Data Mining: Concepts and Techniques

3rd Edition

Solution Manual

Jiawei Han, Micheline Kamber, Jian Pei

The University of Illinois at Urbana-Champaign

Simon Fraser University

Version January 2, 2012

Preface

For a rapidly evolving field like data mining, it is difficult to compose "typical" exercises and even more difficult to work out "standard" answers. Some of the exercises in *Data Mining: Concepts and Techniques* are themselves good research topics that may lead to future Master or Ph.D. theses. Therefore, our solution manual is intended to be used as a guide in answering the exercises of the textbook. You are welcome to enrich this manual by suggesting additional interesting exercises and/or providing more thorough, or better alternative solutions.

While we have done our best to ensure the correctness of the solutions, it is possible that some typos or errors may exist. If you should notice any, please feel free to point them out by sending your suggestions to *hanj@cs.uiuc.edu*. We appreciate your suggestions.

To assist the teachers of this book to work out additional homework or exam questions, we have added one additional section "**Supplementary Exercises**" to each chapter of this manual. This section includes additional exercise questions and their suggested answers and thus may substantially enrich the value of this solution manual. Additional questions and answers will be incrementally added to this section, extracted from the assignments and exam questions of our own teaching. To this extent, our solution manual will be incrementally enriched and subsequently released in the future months and years.

Notes to the current release of the solution manual.

Due to the limited time, this release of the solution manual is a preliminary version. Many of the newly added exercises in the third edition have not provided the solutions yet. We apologize for the inconvenience. We will incrementally add answers to those questions in the next several months and release the new versions of updated solution manual in the subsequent months.

Acknowledgements

For each edition of this book, the solutions to the exercises were worked out by different groups of teach assistants and students. We sincerely express our thanks to all the teaching assistants and participating students who have worked with us to make and improve the solutions to the questions. In particular, for the first edition of the book, we would like to thanks Denis M. C. Chai, Meloney H.-Y. Chang, James W. Herdy, Jason W. Ma, Jiuhong Xu, Chunyan Yu, and Ying Zhou who took the class of *CMPT-459: Data Mining and Data Warehousing* at Simon Fraser University in the Fall semester of 2000 and contributed substantially to the solution manual of the first edition of this book. For those questions that also appear in the first edition, the answers in this current solution manual are largely based on those worked out in the preparation of the first edition.

For the solution manual of the second edition of the book, we would like to thank Ph.D. students and teaching assistants, Deng Cai and Hector Gonzalez, for the course *CS412: Introduction to Data Mining and Data Warehousing*, offered in the Fall semester of 2005 in the Department of Computer Science at the University of Illinois at Urbana-Champaign. They have helped prepare and compile the answers for the new exercises of the first seven chapters in our second edition. Moreover, our thanks go to several students from the *CS412* class in the Fall semester of 2005 and the *CS512: Data Mining: Principles and Algorithms* classes

in the Spring semester of 2006. Their answers to the class assignments have contributed to the advancement of this solution manual.

For the solution manual of the third edition of the book, we would like to thank Ph.D. students, Jialu Liu, Brandon Norick and Jingjing Wang, in the course *CS412: Introduction to Data Mining and Data Warehousing*, offered in the Fall semester of 2011 in the Department of Computer Science at the University of Illinois at Urbana-Champaign. They have helped checked the answers of the previous editions and did many modifications, and also prepared and compiled the answers for the new exercises in this edition. Moreover, our thanks go to teaching assistants, Xiao Yu, Lu An Tang, Xin Jin and Peixiang Zhao, from the *CS412* class and the *CS512: Data Mining: Principles and Algorithms* classes in the years of 2008–2011. Their answers to the class assignments have contributed to the advancement of this solution manual.

Contents

Chapter 1

Introduction

1.1 Exercises

1. What is *data mining*? In your answer, address the following:

   (a) Is it another hype?

   (b) Is it a simple transformation or application of technology developed from *databases*, *statistics*, *machine learning*, and *pattern recognition*?

   (c) 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 result 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 fields of *statistics* and *pattern recognition*.

   (d) Describe the steps involved in data mining when viewed as a process of knowledge discovery.

   Answer:

   Data mining refers to the process or method that extracts or "mines" interesting knowledge or patterns from large amounts of data.

   (a) Is it another hype?

   Data mining is not another hype. Instead, the need for data mining has arisen due to the wide availability of huge amounts of data and the imminent need for turning such data into useful information and knowledge. Thus, data mining can be viewed as the result of the natural evolution of information technology.

