Assignment 1 Problems

Advanced Data Mining : Winter 1401 : Dr. Minaei
Due Monday, Esfand 29, 1401

Farbod Davoodi
Faezeh Sadeghi

Problem 1

Indicate with reasons if each of the following tasks is a Data Mining task or not. If they are, classify what kind of a Data Mining task it is:

1. Dividing a company's customers by gender.
2. Dividing the customers of a company according to their profitability.
3. Count how many emails are tagged with spam.
4. Calculating the total sales of a company.
5. Predicting the future stock price of a company using past records.
6. Find the answer to each question by analyzing the corresponding image.
7. Monitoring the heart rate of a patient for abnormalities.
8. Monitoring seismic waves for earthquake activities.
9. Determining the association rules of market transactions.
10. Summarization of arguments about a certain topic.
11. Sorting a student database based on student identification numbers.
12. Predicting the outcomes of tossing a (fair) pair of dice.
13. Extracting the frequencies of a sound wave.
14. Save the identification numbers in a database.
15. Look up the phone number in the phone directory.
16. Query a Web search engine for information about "Iran University of Science and Technology".
17. Certain names are more prevalent in certain US locations.
18. Group together similar documents returned by the search engines according to their context.

Problem 2

Classify the following attributes as binary, discrete, or continuous. Also, classify them with reasons as qualitative (nominal or ordinal) or quantitative (interval or ratio). Some cases may have more than one interpretation, so briefly indicate your reasoning if you think there may be some ambiguity.

Example: Age in years. Answer: Discrete, quantitative, ratio because it has all four properties.

a. Time in terms of AM or PM.
b. Brightness as measured by a light meter.
c. Brightness as measured by people's judgments.
d. Satisfaction of customers. (Unsatisfied, Neutral, Satisfied)
e. Bronze, Silver, and Gold medals as awarded at the Olympics.
f. Height of a person. (tall, short)
g. Height of a person as measured by centimeters.
h. The temperature of a home. (cold, hot, warm)
i. The temperature of a home as measured by a thermometer.

j. Military rank.
k. University rank.
l. Angles as measured in degrees between 0◦ and 360◦.
m. Height above sea level.
n. Number of patients in a hospital.
o. ISBN numbers for books. (Look up the format on the Web.)
p. Ability to pass light in terms of the following values: opaque, translucent, transparent.
q. Distance from the center of campus.
r. Density of a substance in grams per cubic centimeter.
s. Coat check number. (When you attend an event, you can often give your coat to someone who, in turn, gives you a number that you can use to claim your coat when you leave.)

Problem 3

What is the difference between the definition of classification and clustering? Name four applications of classification and two applications of clustering with their goal and approach.

Problem 4

Name all types of sampling and explain their advantages and disadvantages. Is random sampling without replacement a good method? Explain your answer.

Problem 5

One of the most important steps in data mining that takes a lot of time from users is data preprocessing. In the exercise, we are going to pre-process and visualize the data. For this purpose, the Corona disease dataset "owid-covid-data.csv" has been considered for this exercise, and you can download this dataset from the link below.

https://github.com/owid/covid-19-data/tree/master/public/data

Answer each of the following questions according to the Corona disease dataset:

1. One of the steps taken at the beginning of all data analysis projects is data quality assessment, during which a general view of the ratio of fields and their values is obtained and familiarity with the data takes place.

Calculate each of the following for 10 arbitrary fields from the Corona disease dataset.

a. Number of rows without value (Null)
b. Maximum and minimum values of each column along with its country name
c. Median and mean of each column
d. Checking the invalid values of each column (negative data and out of range, etc.)

2. Draw a bar chart for the values of new_cases and new_deaths columns of Iran in daily, weekly and monthly intervals.

3. In the new_cases column of France, negative values can be seen, assuming that these data are errors, what is your best suggestion to deal with these data? State two methods of finding missing data (missing values), implement one method. Create and report suggested values to replace these data.

4.

a. For the values of the new_cases column, draw a box whisker diagram of the data of Iran and two neighboring countries of Iran and two European countries in one diagram. Note that the middle is also visible. As in Figure 2:

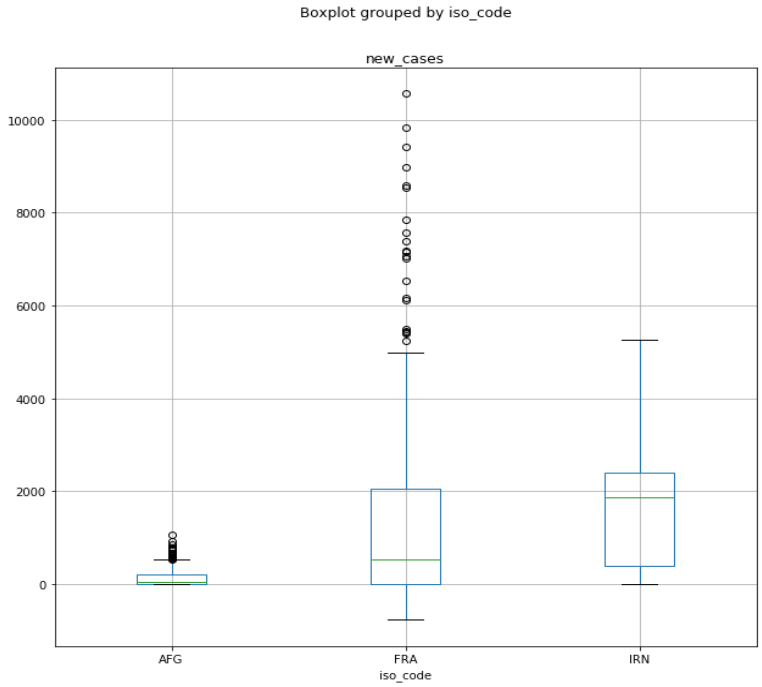

Figure 1: box whisker diagram of countries Iran, France and Afghanistan

b. Calculate and report the value of Q1, Q3, IQR, top whisker and bottom whisker for the new_cases column values of Iran.
c. Find the 10 most outlier data using box whisker plot. The most outlier data means those that have the greatest distance from the top whisker and bottom whisker.

Problem 6

After getting familiar with the ARMA model, in this section we are trying to implement ARIMA using *python* and ***statsmodels*** library and use it to predict the probable distance that a person is going to cover in the upcoming days. The steps to perform the task are explained below.

1. Data Preprocessing

In this exercise, the dataset includes latitude and longitude, time, and an accumulated distance for each row. Specifically, we intend to use the available information to predict the distance traveled in future days.

1.1. Importing data

- Read the data from the file.
- Perform necessary preprocessing on the data.
- Since the distance in the dataset is stored cumulatively, it is necessary to calculate the distance for each day so that we can use the date and distance of that day to calculate the distance of future days using the ARIMA model.

2. Determining the Parameters

- In ARIMA, we have three parameters p, d and q.
- Using the available dataset, estimate these parameters.

3. Model Training

In this step, train the ARIMA model using the estimated values for the p, q and d.

4. Plotting the Model and Analysis

- Using matplotlib, plot and analyze the obtained models.

- Explain how the difference in the value of the three mentioned parameters affects the output and what is the reason for it.

Notes

- If you have any questions, feel free to ask. You can ask your questions in the Telegram group.
- Please upload your assignments as a zipped folder with all necessary components. Upload your file in HW1-ADM-YourStudentID-YourName.zip format.