COSI 126A Homework 0

Due by Sept. 7th

Yu Huai

September 7, 2018

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.
False.

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

(C) Computing the total sales of company.
False

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

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

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

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

(H) Monitoring seismic waves for earthquake activities.
True

(I) Extracting the frequencies of a sound wave.
True

Problem 2 (10 points)

Suppose that you are employed as a data mining consultant for an Internet search engine company. Describe how data mining can help the company by giving specific examples of how techniques, such as clustering, classification, association rule mining, and anomaly detection can be applied.

Data mining could help the company in following ways.

(A) Clustering. Web company could group together similar web pages using clustering methods. For example, when users search a key word, web page could show different groups of web pages with each group containing several web sources, such as recent news reports, social-network links, old news related to it.

(B) Classification. Web company could classify source pages into different groups, and retrieve information from each group to show on searching results. Also, company could use classification algorithms to filter harmful and illegal websites.

(C) Association rule mining. Based on users' searching behavior, web company could predict users' next probable behavior. Finding combinations of users' interests could help the company to recommend resources in advance.

(D) Anomaly detection. Search engine company needs to monitor anomalous IP addresses and unusual visiting frequency.

To sum up, data mining could guarantee website security and help building recommendation systems for useful links and advertisements by predicting user behaviors.

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.

Answer

Data Privacy is not an important issue for census. Census is executed by the government. The Census Bureau has several policies to ensure the data collected is well-protected. Usually, census data is confidential, and this applies to the individuals, households, and businesses. If the government expose this data to other institution for profiting or other purposes, then the data privacy would be a problem.

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

Answer

For web companies, collecting IP addresses and visit times of Web users is not an invasion of data privacy, because companies need to use these data to make improvements to their websites and make anomaly detection for security. However, IP addresses are relate to users' real residency addresses, so if web companies

(C) Images from Earth-orbiting satellites.

Answer

It's been controversial for a long time about whether Street View and satellites images invade personal privacy. The point is that people don't know when and where they are photographed. Not only the government but also for-profit companies could access satellite-tech. Since satellite imagery, facial recognition technology, and image distribution speed is rapidly improving, individuals are more possible to be identified. There should be some regulations to prevent privacy invasion.

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

Answer

Yes. Telephone book is personal. Addresses and names are closely related to people's safety for life and property. Especially if information from people's telephone books is exposed to public or advertisement companies or harmful groups, it would cause serious problems.

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

Answer

Yes. When it comes to people's contact emails or phone numbers, it's always a serious problem. If companies collect emails data not only for research purpose, but also selling to advertisement companies or other institutions, it will be invasion of privacy. People would receive unexpected spam emails. Moreover, people tend to use their emails as personal accounts, particularly as bank and e-commercial accounts. So when emails are collected unwell, it may cause loss of property.

Problem 4 (15 points)

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

Answer

- rank(A) = 1 ¡ 3, so the A is not invertible.A^{-1} does not exist.
- rank(A) = 1 ¡ 3, so the A is not invertible.A^{-1} does not exist
  Since A is not invertible, d = the number of rank-defect = 3-1 = 2.
  delete 2 rows and columns from $\begin{bmatrix}1&2&3\\2&4&6\\3&6&9\end{bmatrix}$ $A_1 = \begin{bmatrix}9\end{bmatrix}$ $A_1^{-1} = \begin{bmatrix}\frac{1}{9}\end{bmatrix}$
  padding 2 rows and columns with 0 to A_1^{-1},
  so $A^+ = \begin{bmatrix}0&0&0\\0&0&0\\0&0&\frac{1}{9}\end{bmatrix}$
  To verify the answer, AA^+A = A

  $$AA^+A = \begin{bmatrix}1&2&3\\2&4&6\\3&6&9\end{bmatrix}\begin{bmatrix}0&0&0\\0&0&0\\0&0&\frac{1}{9}\end{bmatrix}\begin{bmatrix}1&2&3\\2&4&6\\3&6&9\end{bmatrix}$$
  $$= \begin{bmatrix}0&0&\frac{1}{3}\\0&0&\frac{2}{3}\\0&0&1\end{bmatrix}\begin{bmatrix}1&2&3\\2&4&6\\3&6&9\end{bmatrix} = \begin{bmatrix}1&2&3\\2&4&6\\3&6&9\end{bmatrix} = A$$