   (b) Is it a simple transformation of technology developed from databases, statistics, and machine learning?

   No. Data mining is more than a simple transformation of technology developed from databases, statistics, and machine learning. Instead, data mining involves an integration, rather than a simple transformation, of techniques from multiple disciplines such as database technology, statistics, machine learning, high-performance computing, pattern recognition, neural networks, data visualization, information retrieval, image and signal processing, and spatial data analysis.

   (c) Explain how the evolution of database technology led to data mining.

   Database technology began with the development of data collection and database creation mechanisms that led to the development of effective mechanisms for data management including data storage and retrieval, and query and transaction processing. The large number of database systems offering query and transaction processing eventually and naturally led to the need for data analysis and understanding. Hence, data mining began its development out of this necessity.

(d) Describe the steps involved in data mining when viewed as a process of knowledge discovery.
The steps involved in data mining when viewed as a process of knowledge discovery are as follows:

- **Data cleaning**, a process that removes or transforms noise and inconsistent data
- **Data integration**, where multiple data sources may be combined
- **Data selection**, where data relevant to the analysis task are retrieved from the database
- **Data transformation**, where data are transformed or consolidated into forms appropriate for mining
- **Data mining**, an essential process where intelligent and efficient methods are applied in order to extract patterns
- **Pattern evaluation**, a process that identifies the truly interesting patterns representing knowledge based on some interestingness measures
- **Knowledge presentation**, where visualization and knowledge representation techniques are used to present the mined knowledge to the user

■

2. How is a *data warehouse* different from a database? How are they similar?

Answer:

Differences between a data warehouse and a database: A **data warehouse** is a repository of information collected from multiple sources, over a history of time, stored under a unified schema, and used for data analysis and decision support; whereas a **database**, is a collection of interrelated data that represents the current status of the stored data. There could be multiple heterogeneous databases where the schema of one database may not agree with the schema of another. A database system supports ad-hoc query and on-line transaction processing. For more details, please refer to the section "Differences between operational database systems and data warehouses."

Similarities between a data warehouse and a database: Both are repositories of information, storing huge amounts of persistent data.

■

3. Define each of the following *data mining functionalities*: characterization, discrimination, association and correlation analysis, classification, regression, clustering, and outlier analysis. Give examples of each data mining functionality, using a real-life database that you are familiar with.

Answer:

Characterization is a summarization of the general characteristics or features of a target class of data. For example, the characteristics of students can be produced, generating a profile of all the University first year computing science students, which may include such information as a high GPA and large number of courses taken.

Discrimination is a comparison of the general features of target class data objects with the general features of objects from one or a set of contrasting classes. For example, the general features of students with high GPA's may be compared with the general features of students with low GPA's. The resulting description could be a general comparative profile of the students such as 75% of the students with high GPA's are fourth-year computing science students while 65% of the students with low GPA's are not.

Association is the discovery of association rules showing attribute-value conditions that occur frequently together in a given set of data. For example, a data mining system may find association rules like

$$major(X, \text{“}computing\ science\text{”}\text{”}) \Rightarrow owns(X, \text{“}personal\ computer\text{”})$$

$$[support = 12\%, confidence = 98\%]$$

where X is a variable representing a student. The rule indicates that of the students under study, 12% (**support**) major in computing science and own a personal computer. There is a 98% probability (**confidence**, or certainty) that a student in this group owns a personal computer. Typically, association rules are discarded as uninteresting if they do not satisfy both a **minimum support threshold** and a **minimum confidence threshold**. Additional analysis can be performed to uncover interesting statistical **correlations** between associated attribute-value pairs.

Classification is the process of finding a **model** (or function) that describes and distinguishes data classes or concepts, for the purpose of being able to use the model to predict the class of objects whose class label is unknown. It predicts categorical (discrete, unordered) labels.

Regression, unlike **classification**, is a process to model continuous-valued functions. It is used to predict missing or unavailable *numerical data values* rather than (discrete) class labels.

Clustering analyzes data objects without consulting a known class label. The objects are clustered or grouped based on the principle of maximizing the intraclass similarity and minimizing the interclass similarity. Each cluster that is formed can be viewed as a class of objects. Clustering can also facilitate *taxonomy formation*, that is, the organization of observations into a hierarchy of classes that group similar events together.

Outlier analysis is the analysis of **outliers**, which are objects that do not comply with the general behavior or model of the data. Examples include fraud detection based on a large dataset of credit card transactions. ■

4. Present an example where data mining is crucial to the success of a business. What *data mining functionalities* does this business need (e.g., think of the kinds of patterns that could be mined)? Can such patterns be generated alternatively by data query processing or simple statistical analysis?

