



JAN 27, 2020


COSI126A_HW0

CHUYUE WU

INSTRUCTOR: DR. HONGFU LIU

SCHOOL: BRANDEIS UNIVERSITY

Course Name: Introduction to Data Mining



Problem 1 (9 points)

Discuss whether or not each of the following activities is a data mining task.

- (A) Dividing the customers of a company according to their gender.

No. Because data mining is non-trivial extraction of implicit, previously unknown and potentially useful information from data, in this case, the customers or each gender are already known.

- (B) Dividing the customers of a company according to their profitability.

No. Because data mining is non-trivial extraction of implicit, previously unknown and potentially useful information from data, in this case, the profitability of customers is already known.

- (C) Computing the total sales of company.

No. Because the total sales of company are easy to know by using sum function.

- (D) Sorting a student database based on student identification numbers.

No. Because there is no potential useful information to be found in this case.

- (E) Predicting the outcomes of tossing a fair pair of dice.

Yes. Because the outcomes of tossing are what we could calculate by probability theory. Predicting is a typical task in data mining.

- (F) Predicting the future stock price of a company using historical records.

Yes. Because we need to use regression model to figure it out. Predicting and regression is the typical tasks in data mining.

- (G) Monitoring the heart rate of a patient for abnormalities.

No. Monitoring means there is no need to extract some implicit, previously unknown and potentially useful information from data. It's obviously.

- (H) Monitoring seismic waves for earthquake activities.

No. Monitoring means there is no need to extract some implicit, previously unknown and potentially useful information from data. The results are shown obviously.

(I) Extracting the frequencies of a sound wave.

No. Because it is not something unknown needed to be explored, it's specific task needed to be calculated.

Problem 2 (10 points)

Clustering:

Clustering could help search engine company display the results that contain not only the keyword which users input, but also related results.

For example, we can use clustering method to cluster the climate patterns in an area of different years. It can help people have a clearer expectation to the temperature and precipitation they are suffering now or in the future.

Classification:

Classification is the process of finding a set of functions that describe and distinguish data classes or concepts, and using this function to predict the class of object whose class label is unknown. Classification analyzes class-labeled data objects whereas clustering analyzes data objects without consulting a known class label. This is more of an internal implementation.

For example, when we train the classification model with 10000 face figures with “young”, “adult”, “old” labels. Then we can use this model to predict the age label of face figures without labels.

Association rule mining:

Association rule mining is the discovery of association rules showing attribute-value conditions that occur frequently together in a given set of data. A search engine could append additional information in its result based on the keywords entered by the user.

For example, if a user typed “swimming ring” in the search engine, it would also display some words such as “swimsuit”, “swimming goggles”, “swimming cap” as a result by customers’ historical purchase behavior.

Anomaly detection:

Anomalies are the data objects that do not conform to the general behavior of the data. The analysis of anomalies is known as anomaly detection.

For example, credit card can use anomalies to detect abnormal purchase behavior and remand the card host to check their transaction record.

Problem 3 (10 points)

For each of the following data sets, explain whether or not data privacy is an important issue.

(A) Census data collected from 1900-1950.

Data privacy **is not** an important issue in this case. Because 1900 – 1950 is too long, census from over 60 years ago is not too important.

(B) IP addresses and visit times of Web users who visit your Website.

Data privacy **is** an important issue in this case. Because by IP address, personal information could be found. Sometimes people visit a website without wanting others to know it such as sex website.

(C) Images from Earth-orbiting satellites.

Data privacy **is** an important issue in this case. Because some images in important area such Military Bases would make people hard to .

(D) Names and addresses of people from the telephone book.

Data privacy **is** an important issue in this case. Because information in telephone book is private.

(E) Names and email addresses collected from the Web.

Data privacy **is not** an important issue. Because once your names and email addresses have been posted on the website, it means they are open to everyone.

Problem 4 (15 points)

Matrix $A = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 4 & 6 \\ 3 & 6 & 9 \end{bmatrix}$, calculate A^{-1} , A^{+1} , A^{100} .

(1) In this case there is no A^{-1} , because $|A| = 0$.

