CS 412: Introduction to Data Mining
Quiz 2

Full Name: ______________________

NetID: ______________________

Instructions:

- There are 6 problems in this quiz, worth 35 points in total.
- You must not communicate with other students during this test.
- Books or notes are not allowed.
- Electronic devices except calculators are not allowed.
- You cannot use your mobile phone as calculator.
- The duration of this quiz is 45 minutes.
- Wait for instructions before turning this page.

1 Normalization

Consider the following data points: 697, 250, 775, 690, 280, 872. (i) If we use min-max normalization to transform the data onto the range [0.5, 1.5], what would be the normalized value for x? (ii) If we use z-score normalization to transform the data, what would be the normalized value for x? (iii) If we use z-score normalization using mean absolute deviation instead of standard deviation, what would be the normalized value for x? [5 points]

Answer:

(i) 0.5 + (x - 250) * (1.5 - 0.5) / (872 - 250) = 0.5 + (x - 250) / 622
(ii) (x - Avg(697, 250, 775, 690, 280, 872)) / SD(697, 250, 775, 690, 280, 872) = (x - 594) / 263.364
(ii) (x - Avg(697, 250, 775, 690, 280, 872)) / MAD(697, 250, 775, 690, 280, 872) = (x - 594) / 219.33

2 Datacube Update

Suppose we need to record a single measure in a datacube: median. Design an efficient computation and storage method for median given that the cube allows data to be deleted incrementally (i.e., in small portions at a time). **Hint:** Consider what additional information can be stored so that median can be recalculated in most cases after a deleting a small portion of data. [10 points]

Answer:

For median, keep a small number p of centered values, (e.g. p = 10), plus two counts: up count and down count. Each removal may change the count or remove a centered value. If the median no longer falls among these centered values, recalculate the set. Otherwise, the median can be easily calculated from the above set.

3 Distance Measure

The N-Puzzle is a board game for a single player. It consists of $(N^2 - 1)$ numbered squared tiles in random order, and one blank space (a missing tile). The object of the puzzle is to rearrange the tiles in order by making sliding moves that use the empty space, using the fewest moves. Moves of the puzzle are made by sliding an adjacent tile into the empty space. Only tiles that are horizontally or vertically adjacent to the blank space (not diagonally adjacent) may be moved. You are to design a distance measure for calculating the distance between two different states (configurations) of the game. The distance measure should closely reflect how many moves are required to move the game from one state to to the other. For the two states shown in figure, use your distance measure to calculate the distance between the two states. [5 points]

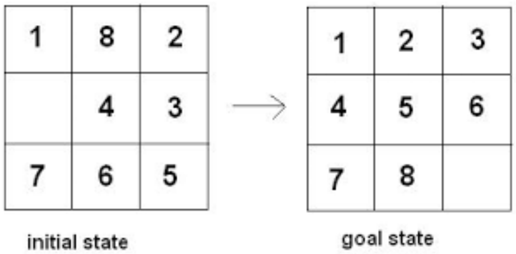

Figure 1: 3-Puzzle

Answer:

This is a classical example of designing a distance measure based on a heuristic. The commonly employed heuristic for this problem is Manhattan distance. The distance between any two states of the N-Puzzle game can be defined as the sum of Manhattan distance of the N^2 tiles, based on the position of these tiles in the two states. For example, consider the two states shown in the figure. For the tile numbered 1, the Manhattan distance for two states is 0. For the tile numbered numbered 2, the Manhattan distance for the two states is 1. If we add these distances for all tiles (1-8 and blank), we get the distance between the two states.

4 Querying Datacube

Starting with the base cuboid [day, doctor, patient], what specific OLAP operations should be performed in order to list the total fee collected by each doctor in 2010? In each dimension, there exists a hierarchy, which is presented in the figure as a top down order. [5 points]

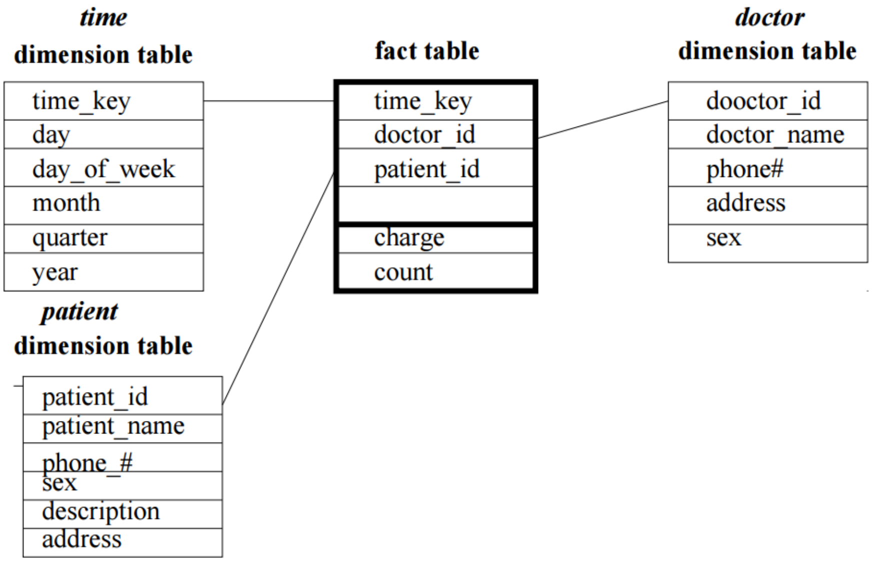

Figure 2: Star schema for data warehouse

Answer:

(i) Roll up on time from day to year. (ii) Slice for time = 2010. (ii) Roll up on patient from individual patient to all.