   Answer:

   A department store, for example, can use data mining to assist with its target marketing mail campaign. Using data mining functions such as association, the store can use the mined strong association rules to determine which products bought by one group of customers are likely to lead to the buying of certain other products. With this information, the store can then mail marketing materials only to those kinds of customers who exhibit a high likelihood of purchasing additional products. Data query processing is used for data or information retrieval and does not have the means for finding association rules. Similarly, simple statistical analysis cannot handle large amounts of data such as those of customer records in a department store.

   ■

5. What is the difference between discrimination and classification? Between characterization and clustering? Between classification and regression? For each of these pairs of tasks, how are they similar?

   Answer:

   Discrimination differs from **classification** in that the former refers to a comparison of the general features of target class data objects with the general features of objects from one or a set of contrasting classes, while the latter is the process of finding a set of models (or functions) that describe and distinguish data classes or concepts for the purpose of being able to use the model to predict the class of objects whose class label is unknown. Discrimination and classification are similar in that they both deal with the analysis of class data objects.

   Characterization differs from **clustering** in that the former refers to a summarization of the general characteristics or features of a target class of data while the latter deals with the analysis of data objects without consulting a known class label. This pair of tasks is similar in that they both deal with grouping together objects or data that are related or have high similarity in comparison to one another.

Classification differs from **regression** in that the former predicts categorical (discrete, unordered) labels while the latter predicts missing or unavailable, and often numerical, data values. This pair of tasks is similar in that they both are tools for prediction. ■

6. Based on your observation, describe another possible kind of knowledge that needs to be discovered by data mining methods but has not been listed in this chapter. Does it require a mining methodology that is quite different from those outlined in this chapter?

   Answer:

   There is no standard answer for this question and one can judge the quality of an answer based on the freshness and quality of the proposal. For example, one may propose *partial periodicity* as a new kind of knowledge, where a pattern is partial periodic if only some offsets of a certain time period in a time series demonstrate some repeating behavior. ■

7. *Outliers* are often discarded as noise. However, one person's garbage could be another's treasure. For example, exceptions in credit card transactions can help us detect the fraudulent use of credit cards. Using fraudulence detection as an example, propose two methods that can be used to detect outliers and discuss which one is more reliable.

   Answer:

   There are many outlier detection methods. More details can be found in Chapter 12. Here we propose two methods for fraudulence detection:

   a) **Statistical methods** (also known as **model-based methods**): Assume that the normal transaction data follow some statistical (stochastic) model, then data not following the model are outliers.

   b) **Clustering-based methods**: Assume that the normal data objects belong to large and dense clusters, whereas outliers belong to small or sparse clusters, or do not belong to any clusters.

   It is hard to say which one is more reliable. The effectiveness of statistical methods highly depends on whether the assumptions made for the statistical model hold true for the given data. And the effectiveness of clustering methods highly depends on which clustering method we choose. ■

8. Describe three challenges to data mining regarding *data mining methodology* and *user interaction issues.*

   Answer:

   Challenges to data mining regarding data mining methodology and user interaction issues include the following: mining different kinds of knowledge in databases, interactive mining of knowledge at multiple levels of abstraction, incorporation of background knowledge, data mining query languages and ad hoc data mining, presentation and visualization of data mining results, handling noisy or incomplete data, and pattern evaluation. Below are the descriptions of the first three challenges mentioned:

   Mining different kinds of knowledge in databases: Different users are interested in different kinds of knowledge and will require a wide range of data analysis and knowledge discovery tasks such as data characterization, discrimination, association, classification, clustering, trend and deviation analysis, and similarity analysis. Each of these tasks will use the same database in different ways and will require different data mining techniques.

   Interactive mining of knowledge at multiple levels of abstraction: Interactive mining, with the use of OLAP operations on a data cube, allows users to focus the search for patterns, providing and refining data mining requests based on returned results. The user can then interactively view the data and discover patterns at multiple granularities and from different angles.

   Incorporation of background knowledge: Background knowledge, or information regarding the domain under study such as integrity constraints and deduction rules, may be used to guide the

discovery process and allow discovered patterns to be expressed in concise terms and at different levels of abstraction. This helps to focus and speed up a data mining process or judge the interestingness of discovered patterns. ■

9. What are the major challenges of mining a huge amount of data (such as billions of tuples) in comparison with mining a small amount of data (such as a few hundred tuple data set)?