(2)

```
17 {r}
18 ginv(a)
19
```

	[,1]	[,2]	[,3]
[1,]	0.005102041	0.01020408	0.01530612
[2,]	0.010204082	0.02040816	0.03061224
[3,]	0.015306122	0.03061224	0.04591837

(3)

```
20 {r}
21 a ** 100
22
```

	[,1]	[,2]	[,3]
[1,]	1.000000e+00	1.267651e+30	5.153775e+47
[2,]	1.267651e+30	1.606938e+60	6.533186e+77
[3,]	5.153775e+47	6.533186e+77	2.656140e+95

Another Method:

$A = a \times b$, where $a = (1 \ 2 \ 3)^{-1}$, and $b = (1 \ 2 \ 3)$.

$$A^{100} = (a \times b)^{100}$$

$$= a \times b \times a \times b \cdots 100 \text{ times}$$

$$= a \times (b \times a) \times (b \times a) \times \cdots \times b$$

$$= 14^{99} a \times b = 14^{99} A$$

Problem 5 (14 points)

Assume there three students, X, Y, Z. Only one of them gets a score A+. X asks Teacher if he gets A+. Teacher refuses to tell X his score. Instead, Teacher says that Y does not get A+. Calculate $P(Z \text{ gets } A+)$

$$P(Z \text{ get } A+) = P(Z \text{ get } A+ | X \text{ get } A+) P(X \text{ get } A+) + P(Z \text{ get } A+ | X \text{ NOT get } A+) P(X \text{ NOT get } A+) = 0 \times 1/3 + 1 \times 2/3 = 2/3$$

Problem 6 (14 points)

There are two kinds of products in a warehouse, A and B. The percentage of A is 70%, B is 30%. The probability of substandard products in A is $P(A = \text{sub}) = 2.5\%$, for B, it's $P(B = \text{sub}) = 5\%$. Warehouse tests 4 products and one of them is substandard. What is the probability that this product is from A?

Choosing one which is substandard: $P = 0.7 * 0.025 + 0.3 * 0.05 = 0.0325$

Choosing one which is not substandard: $P = 1 - (0.7 * 0.025 + 0.3 * 0.05) = 1 - 0.0325 = 0.9675$

P (this product is from A)

= P (from A and one out of four is substandard) / P (one out of four is substandard)

= $(4 * 0.7 * 0.025 * 0.9675 ** 3) / (4 * 0.0325 * 0.9675 ** 3)$

= 0.5384615

Problem 7 (14 points)

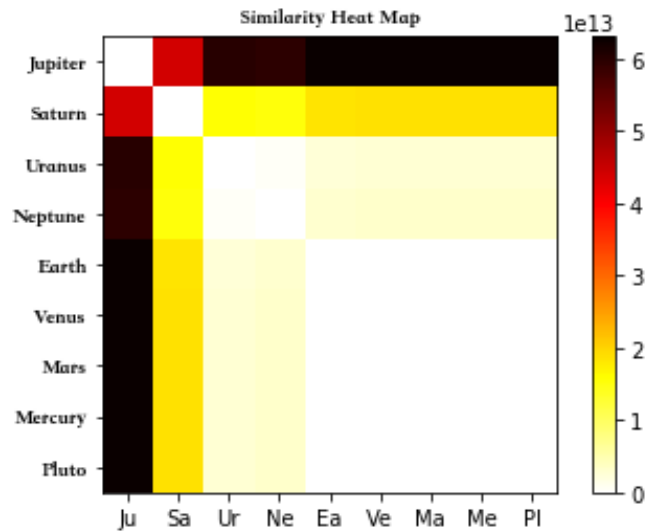
Calculate the similarity matrix between 9 planets. The data of planets is in Table 1. You can use $s(p_1, p_2) = \sqrt{a_0(d_1 - d_2)^2 + a_1(r_1 - r_2)^2 + a_2(m_1 - m_2)^2}$ as the metric, where $a_0 = 3.5 * 10^{-7}$, $a_1 = 1.6 * 10^{-5}$, $a_2 = 1.1 * 10^{-27}$. Set a threshold to separate 9 planets into different groups. What is the relationship between threshold and groups.

Table 1: Data of Nine Planets

Planet p	Distance to Sun (km) d	Radius (km) r	Mass (kg) m
Jupiter	778000	71492	1.90e27
Saturn	1429000	60268	5.69e26
Uranus	2870990	25559	8.69e25
Neptune	4504300	24764	1.02e26
Earth	149600	6378	5.98e24
Venus	108200	6052	4.87e24
Mars	227940	3398	6.42e23
Mercury	57910	2439	3.30e23
Pluto	5913520	1160	1.32e22

