1/29/2020

Chris Tran

university of Central Florida

Machine Learning: Titanic

Task 1 Attributes

Classify the following attributes as binary, discrete, or continuous. Further classify the

attributes as nominal, ordinal, interval, ratio.

1. Rating of an Amazon product by a person on a scale of 1 to 5
   1. Discrete/Ordinal
2. The Internet Speed
   1. Continuous/ ratio
3. Number of customers in a store.
   1. Continuous/ratio
4. your Student ID
   1. discrete/nominal
5. Distance
   1. Continuous/Ratio
6. your letter grade (A, B, C, D)
   1. Discrete/Ordinal
7. The temperature in the campus
   1. Continuous/Interval

Task 2 Distance/Similarity Measures

1. Which proximity measure would you use to group the boxes based on their shapes (length-width ratio)? Justify your answer.
   1. Based on their shapes (rectangles vs squares), I would base their grouping on the length and width ratio because the rectangles have a 2:1 ratio while the squares have a 1:1.
2. Which proximity measure would you use to group the boxes based on their size? Justify your answer.
   1. Based on size, using the width would group the boxes appropriately by splitting the bottom two and the top two since both top boxes have the same width and the bottom boxes also match widths.

Task 3 Data Preprocessing of Titanic – Part 1

Subtask 1: Analyze by describing data

Q1: Which features are available in the dataset?

There are 12 total features available in the dataset once both are combined.

Q2: Which features are categorical?

Cabin, Embarked, Name, Sex, Survived

Q3: Which features are numerical?

Age, Fare,Parch, PassengerId, Pclass, SibSp, Survived

Q4: Which features are mixed data types?

Cabin

Embarked

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

Age 263 Cabin 1014 Embarked 2 Fare 1 Survived 418

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

Age float64

Cabin object

Embarked object

Fare float64

Name object

Parch int64

PassengerId int64

Pclass int64

Sex object

SibSp int64

Survived float64

Ticket object

dtype: object

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?

|  | **Age** | **Fare** | **Parch** | **PassengerId** | **Pclass** | **SibSp** | **Survived** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| count | 1046.000000 | 1308.000000 | 1309.000000 | 1309.000000 | 1309.000000 | 1309.000000 | 891.000000 |
| mean | 29.881138 | 33.295479 | 0.385027 | 655.000000 | 2.294882 | 0.498854 | 0.383838 |
| std | 14.413493 | 51.758668 | 0.865560 | 378.020061 | 0.837836 | 1.041658 | 0.486592 |
| min | 0.170000 | 0.000000 | 0.000000 | 1.000000 | 1.000000 | 0.000000 | 0.000000 |
| 25% | 21.000000 | 7.895800 | 0.000000 | 328.000000 | 2.000000 | 0.000000 | 0.000000 |
| 50% | 28.000000 | 14.454200 | 0.000000 | 655.000000 | 3.000000 | 0.000000 | 0.000000 |
| 75% | 39.000000 | 31.275000 | 0.000000 | 982.000000 | 3.000000 | 1.000000 | 1.000000 |
| max | 80.000000 | 512.329200 | 9.000000 | 1309.000000 | 3.000000 | 8.000000 | 1.000000 |

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 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).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Count | Unique | Top | Freq |
| Age | 1309 | 98 | 22 |  |
| Cabin | 1309 |  | NaN |  |
| Embarked | 1309 | 4 |  |  |
| Fare | 1309 |  |  |  |
| Name | 1309 |  |  |  |
| Parch | 1309 | 8 |  |  |
| PassengerId | 1309 |  |  |  |
| Pclass | 1309 |  |  |  |
| Sex | 1309 | 2 |  |  |
| SibSp | 1309 | 7 |  |  |
| Survived | 1309 |  |  |  |
| Ticket | 1309 |  |  |  |

Github.com/ctran301/ML/HW1