   Answer:

   One challenge to data mining regarding performance issues is the *efficiency and scalability* of data mining algorithms. Data mining algorithms must be efficient and scalable in order to effectively extract information from large amounts of data in databases within predictable and acceptable running times. Another challenge is the *parallel, distributed, and incremental* processing of data mining algorithms. The need for parallel and distributed data mining algorithms has been brought about by the huge size of many databases, the wide distribution of data, and the computational complexity of some data mining methods. Due to the high cost of some data mining processes, incremental data mining algorithms incorporate database updates without the need to mine the entire data again from scratch. ■

10. Outline the major research challenges of data mining in one specific application domain, such as stream/sensor data analysis, spatiotemporal data analysis, or bioinformatics.

    Answer:

    Let's take spatiotemporal data analysis for example. With the ever increasing amount of available data from sensor networks, web-based map services, location sensing devices etc., the rate at which such kind of data are being generated far exceeds our ability to extract useful knowledge from them to facilitate decision making and to better understand the changing environment. It is a great challenge how to utilize existing data mining techniques and create novel techniques as well to effectively exploit the rich spatiotemporal relationships/patterns embedded in the datasets because both the temporal and spatial dimensions could add substantial complexity to data mining tasks. First, the spatial and temporal relationships are information bearing and therefore need to be considered in data mining. Some spatial and temporal relationships are implicitly defined, and must be extracted from the data. Such extraction introduces some degree of fuzziness and/or uncertainty that may have an impact on the results of the data mining process. Second, working at the level of stored data is often undesirable, and thus complex transformations are required to describe the units of analysis at higher conceptual levels. Third, interesting patterns are more likely to be discovered at the lowest resolution/granularity level, but large support is more likely to exist at higher levels. Finally, how to express domain independent knowledge and how to integrate spatiotemporal reasoning mechanisms in data mining systems are still open problems [1].

    [1] J. Han and J. Gao, Research Challenges for Data Mining in Science and Engineering, in H. Kargupta, et al., (eds.), Next Generation of Data Mining, Chapman & Hall, 2009.

    ■

1.2 Supplementary Exercises

1. Suppose your task as a software engineer at *Big-University* is to design a data mining system to examine their university course database, which contains the following information: the name, address, and status (e.g., undergraduate or graduate) of each student, the courses taken, and their cumulative grade point average (GPA).

   Describe the *architecture* you would choose. What is the purpose of each component of this architecture?

   Answer:

A data mining architecture that can be used for this application would consist of the following major components:

- A **database, data warehouse, or other information repository**, which consists of the set of databases, data warehouses, spreadsheets, or other kinds of information repositories containing the student and course information.
- A **database or data warehouse server** which fetches the relevant data based on users' data mining requests.
- A **knowledge base** that contains the domain knowledge used to guide the search or to evaluate the interestingness of resulting patterns. For example, the knowledge base may contain metadata which describes data from multiple heterogeneous sources.
- A **data mining engine**, which consists of a set of functional modules for tasks such as classification, association, classification, cluster analysis, and evolution and deviation analysis.
- A **pattern evaluation module** that works in tandem with the data mining modules by employing interestingness measures to help focus the search towards interestingness patterns.
- **A graphical user interface** that allows the user an interactive approach to the data mining system.

■

2. Briefly describe the following *advanced database systems* and applications: object-relational databases, spatial databases, text databases, multimedia databases, the World Wide Web.

   Answer:

   An objected-oriented database is designed based on the object-oriented programming paradigm where data are a large number of objects organized into classes and class hierarchies. Each entity in the database is considered as an object. The object contains a set of variables that describe the object, a set of messages that the object can use to communicate with other objects or with the rest of the database system, and a set of methods where each method holds the code to implement a message.

   A spatial database contains spatial-related data, which may be represented in the form of raster or vector data. Raster data consists of n-dimensional bit maps or pixel maps, and vector data are represented by lines, points, polygons or other kinds of processed primitives, Some examples of spatial databases include geographical (map) databases, VLSI chip designs, and medical and satellite images databases.

   A text database is a database that contains text documents or other word descriptions in the form of long sentences or paragraphs, such as product specifications, error or bug reports, warning messages, summary reports, notes, or other documents.

   A multimedia database stores images, audio, and video data, and is used in applications such as picture content-based retrieval, voice-mail systems, video-on-demand systems, the World Wide Web, and speech-based user interfaces.

