CS 412: Homework #1

Due on Friday Sept. 30th

Kurt Rudolph
rudolph9

Problem 1

As a highly application-driven discipline, data mining has been widely applied in many areas. We briey presented two highly successful and popular application examples of data mining: business intelligence and Web search engines, in our textbooks. Do you think that data mining can also be applied to the following areas? If yes, please provide a brief yet concrete example, if not, please briefly state your reasons.

(1) Software Engineering

Solution

(2) Transportation

Solution

(3) Sociology

Solution

Problem 2

Suppose a student collected the price and weight of 20 products in a shop with the following result

price	$11.78	$85.12	$10.47	$298.00	$38.45	$102.14	$123.62	$203.29	$65.00	$225.50
weight	3.2	3.4	4.5	35.4	9.1	5.7	1.5	23.8	8.6	42.3

price	$9.25	$164.32	$102.45	$120.45	$73.15	$625.00	$125.00	$242.64	$441.76	$325.45
weight	5.9	12.3	6.5	11.8	12.2	32.9	11.6	48.0	52.9	78.2

(1) Calculate the mean, Q1, median, Q3, and standard deviation of price and weight

Solution

(2) Draw the boxplots for price and weight

Solution

(3) Draw scatter plot and Q-Q plot based on these two variables

Solution

(4) Normalize the two variables based on the min-max normalization ($min = 1, max = 10$)

Solution

(5) Normalize the two variables based on the z-score normalization

Solution

(6) Calculate the Pearson correlation coefficient. Are these two variables positively or negatively correlated?

Solution

(7) Take the price of the above 20 products, partition them into four bins by each of the following methods

- equal-width partitioning
- equal-depth (equal-frequency) partitioning

Solution

Problem 3

Design a data warehouse for a university's gradebook data to analyze the class performances. Suppose the data warehouse consisting of the following dimensions: department, semester, course, student, instructor, and gradebook; and a set of measures you would like to define.

(a) Draw a star-schema, based on your consideration of power and convenience of analysis of the warehouse

Solution

(b) Suppose we want to present the standard-deviation of final scores by course and year, and freely drilling up and down in multidimensional space, describe how this measure can be computed efficiently

Solution

(c) Is top 10% in a class a holistic or algebraic measure? Discuss how to develop an efficient (maybe approximate) methods to compute a query like: find those Engineering students whose final score is within top 10% in class in at least 80% of the CS courses that he or she has taken?

Solution

(d) Is it a good idea to merge this data warehouse and the current university's gradebook database system together into one big data management/analysis system? Why?

Solution

Problem 4

A location-based social networking website which provides check-in services hires you to help them build a data warehouse.

Users of this service can "check-in" at venues using mobile device applications by running the applications and selecting from a list of venues that the application locates nearby. Also, users can "add" each other as "friends". The website also has sufficient information about venues, including address, GPS location, and category of the venue (e.g., a Japanese restaurant), and users tend to provide their personal information to the website when they register.

(a) Design a data warehouse that may facilitate effective on-line analytical processing for this website (provide both schema and measures, also explain why).

Solution

(b) Check-in data collected from the website and mobile applications are noisy. Besides network and device errors, are there any other reasons which might cause noises in this data set? For the reason you come up with, discuss a method that can clean-up check-in data effectively in the data warehouse.

Solution

(c) One may like to performance on-line analytical processing to the checks-in data at different venues by month, by cities and by categories (Italian or Japanese, etc.). How can this be done efficiently in the data warehouse?

Solution

(d) Hackers create fake profiles on this website. They are using bots to manipulate fake profiles, generate fake check-in data and try to add everyone as their friends (yes this is a common problem for many social network websites, and no, I am not telling you how to write bots). Although bots are trying to mimic real users, they still behave differently, e.g., they check-in at random places (Chicago this minute, Las Vegas next minute), they check-in way too often than real users, and their social network structures are usually very large but also very sparse (your friends on facebook tend to form communities but bots don't do that). Discuss possible solutions on how to identify fake profiles (bots) in your data warehouse.

Solution