	0	1	2	3	4	5	6	7	8
0	0	4.41443e+13	6.01337e+13	5.96329e+13	6.28175e+13	6.28544e+13	6.29946e+13	6.30049e+13	6.30154e+13
1	4.41443e+13	0	1.59894e+13	1.54886e+13	1.86733e+13	1.87101e+13	1.88503e+13	1.88607e+13	1.88712e+13
2	6.01337e+13	1.59894e+13	0	5.0081e+11	2.68381e+12	2.72063e+12	2.86085e+12	2.8712e+12	2.88171e+12
3	5.96329e+13	1.54886e+13	5.0081e+11	0	3.18462e+12	3.22144e+12	3.36166e+12	3.37201e+12	3.38252e+12
4	6.28175e+13	1.86733e+13	2.68381e+12	3.18462e+12	0	3.68145e+10	1.77041e+11	1.87389e+11	1.97896e+11
5	6.28544e+13	1.87101e+13	2.72063e+12	3.22144e+12	3.68145e+10	0	1.40227e+11	1.50575e+11	1.61082e+11
6	6.29946e+13	1.88503e+13	2.86085e+12	3.36166e+12	1.77041e+11	1.40227e+11	0	1.03479e+10	2.08549e+10
7	6.30049e+13	1.88607e+13	2.8712e+12	3.37201e+12	1.87389e+11	1.50575e+11	1.03479e+10	0	1.05071e+10
8	6.30154e+13	1.88712e+13	2.88171e+12	3.38252e+12	1.97896e+11	1.61082e+11	2.08549e+10	1.05071e+10	0

(Similarity Sheet)



(Similarity Heat Map)

According to the results of matrix and heat map, the threshold could be set as $1e+12$. Under the threshold, there are Earth, Venus, Mars, Mercury and Pluto. Earth, Venus, Mars and Mercury are all terrestrial planet, and Pluto is also like terrestrial because it also has solid surface. Above the threshold, there are Jupiter, Saturn, Uranus and Neptune. They are gas giants, which are composed mainly of hydrogen and helium

Problem 8 (14 points)

Given N documents. Write a Python program to find the most frequent

1. < word >

2. < word1, word2 >

3. < word1, word2, word3 >

e.g. D1 = {aa aa a aaa}, D2 = {aa aa aaa}, D3 = {aaa}, most frequent < word > is < aaa > whose frequency is 3, < word1, word2 > is < aa, aaa > whose frequency is 2, < word1, word2, word3 > is < a, aa, aaa > whose frequency is 1

```

9 import pandas as pd
10 import numpy as np
11 import os
12 import re
13
14 os.getcwd() #Get the current working path and see if it is your own
15 os.chdir('/Users/cyfile/Documents/Brandeis/Data Mining/HW0') #if it is not, it will
16 path = '/Users/cyfile/Documents/Brandeis/Data Mining/HW0/docs'
17 os.listdir(path) #See what data is in the target path
18
19 datalist = []
20 for i in os.listdir(path):
21     if os.path.splitext(i)[1] == '.txt': #把文件分为文件名和扩展名, 选
22         datalist.append(i)
23 datalist #查看datalist
24
25 alltext = []
26 for txt in datalist:
27     data_path = os.path.join(path,txt) #path data_path
28     x0 = open(data_path)
29     f = x0.read().lower()
30     alltext.append(f)
31
32 # data cleaning
33 lists_new = []
34 for j in alltext:
35     a = re.sub('[()~.,\!@":;_?\\t\\n1234567890&]', '', j)
36     lists_new.append(a)
37
38 list_formal = []
39 for item in lists_new:
40     list_formal.append(list(item.split()))
41
42 # count 1 word
43 list_oneword = []
44 for h in list_formal:
45     setword = set(h)
46     list_oneword.extend(setword)
47
48 dic = {}
49 for each in list_oneword:
50     if each in dic:
51         dic[each] = dic[each] + 1
52     else:
53         dic[each] = 1
54 # print all value
55 # print(sorted(dic.items(),key=lambda x:x[1],reverse=True))
56
57 # only print biggest value
58 HighValue = 0
59 HighKey = None
60 for each in dic:
61     if dic[each] > HighValue:
62         HighValue = dic[each]
63         HighKey = each
64
65 for each in dic:
66     if dic[each] == HighValue:
67         Highkey = each
68     print(Highkey, HighValue)

```

The results are shown below: the most frequent words are “and”, “is”, “the”, “in”, “a”, “university”, “of”. They all appears 93 times, which means, they all appear in every txt file.

```

and 93
is 93
the 93
in 93
a 93
university 93
of 93

```

Based on this, we could know the most frequent two words are combinations of two words from these 7 words, there are 21 combinations.

<and, is> <and, the> <and, in> <and, a> <and, university> <and, of>

<is, the> <is, in> <is, a> <is, university> <is, of>

<the, in> <the, a> <the, university> <the, of>

<in, a> <in, university> <in, of>

<a, university> <a, of>

<university, of>

And the most frequent three words are combinations of three words from these 7 words, there are 35 combinations. The logic to list all the combinations is the same as previous one.