   The **World-Wide Web** provides rich, world-wide, on-line information services, where data objects are linked together to facilitate interactive access. Some examples of distributed information services associated with the World-Wide Web include America Online, Yahoo!, AltaVista, and Prodigy. ■

3. List and describe the five *primitives* for specifying a data mining task.

   Answer:

   The five primitives for specifying a data-mining task are:

- **Task-relevant data:** This primitive specifies the data upon which mining is to be performed. It involves specifying the database and tables or data warehouse containing the relevant data, conditions for selecting the relevant data, the relevant attributes or dimensions for exploration, and instructions regarding the ordering or grouping of the data retrieved.
- **Knowledge type to be mined:** This primitive specifies the specific data mining function to be performed, such as characterization, discrimination, association, classification, clustering, or evolution analysis. As well, the user can be more specific and provide pattern templates that all discovered patterns must match. These templates, or metapatterns (also called metarules or metaqueries), can be used to guide the discovery process.
- **Background knowledge:** This primitive allows users to specify knowledge they have about the domain to be mined. Such knowledge can be used to guide the knowledge discovery process and evaluate the patterns that are found. Of the several kinds of background knowledge, this chapter focuses on concept hierarchies.
- **Pattern interestingness measure:** This primitive allows users to specify functions that are used to separate uninteresting patterns from knowledge and may be used to guide the mining process, as well as to evaluate the discovered patterns. This allows the user to confine the number of uninteresting patterns returned by the process, as a data mining process may generate a large number of patterns. Interestingness measures can be specified for such pattern characteristics as simplicity, certainty, utility and novelty.
- **Visualization of discovered patterns:** This primitive refers to the form in which discovered patterns are to be displayed. In order for data mining to be effective in conveying knowledge to users, data mining systems should be able to display the discovered patterns in multiple forms such as rules, tables, cross tabs (cross-tabulations), pie or bar charts, decision trees, cubes or other visual representations.

■

4. Describe why *concept hierarchies* are useful in data mining.

   Answer:

   Concept hierarchies define a sequence of mappings from a set of lower-level concepts to higher-level, more general concepts and can be represented as a set of nodes organized in a tree, in the form of a lattice, or as a partial order. They are useful in data mining because they allow the discovery of knowledge at multiple levels of abstraction and provide the structure on which data can be generalized (rolled-up) or specialized (drilled-down). Together, these operations allow users to view the data from different perspectives, gaining further insight into relationships hidden in the data. Generalizing has the advantage of compressing the data set, and mining on a compressed data set will require fewer I/O operations. This will be more efficient than mining on a large, uncompressed data set. ■

5. Recent applications pay special attention to spatiotemporal data streams. A *spatiotemporal data stream* contains spatial information that changes over time, and is in the form of stream data, i.e., the data flow in-and-out like possibly infinite streams.

   (a) Present three application examples of spatiotemporal data streams.

   (b) Discuss what kind of interesting knowledge can be mined from such data streams, with limited time and resources.

   (c) Identify and discuss the major challenges in spatiotemporal data mining.

   (d) Using one application example, sketch a method to mine one kind of knowledge from such stream data efficiently.

   Answer: New question. Answer needs to be worked out. ■

6. Describe the differences between the following approaches for the integration of a data mining system with a database or data warehouse system: *no coupling, loose coupling, semitight coupling,* and *tight coupling.* State which approach you think is the most popular, and why.

   Answer: The differences between the following architectures for the integration of a data mining system with a database or data warehouse system are as follows.

   - **No coupling:** The data mining system uses sources such as flat files to obtain the initial data set to be mined since no database system or data warehouse system functions are implemented as part of the process. Thus, this architecture represents a poor design choice.
   - **Loose coupling:** The data mining system is not integrated with the database or data warehouse system beyond their use as the source of the initial data set to be mined, and possible use in storage of the results. Thus, this architecture can take advantage of the flexibility, efficiency and features such as indexing that the database and data warehousing systems may provide. However, it is difficult for loose coupling to achieve high scalability and good performance with large data sets as many such systems are memory-based.
   - **Semitight coupling:** Some of the data mining primitives such as aggregation, sorting or precomputation of statistical functions are efficiently implemented in the database or data warehouse system, for use by the data mining system during mining-query processing. Also, some frequently used intermediate mining results can be precomputed and stored in the database or data warehouse system, thereby enhancing the performance of the data mining system.
   - **Tight coupling:** The database or data warehouse system is fully integrated as part of the data mining system and thereby provides optimized data mining query processing. Thus, the data mining subsystem is treated as one functional component of an information system. This is a highly desirable architecture as it facilitates efficient implementations of data mining functions, high system performance, and an integrated information processing environment.

