CAP 5610 Machine Learning

Homework-1: Data and Data Preprocessing

Name: Lam Nguyen
NID: la815794

Problem 1: Types of Attributes

Classify the following attributes as nominal, ordinal, interval, ratio. **Explain why.**

(a) Rating of an Amazon product by a person on a scale of 1 to 5

This attribute is classified as ORDINAL. Ordinal scale is defined as a variable measurement scale used to depict the order of variables and not the difference between each variable. It maintains descriptional qualities along with an intrinsic order but is void of an origin and the distance between variables can't be calculated.

(b) The Internet Speed

This attribute is classified as RATIO. A ratio variable has all the properties of an interval variable, but the ratio between two measurements make a difference. For example 1 mile is twice as long as .5 miles. Whereas for a birthday, the numbers can't be directly divided to form a meaningful ratio.

(c) Number of customers in a store.

This attribute is classified as RATIO. A ratio variable has all the properties of an interval variable, but the ratio between two measurements make a difference. For example 1 mile is twice as long as .5 miles. Whereas for a birthday, the numbers can't be directly divided to form a meaningful ratio.

(d) UCF Student ID

This attribute is calssified as NOMINAL. It is used to categorize data into distinct groups and there is no particular quantitative value or order.

(e) Distance

This attribute is classified as RATIO. A ratio variable has all the properties of an interval variable, but the ratio between two measurements make a difference. For example 1 mile is twice

as long as .5 miles. Whereas for a birthday, the numbers can't be directly divided to form a meaningful ratio.

(f) Letter grade (A, B, C, D)

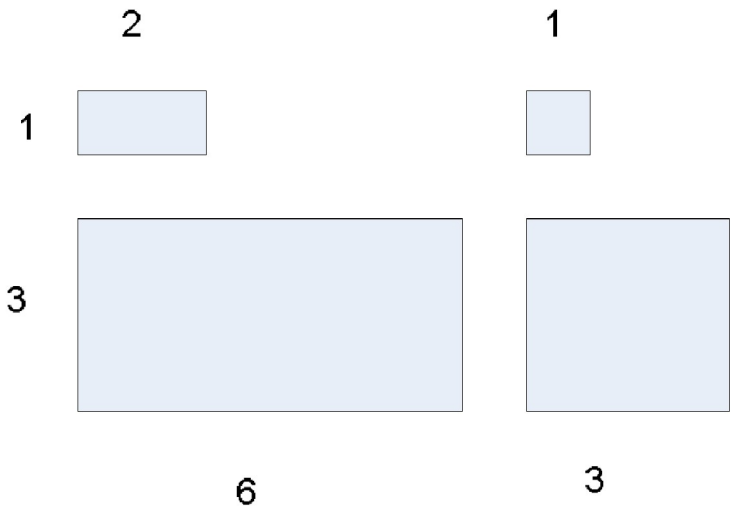
This attribute is classified as ORDINAL. Ordinal scale is defined as a variable measurement scale used to depict the order of variables and not the difference between each variable. IT maintains descriptional qualities along with an intrinsic order but is void of an origin and the distance between variables can't be calculated.

(g) Your birthday

This is classified as an INTERVAL. The order and difference between the birthday is meaningful, but the actual numbers can't be divided into each to form a meaningful ratio. For example dividing a persons birth year of 1989 by another person's birth year of 1998 doesn't yield a meaningful ratio.

Problem 2: Distance/Similarity Measures

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. **Please explain your choice.**



(a) Which proximity measure would you use to group the boxes based on their shapes (length-width ratio)?

CORRELATION would be used since the formula needs the ratio of the sides of the quadrilaterals. The more similar the ratio of the values in the vector, the closer the proportionality of the quadrilateral, which means that the shapes are more similar.

(b) Which proximity measure would you use to group the boxes based on their size?

EUCLIDEAN distance would be used to measure the size because the formula is the distance formula which gives sort of magnitude of the object. The more similar the magnitude, the more similar the size of the boxes.