Homework: 1

CS 5402

By: Md Yasin Kabir, email: mkdv6@mst.edu

Github Link: https://github.com/mykabir/CS5402/blob/master/homework1/Homework1.ipynb

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
- Discrete; Ordinal

(b) The Internet Speed
- Continuous; Ratio

(c) Number of customers in a store.
- Discrete; Ratio

(d) MST Student ID
- Discrete; Nominal

(e) Distance
- Continuous; Ratio

(f) MST letter grade (A, B, C, D)
- Discrete; Ordinal

(g) The temperature at Rolla
- Continuous; Interval

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.

- I would like to use Correlation distance to group the boxes. From the 4 boxes we can clearly see that some boxes have the width and length ratio of 1. Those boxes are essentially squars. If I consider this problem as object detection and hence wanted to divide the groups into rectangles and squars, the correlation distance measurement will be helpful. Because correlation is unit independent; if we scale one of the objects multiples times, we will get different Euclidean distances but same correlation distances.

Task 3 Data Preprocessing of Titanic

Subtask 1: Analyze by describing data

Q1: Which features are available in the dataset?

'PassengerId', 'Survived', 'Pclass', 'Name', 'Sex', 'Age', 'SibSp', 'Parch', 'Ticket', 'Fare', 'Cabin', 'Embarked';

Q2: Which features are categorical?

Survived, Sex, Pclass and Embarked.

Q3: Which features are numerical?

Age, SibSp, Parch and Fare.

Q4: Which features are mixed data types?

Ticket and Cabin.

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

- In train: Age, Cabin and Embarked contain empty, null or empty values.
- In test: Age, Fare, and Cabin contain empty, null or empty values.

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

The data types are:

- integers: PassengerID, Survived, Pclass, SibSp, Parch
- floats: Age, Fare
- String/Object: Name, Sex, Ticket, Cabin, Embarked

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?

```
In [6]: train_df.describe()
```

Out[6]:

	PassengerId	Survived	Pclass	Age	SibSp	Parch	Fare
count	891.000000	891.000000	891.000000	714.000000	891.000000	891.000000	891.000000
mean	446.000000	0.383838	2.308642	29.699118	0.523008	0.381594	32.204208
std	257.353842	0.486592	0.836071	14.526497	1.102743	0.806057	49.693429
min	1.000000	0.000000	1.000000	0.420000	0.000000	0.000000	0.000000
25%	223.500000	0.000000	2.000000	20.125000	0.000000	0.000000	7.910400
50%	446.000000	0.000000	3.000000	28.000000	0.000000	0.000000	14.454200
75%	668.500000	1.000000	3.000000	38.000000	1.000000	0.000000	31.000000
max	891.000000	1.000000	3.000000	80.000000	8.000000	6.000000	512.329200

```
In [7]: test_df.describe()
```

Out[7]:

	PassengerId	Pclass	Age	SibSp	Parch	Fare
count	418.000000	418.000000	332.000000	418.000000	418.000000	417.000000
mean	1100.500000	2.265550	30.272590	0.447368	0.392344	35.627188
std	120.810458	0.841838	14.181209	0.896760	0.981429	55.907576
min	892.000000	1.000000	0.170000	0.000000	0.000000	0.000000
25%	996.250000	1.000000	21.000000	0.000000	0.000000	7.895800
50%	1100.500000	3.000000	27.000000	0.000000	0.000000	14.454200
75%	1204.750000	3.000000	39.000000	1.000000	0.000000	31.500000
max	1309.000000	3.000000	76.000000	8.000000	9.000000	512.329200

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?

```
In [8]: train_df.describe(include=["O"])
```

Out[8]:

	Name	Sex	Ticket	Cabin	Embarked
count	891	891	891	204	889
unique	891	2	681	147	3
top	Thorne, Mrs. Gertrude Maybelle	male	347082	B96 B98	S
freq	1	577	7	4	644

```
In [9]: test_df.describe(include=["O"])
```

Out[9]:

	Name	Sex	Ticket	Cabin	Embarked
count	418	418	418	91	418
unique	418	2	363	76	3
top	Stengel, Mr. Charles Emil Henry	male	PC 17608	B57 B59 B63 B66	S
freq	1	266	5	3	270

Subtask 2: Analyze by pivoting features

Q9: Can you observe significant correlation (>0.5) among Pclass=1 and Survived? If Pclas has significant correlation with Survivied, we should include this feature in the predictive model. Based on your computation, will you include this feature in the predictive model?

I can observe significant correlation (0.629630) among Pclass = 1 and Survived. Hence, I will include this feature in the predictive model.

Q10: Are Women (Sex=female) were more likely to have survived?

Female has a correlation of 0.742038 with the Survived which refere they are more likely to survived than male.

