COSI 126A Homework 0

Due by January 21st

January 12, 2020

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.

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

(C) Computing the total sales of company.

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

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

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

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

(H) Monitoring seismic waves for earthquake activities.

(I) Extracting the frequencies of a sound wave.

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.

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.

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

(C) Images from Earth-orbiting satellites.

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

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

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}

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 $P(Z \text{ gets } A^{+})$

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, $P(\text{this sub from } A)$

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

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