Task 1 Attributes

Classify the following attributes as binary, discrete, or continuous. Further classify the attributes as nominal, ordinal, interval, ratio.

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

(b) The Internet Speed

(c) Number of customers in a store.

(d) your Student ID

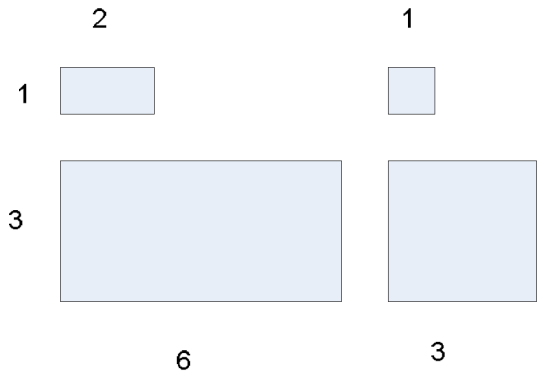
(e) Distance

(f) your letter grade (A, B, C, D)

(g) The temperature in the campus

Task 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. Briefly explain your choice.



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

- Which proximity measure would you use to group the boxes based on their size? Justify your answer.

Task 3 Data Preprocessing of Titanic – Part 1

You can download the Kaggle Titanic dataset from files/data/ Titanic.zip. You can refer to https://www.kaggle.com/c/titanic/data for more details. The data has been split into two groups:

- training set (train.csv)
- test set (test.csv)

The training set should be used to build your machine learning models. For the training set, we provide the outcome (also known as the "ground truth") for each passenger. Your model will be based on "features" like passengers' gender and class. You can also use feature engineering to create new features.
The test set should be used to see how well your model performs on unseen data. For the test set, we do not provide the ground truth for each passenger. It is your job to predict these outcomes. For each passenger in the test set, use the model you trained to predict whether or not they survived the sinking of the Titanic.
We also include gender_submission.csv, a set of predictions that assume all and only female passengers survive, as an example of what a submission file should look like.

Data Dictionary

VariableDefinitionKey survival Survival 0 = No, 1 = Yes pclass Ticket class 1 = 1st, 2 = 2nd, 3 = 3rd sex Sex Age Age in years sibsp # of siblings / spouses aboard the Titanic parch # of parents / children aboard the Titanic ticket Ticket number fare Passenger fare cabin Cabin number embarked Port of Embarkation C = Cherbourg, Q = Queenstown, S = Southampton

Variable Notes

pclass: A proxy for socio-economic status (SES)
1st = Upper
2nd = Middle
3rd = Lower

age: Age is fractional if less than 1. If the age is estimated, is it in the form of xx.5

sibsp: The dataset defines family relations in this way...
Sibling = brother, sister, stepbrother, stepsister
Spouse = husband, wife (mistresses and fiancés were ignored)

parch: The dataset defines family relations in this way...
Parent = mother, father
Child = daughter, son, stepdaughter, stepson
Some children travelled only with a nanny, therefore parch=0 for them.

Let us start with acquiring data: The Python Pandas packages helps us work with our datasets. We start by acquiring the training and testing datasets into Pandas DataFrames. We also combine these datasets to run certain operations on both datasets together.

```
train_df = pd.read_csv('../input/train.csv')
test_df = pd.read_csv('../input/test.csv')
combine = [train_df, test_df]
```

Subtask 1: Analyze by describing data

Q1: Which features are available in the dataset?

Q2: Which features are categorical?

Q3: Which features are numerical?

Q4: Which features are mixed data types?

Q5: Which features contain blank, null or empty values?

Q6: What are the data types (e.g., integer, floats or strings for various features?

Q7: To understand what is the distribution of numerical feature values across the samples, please list the properties (count, mean, std, min, 25% percentile, 50% percentile, 75% percentile, max) of numerical features?

Q8: To understand what is the distribution of categorical features, we define: count is the total number of categorical values per column; unique is the total number of unique categorical values per column; top is the most frequent categorical value; freq is the total number of the most frequent categorical value. Please list the properties (count, unique, top, freq) of categorical features?

Please submit a report (PDF or word) that includes a link to your code, your answers/results, and your explanations or interpretations (if any).