   From the descriptions of the architectures provided above, it can be seen that tight coupling is the best alternative without respect to technical or implementation issues. However, as much of the technical infrastructure needed in a tightly coupled system is still evolving, implementation of such a system is non-trivial. Therefore, the most popular architecture is currently semitight coupling as it provides a compromise between loose and tight coupling. ■

Chapter 2

Getting to Know Your Data

2.1 Exercises

1. Give three additional commonly used statistical measures (i.e., not illustrated in this chapter) for the characterization of *data dispersion,* and discuss how they can be computed efficiently in large databases.

   Answer:

   Data dispersion, also known as variance analysis, is the degree to which numeric data tend to spread and can be characterized by such statistical measures as *mean deviation, measures of skewness* and the *coefficient of variation.*

   The **mean deviation** is defined as the arithmetic mean of the absolute deviations from the means and is calculated as:

   $$mean\ deviation = \frac{\sum_{i=1}^{n} |x - \bar{x}|}{n} \tag{2.1}$$

   where, $\bar{x}$ is the arithmetic mean of the values and n is the total number of values. This value will be greater for distributions with a larger spread.

   A common **measure of skewness** is:

   $$\frac{\bar{x} - mode}{s} \tag{2.2}$$

   which indicates how far (in standard deviations, s) the mean ($\bar{x}$) is from the mode and whether it is greater or less than the mode.

   The **coefficient of variation** is the standard deviation expressed as a percentage of the arithmetic mean and is calculated as:

   $$coefficient\ of\ variation = \frac{s}{\bar{x}} \times 100 \tag{2.3}$$

   The variability in groups of observations with widely differing means can be compared using this measure.

   Note that all of the input values used to calculate these three statistical measures are algebraic measures (Chapter 4). Thus, the value for the entire database can be efficiently calculated by partitioning the database, computing the values for each of the separate partitions, and then merging theses values into an algebraic equation that can be used to calculate the value for the entire database.

The measures of dispersion described here were obtained from: Statistical Methods in Research and Production, fourth ed., Edited by Owen L. Davies and Peter L. Goldsmith, Hafner Publishing Company, NY:NY, 1972. ■

2. Suppose that the data for analysis includes the attribute *age*. The *age* values for the data tuples are (in increasing order) 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70.

   (a) What is the *mean* of the data? What is the *median*?

   (b) What is the *mode* of the data? Comment on the data's modality (i.e., bimodal, trimodal, etc.).

   (c) What is the *midrange* of the data?

   (d) Can you find (roughly) the first quartile ($Q1$) and the third quartile ($Q3$) of the data?

   (e) Give the *five-number summary* of the data.

   (f) Show a *boxplot* of the data.

   (g) How is a *quantile-quantile plot* different from a *quantile plot*?

   Answer:

   (a) What is the *mean* of the data? What is the *median*?
   The (arithmetic) mean of the data is: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = 809/27 = 30$. The median (middle value of the ordered set, as the number of values in the set is odd) of the data is: 25.

   (b) What is the *mode* of the data? Comment on the data's modality (i.e., bimodal, trimodal, etc.).
   This data set has two values that occur with the same highest frequency and is, therefore, bimodal. The modes (values occurring with the greatest frequency) of the data are 25 and 35.

   (c) What is the *midrange* of the data?
   The midrange (average of the largest and smallest values in the data set) of the data is: $(70 + 13)/2 = 41.5$

   (d) Can you find (roughly) the first quartile ($Q1$) and the third quartile ($Q3$) of the data?
   The first quartile (corresponding to the 25th percentile) of the data is: 20. The third quartile (corresponding to the 75th percentile) of the data is: 35.

   (e) Give the *five-number summary* of the data.
   The five number summary of a distribution consists of the minimum value, first quartile, median value, third quartile, and maximum value. It provides a good summary of the shape of the distribution and for this data is: 13, 20, 25, 35, 70.

   (f) Show a *boxplot* of the data.
   See Figure 2.1.

   (g) How is a *quantile-quantile plot* different from a *quantile plot*?
   A quantile plot is a graphical method used to show the approximate percentage of values below or equal to the independent variable in a univariate distribution. Thus, it displays quantile information for all the data, where the values measured for the independent variable are plotted against their corresponding quantile.
   A quantile-quantile plot however, graphs the quantiles of one univariate distribution against the corresponding quantiles of another univariate distribution. Both axes display the range of values measured for their corresponding distribution, and points are plotted that correspond to the quantile values of the two distributions. A line ($y = x$) can be added to the graph along with points representing where the first, second and third quantiles lie, in order to increase the graph's informational value. Points that lie above such a line indicate a correspondingly higher value for the distribution plotted on the y-axis, than for the distribution plotted on the x-axis at the same quantile. The opposite effect is true for points lying below this line.

