers are legitimate data objects and hence may be of use. Hence outliers are acceptable.

Q6

1. Comment briefly on what you have learned about each variable. Include comparisons between x1 and x2 using location measures (mean, median, etc.), spread measures (standard deviation), and the shape of the histograms. Qualitatively, does either variable appear to have come from a normal distribution? Why?

**Soln.**

The mean value of 1st variable is approximately same as medium which means that the distribution is approximately symmetrical.

The mean value of 2nd variable is not close to the median, mean > median, which means that the distribution is negatively skewed.

The standard deviation of 1st variable is 1.46 which means the data is not skewed but the standard deviation of 2nd variable is 5.23 which means data is skewed.

The histogram shape of the first variable shows that data is evenly distributed about the mean that means it is normalized and shape of second variable shows that data is not evenly distributed about the mean and is not normalized.

The first variable values come from a normalized data as the qq-plot of that data is more evenly distributed along the line vector as compared to 2nd variable where the data points are more scattered from the line vector.