Q11:

- Do infants (Age <=4) have high survival rate?
  - Yes
- Do oldest passengers (Age = 80) survive?
  - Yes, the oldest passengers survived.
- Do large number of 15-25 year olds not survive?
  - Unfortunately, the fatality rate is high in 15-20. Most of them wasn't able to survive.
- Should we consider Age in our model training? (If yes, then we should complete the Age feature for null values.)
  - Yes, we shoulod complete the Age feature for null values.
- Should we should band age groups?
  - Yes

Q12:

- Does Pclass=3 have most passengers, however most did not survive?
  - Yes. Although Pclass 3 have most passengers, most of them wasn't able to survive.
- Do infant passengers in Pclass=2 and Pclass=3 mostly survive?
  - Yes.
- Do most passengers in Pclass=1 survive?
  - Yes
- Does Pclass vary in terms of Age distribution of passengers?
  - Yes.
- Should we consider Pclass for model training?
  - Definitely.

Q13:

- Do higher fare paying passengers have better survival?
  - Yes.

- Port of embarkation correlates with survival rates:
  - It's a little bit confusing. Although it seems Embarked C has higher survival rate. However, if I consider the ratio between survived and not-survived it seems similar for S and C.
- Should we consider banding fare feature?
  - Yes, without a doubt.

Q14: What is the rate of duplicates for the Ticket feature? Is there a correlation between Ticket and survival? Should we drop the Ticket feature?

23% of the instances are duplicates in the Ticket feature. Also, there is no correlation between Ticket and Survival. So, we can drop Ticket feature.

Q15: Is the Cabin feature complete? How many null values there are in the Cabin features of the combined dataset of training and test dataset? Should we drop the Cabin feature?

The Cabin feature in not complete. Out of 1309 rows combining both train and test data only 295 rows contain Cabin feature. Among those 186 are unique. The number of null values is: 1014. We should drop the Cabin feature.

Q16: We can convert features which contain strings to numerical values. This is required by most model algorithms. Doing so will also help us in achieving the feature completing goal. In this question, please convert Sex feature to a new feature called Gender where female=1 and male=0.

```
m = {'male' : 0, 'female' : 1}
for df in combine:
   df['Gender'] = df['Sex'].map(m).astype(int)
```

Q17:

```
for df in combine:
   for i in range (0, len(df)):
      if np.isnan(df["Age"][i]) == True:
         df["Age"][i] = np.random.uniform(low=df['Age'].std(), high=df['Age'].median())
```

Q18: Completing a categorical feature: Embarked feature takes S, Q, C values based on port of embarkation. Our training dataset has some missing values. Please simply fill these with the most common occurrences.

```
for df in combine:
   df['Embarked'] = df['Embarked'].fillna(train_df.Embarked.describe().top)
```

Q19: Completing and converting a numeric feature. Please complete the Fare feature for single missing value in test dataset using mode to get the value that occurs most frequently for this feature.

```
test_df['Fare'] = test_df['Fare'].fillna(test_df['Fare'].dropna().median())
```

- I tried to use mode. But mode wasn't working due to the value of 0.

Q20: Convert the Fare feature to ordinal values based on the FareBand

```
train_df['FareBand'] = pd.qcut(train_df['Fare'], 4)
train_df[['FareBand', 'Survived']].groupby(['FareBand']).mean()
val = train_df.FareBand.unique().get_values()
val.sort()
for df in combine:
    for i in range(len(val)):
        df.loc[(df['Fare'] > val[i].left) & (df['Fare'] <= val[i].right), 'Fare'] = i
    df['Fare'] = df['Fare'].astype(int)
```

```
In [26]: train_df.head(10)
```

Out[26]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked	Gender	FareBand
0	1	0	3	Braund, Mr. Owen Harris	male	22.000000	1	0	A/5 21171	0	NaN	S	0	(-0.001, 7.91]
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.000000	1	0	PC 17599	3	C85	C	1	(31.0, 512.329]
2	3	1	3	Heikkinen, Miss. Laina	female	26.000000	0	0	STON/O2. 3101282	1	NaN	S	1	(7.91, 14.454]
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.000000	1	0	113803	3	C123	S	1	(31.0, 512.329]
4	5	0	3	Allen, Mr. William Henry	male	35.000000	0	0	373450	1	NaN	S	0	(7.91, 14.454]
5	6	0	3	Moran, Mr. James	male	14.924794	0	0	330877	1	NaN	Q	0	(7.91, 14.454]
6	7	0	1	McCarthy, Mr. Timothy J	male	54.000000	0	0	17463	3	E46	S	0	(31.0, 512.329]
7	8	0	3	Palsson, Master. Gosta Leonard	male	2.000000	3	1	349909	2	NaN	S	0	(14.454, 31.0]
8	9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27.000000	0	2	347742	1	NaN	S	1	(7.91, 14.454]
9	10	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14.000000	1	0	237736	2	NaN	C	1	(14.454, 31.0]

- **Please visit the following link for the codes and graphs. I used Jupyter notebook for the ease of representation.**
  https://github.com/mykabir/CS5402/blob/master/homework1/Homework1.ipynb