Principles of Data mining

Under the supervision of: Dr. E. Nazerfard

Assignment 1

Winter 2020

1- Explain the following terms

A. Supervised
B. Unsupervised
C. Semi Supervised
D. Outlier
F. Missing Value. Also, Discuss Missing Value Vs. Noise

2- In real-world data, tuples with missing values for some attributes are a common occurrence. Describe various methods for handling this problem.

3- A database has 5 transactions. Let min sup = 60% and min conf = 80%.

TID	*items_bought*
T100	{M, O, N, K, E, Y}
T200	{D, O, N, K, E, Y }
T300	{M, A, K, E}
T400	{M, U, C, K, Y}
T500	{C, O, O, K, I ,E}

Find all frequent itemsets using Apriori and FP-growth, respectively. Compare the efficiency of the two mining processes.

4 - Consider the market basket transactions shown in Table.

Transaction ID	Items Bought
1	{Milk, Beer, Diapers}
2	{Bread, Butter, Milk}
3	{Milk, Diapers, Cookies}
4	{Bread, Butter, Cookies}
5	{Beer, Cookies, Diapers}
6	{Milk, Diapers, Bread, Butter}
7	{Bread, Butter, Diapers}
8	{Beer, Diapers}
9	{Milk, Diapers, Bread, Butter}
10	{Beer, Cookies}

A. What is the maximum number of association rules that can be extracted from this data (including rules that have zero support)?

B. What is the maximum size of frequent itemsets that can be extracted (assuming minsup > 0)?

C. Write an expression for the maximum number of size-3 itemsets that can be derived from this data set.

D. Find an itemset (of size 2 or larger) that has the largest support.

E. Find a pair of items, a and b, such that the rules {a} → {b} and {b} → {a} have the same confidence.

5- What are the purposes of Data Reduction? Explain its strategies.

6- The correlation between two variables is Zero. What does it mean? Are these variables independent?

7 - Suppose that the data mining task is to cluster the following eight points (with (x, y) representing location) into three clusters.

A1(2, 10), A2(2, 5), A3(8, 4), B1(5, 8), B2(7, 5), B3(6, 4), C1(1, 2), C2(4, 9).

The distance function is Euclidean distance. Suppose initially we assign A1, B1, and C1 as the center of each cluster, respectively. Use the k-means algorithm to show:

(a) the three cluster centers after the first round execution

(b) the final three clusters

8- Use an example to show why the k-means algorithm may not find the global optimum, that is, optimizing the within-cluster variation.

Implementation

Goal of this assignment:
learning data analysis methods and working with python libraries

- Pandas
- Numpy
- Matplotlib

At first you need to install python on your system.
If you use the python of your system, you can use “pip” command for installation of libraries and packages.
A better way is to use anaconda. Anaconda is a free and open-source distribution of the Python and R programming languages for scientific computing.

Task:

A. You’ll see a csv file named: ”data” in assignment’s folder which is the epidemiological data of COVID-19 patients in South Korea. Read this file and display it in a table.

B. Describe data by looking into the columns and rows. Print size of data & names of columns.

C. Use python libraries to show **max, mean, and std** of birth_year in a table.

D. Choose one column and show number of fields in it.

E. Check if it is ‘null’ value in data. if yes, use appropriate methods to handle them.

F. Consider an attribute (like birth_year), and show distribution of the data in a histogram. This is called ‘visualization’.Try other visualization methods, too. For more information see:

   - https://machinelearningmastery.com/data-visualization-methods-in-python/

   implement these methods(scatter plot , ...) using different attributes and report the result. Explain completely at each step in your report.

G. Check if it exists ‘outlier’ in data. if yes, explain why it is outlier & decide what should be done with that.
   You can get help from the previous part.

CAUTION:

- For each part, write your codes in a .py file naming with the number of parts, and put it in the "supporting material" folder.
- Deadline is on 27 March 2019 (**8 Farvardin**) and you will lose 10% of your grade after that on each day of delay.
- Report is an important part of your grade. So write it completely and explain your analysis. Your report is only accepted in 'pdf' format. Put it in "report" folder. (There is no force on the language of the report)
- Your codes should be written in python or R. Put them in "supporting material" folder.
- Put all your folders and files like the sample format in a "zip" file and upload it on moodle (http://courses.aut.ac.ir/)
- If you have any question regarding the assignment contact s.sadeghpour97@gmail.com

Please upload your homework in this format:

```
9******_FirstnameLastname_HW1.zip
├── [directory] Report
│   └── 9******_FirstnameLastname_Report1.pdf
└── [directory] Supporting_Material
    └── codes.py
```

Good luck