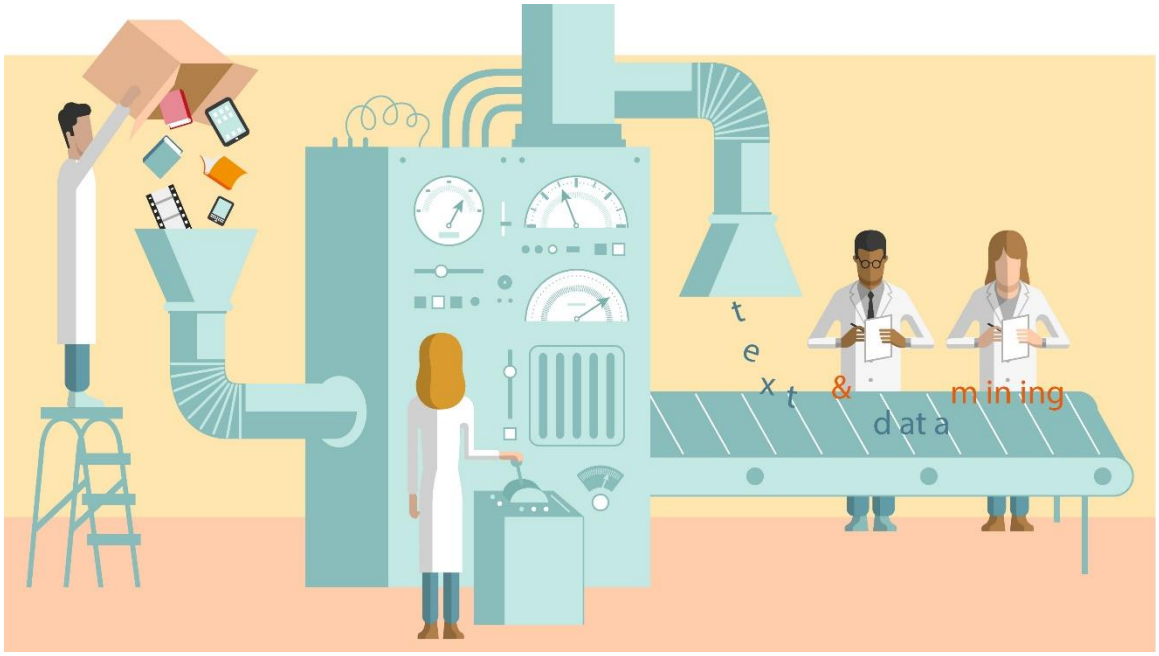
DEPARTMENT OF COMPUTER ENGINEERING AND IT

AMIRKABIR UNIVERSITY OF TECHNOLOGY

Principles of Data mining



Under supervision of:
Dr.E.Nazerfard

Assignment 1

Winter 2019

Theory

1)

This Dataset consists of 9 transactions.

Which association rules could be find in it?

Write all frequent itemsets, explain your solution.

Minimum support = 22%

Minimum confidence =70%

TID	List of item_IDs
T100	I1, I2, I5
T200	I2, I4
T300	I2, I3
T400	I1, I2, I4
T500	I1, I3
T600	I2, I3
T700	I1, I3
T800	I1, I2, I3, I5
T900	I1, I2, I3

2) What are the advantages and disadvantages of Apriori and FP-growth?

3) Explain the following terms

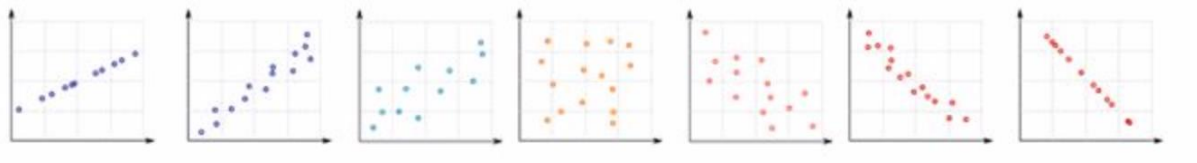
a. Supervised
b. Unsupervised
c. Semi Supervised
d. Outlier
e. Noisy Dataset
f. Missing Value. Also, Discuss Missing Value Vs. Noise
g. What is Correlation ?

4) What are the purposes of Data Reduction? Explain its strategies.

5) Suppose 7 difference plots are given (left to right). Each of these plots is based on two different random variables.

A. Compare correlation value (i.e. sign of the correlation value is not important) between variables of each plot. Write your answer in ascending order.

B. In which plot random variables are most related? Explain your answer



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

Implementation

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

- Pandas
- Numpy
- Matplotlib
- Scikit-learn
- SciPy

At first you need to install python(2 or 3) on your system.
If you use the python of your system, you can use "pip" command for installation of libraries and packages.
Better way is to use anaconda. Anaconda is a free and open-source distribution of the Python and R programming languages for scientific computing.

By installing conda, you can use jupyter notebook for writing and running your codes, too.

Download:

- https://www.anaconda.com/distribution/

for installation:

- Windows
- linux
- Mac

Task

A. You'll see a csv file named:"data" in assignment's folder. 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 and rows.
C. In column "Amount", delete all '$' symbols.
   Convert string fields in this column to numeric.
   Update the column.
   Display new data.
D. Look at the data carefully. One column looks unnecessary. Find and drop it. Update the table and display it.
E. Get the new size of the data. use python libraries to show **max, mean, average and std** of data in a table.
F. Choose one column and show number of fields in it.
G. Check if it is 'null' value in data. if yes, drop it.
H. Consider an attribute(like amount) , 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 , box plot , ….) using different attributes and report the result.

   Explain completely at each step in your report.

I. 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 15 March 2019(24 esfand) 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. 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(https://ceit.aut.ac.ir/courses/)
- Please upload your homework in this format:

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

Good luck