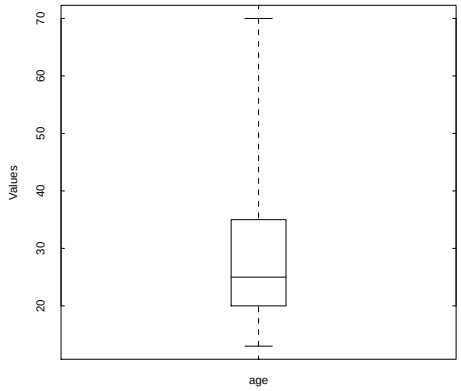

Figure 2.1: A boxplot of the data in Exercise 2.2.

■

3. Suppose that the values for a given set of data are grouped into intervals. The intervals and corresponding frequencies are as follows.

age	*frequency*
1–5	200
6–15	450
16–20	300
21–50	1500
51–80	700
81–110	44

Compute an *approximate median* value for the data.

Answer:

$L_1 = 20$, n= 3194, $(\sum_f)_l = 950$, $freq\_median = 1500$, width = 30, median = 30.94 years. ■

4. Suppose a hospital tested the age and body fat data for 18 randomly selected adults with the following result

age	23	23	27	27	39	41	47	49	50
%fat	9.5	26.5	7.8	17.8	31.4	25.9	27.4	27.2	31.2
age	52	54	54	56	57	58	58	60	61
%fat	34.6	42.5	28.8	33.4	30.2	34.1	32.9	41.2	35.7

(a) Calculate the mean, median and standard deviation of *age* and *%fat*.

(b) Draw the boxplots for *age* and *%fat*.

(c) Draw a *scatter plot* and a *q-q plot* based on these two variables.

Answer:

(a) Calculate the mean, median and standard deviation of *age* and *%fat*.
For the variable *age* the mean is 46.44, the median is 51, and the standard deviation is 12.85. For the variable *%fat* the mean is 28.78, the median is 30.7, and the standard deviation is 8.99.

(b) Draw the boxplots for *age* and *%fat*.
See Figure 2.2.

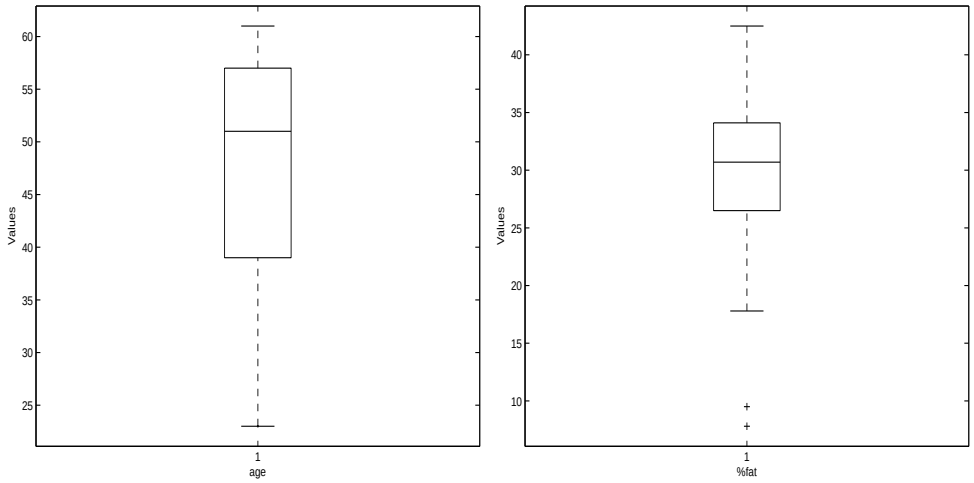

Figure 2.2: A boxplot of the variables *age* and *%fat* in Exercise 2.4.

(c) Draw a *scatter plot* and a *q-q plot* based on these two variables.
See Figure 2.3.