- $A = \begin{bmatrix}1\\2\\3\end{bmatrix}\begin{bmatrix}1&2&3\end{bmatrix}$
  $A^2 = \begin{bmatrix}1\\2\\3\end{bmatrix}\begin{bmatrix}1&2&3\end{bmatrix}\begin{bmatrix}1\\2\\3\end{bmatrix}\begin{bmatrix}1&2&3\end{bmatrix} = 14A$
  $A^3 = 14A^2 = 14^2A$
  $So, A^n = 14^{n-1}A$
  $A^{100} = 14^{99}A = \begin{bmatrix}14^{99}&2\times14^{99}&3\times14^{99}\\2\times14^{99}&4\times14^{99}&6\times14^{99}\\3\times14^{99}&6\times14^{99}&9\times14^{99}\end{bmatrix}$

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 $P(Z$ gets $A^+)$

Answer

$P(Z$ gets $A^+) = \frac{1}{3}$
$P(Z \text{ gets } A^+) = \frac{P(X \text{ gets } A^+ \text{ and } Y \text{doesn't get } A^+)}{P(Teacher\ says\ Y\ doesn't\ get\ A^+)} = \frac{\frac{1}{6}}{\frac{1}{2}} = \frac{1}{3}$

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 $P(A = sub) = 2.5\%$, for B, it's $P(B = sub) = 5\%$. Warehouse tests 4 products and one of them is substandard. What is the probability that this product is from A, $P(\text{this sub from } A)$

Answer

$P(x|x = sub \quad \text{and} \quad x \in A) = 70\% \times 2.5\% = 1.75\%$
$P(x|x = sub \quad \text{and} \quad x \in B) = 30\% \times 5\% = 1.5\%$
$P(\text{this sub from } A) = \frac{1.75\%}{1.75\%+1.5\%} = 53.8\%$

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	**Distance to Sun (km)**	**Radius (km)**	**Mass (kg)**
p	d	r	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

Answer

Similarity Matrix									
Planet	Jupiter	Saturn	Uranus	Neptune	Earth	Venus	Mars	Mercury	Pluto
Jupiter		4.41e+13	6.01e+13	5.96e+13	6.28e+13	6.29e+13	6.30e+13	6.30e+13	6.30e+13
Saturn			1.60e+13	1.55e+13	1.87e+13	1.87e+13	1.89e+13	1.89e+13	1.89e+13
Uranus				5.00e+11	2.68e+12	2.72e+12	2.86e+12	2.87e+12	2.88e+12
Neptune					3.18e+12	3.22e+12	3.36e+12	3.37e+12	3.38e+12
Earth						3.68e+10	1.77e+11	1.87e+11	1.98e+11
Venus							1.40e+11	1.51e+11	1.61e+11
Mars								1.03e+10	2.09e+10
Mercury									1.05e+10
Pluto									
Average S	5.99e+13	2.12e+13	1.13e+13	1.15e+13	1.10e+13	1.10e+13	1.11e+13	1.11e+13	1.11e+13

According to the table,the threshold could be 1.2e+13. There are two groups: Jupiter and Saturn are in group1, others are in group2.
Threshold means the similarity between two planets, the threshold decreases, planets in a group are more similar.

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. $D_1 = \{$aa aa a aaa$\}, D_2 = \{$aa aa aaa$\}, D_3 = \{$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

Answer

This is a frequent itemset mining problem. I use Apriori algorithm. Assumption: If word1 and word2 have same frequency, pick either one. If word1, word2 and word1, word3 have same frequency, pick either one.
To execute it, use command python p8.py path-to-docs