5 Betting Game

Consider the following betting game. A gambler first flips a fair coin to determine the amount of bet: if heads, the bet is \$1; if tails, the bet is \$2. Then the gambler flips again: if heads, he/she wins the amount of bet; if tails, he/she loses it. For example, if the outcome of coin toss is first heads and then tails, the gambler loses \$1; if the outcome is first tails and then heads the gambler wins \$2. Let X be the amount a gambler bets, and Y be his/her net winnings. Determine if variable X and Y are (i) independent or not, (ii) correlated or not. You need to present concise arguments in favor of your answer. [3+2 = 5 points]

Answer:

For simplicity, consider a single game scenario. There are four possible outcomes: HH, HT, TH, TT. Now, P(X=2) = 1/2, P(Y=-2) = 1/4, P(X=2,Y=-2) =1/4. Therefore, P(X=x,Y=y) != P(X=x)P(Y=y), which implies X and Y are not independent. In addition, increasing X does not guarantee the increase (or decrease) of Y, and vice versa. Therefore, the two variables are not correlated. This problem is an example of an important observation: if X and Y are uncorrelated, then they can still be dependent. Please check the following link for a detailed explanation.

https://www.stat.cmu.edu/ cshalizi/uADA/13/reminders/uncorrelated-vs-independent.pdf.

6 Eigenvectors in PCA

What do the eigenvectors of covariance matrix represent in PCA? Given the following square matrix A, decide if any of the following vectors are eigenvectors. Also, determine the eigenvalues corresponding to the eigenvectors. [5 points]

$$A = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 3 & 4 \\ 0 & 4 & 9 \end{bmatrix}$$

$$\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \begin{bmatrix} 3 \\ 0 \\ 0 \end{bmatrix} \begin{bmatrix} 0 \\ 2 \\ -1 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ 2 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \begin{bmatrix} 0 \\ 4 \\ 6 \end{bmatrix} \begin{bmatrix} 1 \\ 4 \\ 6 \end{bmatrix}$$

Answer:

A zero vector is not an eigenvector.

$$\begin{bmatrix} 2 & 0 & 0 \\ 0 & 3 & 4 \\ 0 & 4 & 9 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 2*1+0*0+0*0 \\ 0*1+3*0+4*0 \\ 0*1+4*0+9*0 \end{bmatrix} = \begin{bmatrix} 2 \\ 0 \\ 0 \end{bmatrix}$$

$[1\ 0\ 0]^T$ is eigenvector with eigenvalue 2.
$[3\ 0\ 0]^T$ is eigenvector with eigenvalue 2.

$$\begin{bmatrix} 2 & 0 & 0 \\ 0 & 3 & 4 \\ 0 & 4 & 9 \end{bmatrix} \begin{bmatrix} 0 \\ 2 \\ -1 \end{bmatrix} = \begin{bmatrix} 2*0+0*2+0*(-1) \\ 0*0+3*2+4*(-1) \\ 0*0+4*2+9*(-1) \end{bmatrix} = \begin{bmatrix} 0 \\ 2 \\ -1 \end{bmatrix}$$

$[0\ 2\ -1]^T$ is eigenvector with eigenvalue 1.

$$\begin{bmatrix} 2 & 0 & 0 \\ 0 & 3 & 4 \\ 0 & 4 & 9 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ 2 \end{bmatrix} = \begin{bmatrix} 2*0+0*1+0*2 \\ 0*0+3*1+4*2 \\ 0*0+4*1+9*2 \end{bmatrix} = \begin{bmatrix} 0 \\ 11 \\ 22 \end{bmatrix}$$

$[0\ 1\ 2]^T$ is eigenvector with eigenvalue 11.

$$\begin{bmatrix} 2 & 0 & 0 \\ 0 & 3 & 4 \\ 0 & 4 & 9 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 2*1+0*1+0*1 \\ 0*1+3*1+4*1 \\ 0*1+4*1+9*1 \end{bmatrix} = \begin{bmatrix} 2 \\ 7 \\ 13 \end{bmatrix}$$

$[1\ 1\ 1]^T$ is not an eigenvector.

$$\begin{bmatrix} 2 & 0 & 0 \\ 0 & 3 & 4 \\ 0 & 4 & 9 \end{bmatrix} \begin{bmatrix} 0 \\ 4 \\ 6 \end{bmatrix} = \begin{bmatrix} 2*0+0*4+0*6 \\ 0*0+3*4+4*6 \\ 0*0+4*4+9*6 \end{bmatrix} = \begin{bmatrix} 0 \\ 36 \\ 70 \end{bmatrix}$$

$[0\ 4\ 6]^T$ is not an eigenvector.

$$\begin{bmatrix} 2 & 0 & 0 \\ 0 & 3 & 4 \\ 0 & 4 & 9 \end{bmatrix} \begin{bmatrix} 1 \\ 4 \\ 6 \end{bmatrix} = \begin{bmatrix} 2*1+0*4+0*6 \\ 0*1+3*4+4*6 \\ 0*1+4*4+9*6 \end{bmatrix} = \begin{bmatrix} 2 \\ 36 \\ 70 \end{bmatrix}$$

$[1\ 4\ 6]^T$ is not an eigenvector.