
Submission Requirements

- All answers must be computer generated (including text and diagrams).
- The hand-in version must be ordered correctly and stapled in the top left corner.
- The hand-in version must include a header page (or with sufficient space) indicating: student name, student number, net id, course number and assignment number.

Question 1:

In the introduction chapter, we have presented a certain view on the technology evolution, challenging problems in data mining and potential data mining applications. Now it is time to think over these problems again.

1. We have presented a view that data mining is the result of the evolution of database technology. Do you think that data mining is also the results of the evolution of machine learning research? Can you present such views based on the historical progress of this discipline? Do the same for the field of statistics.
2. Based on your view, what is the most challenging research problems in data mining?
3. What are the cool emerging applications in data mining that have not been presented in (all the chapter of) this textbook?

Question 2:

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

<table>
<thead>
<tr>
<th>price</th>
<th>$5.89</th>
<th>$149</th>
<th>$59.98</th>
<th>$129</th>
<th>$15.89</th>
<th>$56.99</th>
<th>$82.75</th>
<th>$42.19</th>
<th>$31</th>
<th>$125.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>weight</td>
<td>1.4</td>
<td>1.5</td>
<td>2.2</td>
<td>2.7</td>
<td>3.2</td>
<td>3.9</td>
<td>4.1</td>
<td>4.1</td>
<td>4.6</td>
<td>4.8</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>price</th>
<th>$4.5</th>
<th>$22</th>
<th>$52.9</th>
<th>$61</th>
<th>$33.5</th>
<th>$328</th>
<th>$122</th>
<th>$142.19</th>
<th>$229</th>
<th>$189.4</th>
</tr>
</thead>
<tbody>
<tr>
<td>weight</td>
<td>4.9</td>
<td>5.3</td>
<td>5.5</td>
<td>5.8</td>
<td>6.2</td>
<td>8.9</td>
<td>11.6</td>
<td>18.0</td>
<td>22.9</td>
<td>38.2</td>
</tr>
</tbody>
</table>

1. Calculate the mean, Q1, median, Q3, and standard deviation of price and weight;
2. Draw the boxplots for price and weight;
3. Draw scatter plot and Q-Q plot based on these two variables;
4. Normalize the two variables based on the min-max normalization (min = 1, max = 10);
5. Normalize the two variables based on the z-score normalization;
6. Calculate the Pearson correlation coefficient. Are these two variables positively or negatively correlated?
7. Take the price of the above 20 products, partition them into four bins by each of the following methods
   • equal-width partitioning, and
   • equal-depth (equal-frequency) partitioning

**Question 3:**

Design a data warehouse for a university consisting of the following dimensions: student, course, semester, and instructor; and a set of measures you would like to define.

1. Draw a star-schema,
2. Discuss what is the power (i.e., what can be analyzed efficiently) of your designed data warehouse, and
3. Suppose you start from the top (all-summary) of the multi-dimensional hierarchy, what are the **concrete** OLAP operations (drilling, slicing, etc.) you need to find the following:
   • grade point average of each student in the Department of Computer Science in Spring 2008
   • toughest grader among the instructors in the College of Engineering
4. For your designed warehouse, give two meaningful examples for each of the three types of measures: distributive, algebraic, and holistic.

**Question 4.**

A data warehouse for a retail chain may store that embraces the RFID technology. The company registers huge amounts of RFID data in the format of (RFID, location, time), and each RFID represents one object (product) associated with a set of product information, such as product-brand, product-category, place-made, time-made, etc.

People may like to analyze the store, exhibit, and sales of the products in the store in an OLAP manner.

1. Design such a data warehouse that should facilitate future efficient querying and on-line analytical processing in multidimensional space
2. Discuss how RFID data, even it may contain some noise, can be efficiently populated and integrated into such a data warehouse
3. Discuss how such a data warehouse may facilitate OLAP processing

**Question 5.**

Some measures in a data cube may not be easy to be computed efficiently.
1. For a data cube with the three dimensions: time, location, and item, with a measure of sales for each item at a location at a time. Suppose we want to present the standard-deviation of sales by item category, location and week, and freely drilling up and down in multidimensional space, describe how this measure can be computed efficiently.

2. Median and rank are two holistic measures. Discuss how to develop an efficient (maybe approximate) methods to compute these two measures in a multi-dimensional space.