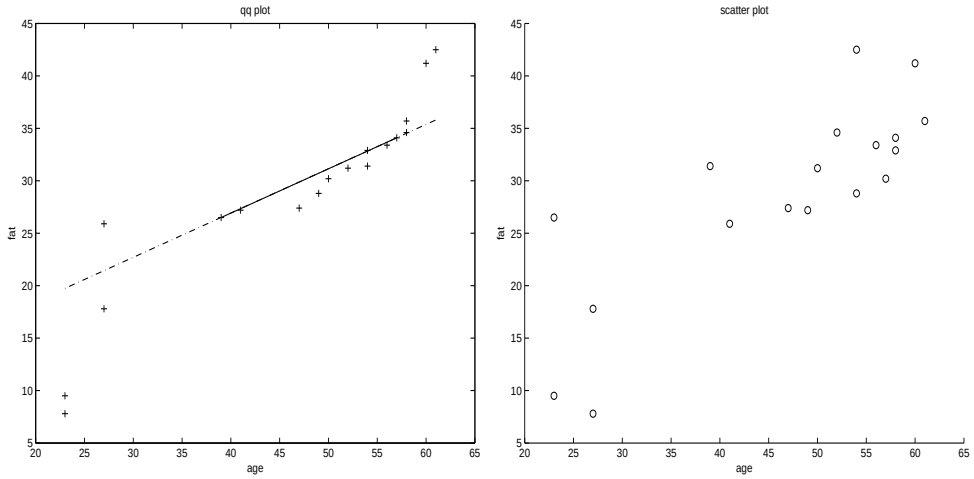

Figure 2.3: A *q-q plot* and a *scatter plot* of the variables *age* and *%fat* in Exercise 2.4.

■

5. Briefly outline how to compute the dissimilarity between objects described by the following:

   (a) Nominal attributes
   (b) Asymmetric binary attributes
   (c) Numeric attributes
   (d) Term-frequency vectors

Answer:

(a) Nominal attributes

A categorical variable is a generalization of the binary variable in that it can take on more than two states.

The dissimilarity between two objects i and j can be computed based on the ratio of mismatches:

$$d(i,j) = \frac{p-m}{p}, \tag{2.4}$$

where m is the number of *matches* (i.e., the number of variables for which i and j are in the same state), and p is the total number of variables.

Alternatively, we can use a large number of binary variables by creating a new binary variable for each of the M nominal states. For an object with a given state value, the binary variable representing that state is set to 1, while the remaining binary variables are set to 0.

(b) Asymmetric binary attributes

If all binary variables have the same weight, we have the contingency Table 2.1.

		object j		
		1	0	sum
object i	1	q	r	$q+r$
	0	s	t	$s+t$
	sum	$q+s$	$r+t$	p

Table 2.1: A contingency table for binary variables.

In computing the dissimilarity between asymmetric binary variables, the number of negative matches, t, is considered unimportant and thus is ignored in the computation, that is,

$$d(i,j) = \frac{r+s}{q+r+s}. \tag{2.5}$$

(c) Numeric attributes

Use **Euclidean distance**, **Manhattan distance**, or **supremum distance**. Euclidean distance is defined as

$$d(i,j) = \sqrt{(x_{i1}-x_{j1})^2 + (x_{i2}-x_{j2})^2 + \cdots + (x_{in}-x_{jn})^2}. \tag{2.6}$$

where $i = (x_{i1}, x_{i2}, \ldots, x_{in})$, and $j = (x_{j1}, x_{j2}, \ldots, x_{jn})$, are two n-dimensional data objects.

The **Manhattan (or city block) distance**, is defined as

$$d(i,j) = |x_{i1}-x_{j1}| + |x_{i2}-x_{j2}| + \cdots + |x_{in}-x_{jn}|. \tag{2.7}$$

The **supremum distance** is

$$d(i,j) = \lim_{h\to\infty} \left(\sum_{f=1}^{p} |x_{if}-x_{jf}|^h \right)^{\frac{1}{h}} = \max_{f}^{p} |x_{if}-x_{jf}|. \tag{2.8}$$

(d) Term-frequency vectors

To measure the distance between complex objects represented by vectors, it is often easier to abandon traditional metric distance computation and introduce a nonmetric similarity function.

For example, the similarity between two vectors, $\boldsymbol{x}$ and $\boldsymbol{y}$, can be defined as a cosine measure, as follows:

$$s(\boldsymbol{x}, \boldsymbol{y}) = \frac{\boldsymbol{x}^t \cdot \boldsymbol{y}}{||\boldsymbol{x}||||\boldsymbol{y}||} \tag{2.9}$$

where $\boldsymbol{x^t}$ is a transposition of vector $\boldsymbol{x}$, $||\boldsymbol{x}||$ is the Euclidean norm of vector $\boldsymbol{x}$,[1] $||\boldsymbol{y}||$ is the Euclidean norm of vector $\boldsymbol{y}$, and s is essentially the cosine of the angle between vectors $\boldsymbol{x}$ and $\boldsymbol{y}$.

■

6. Given two objects represented by the tuples (22, 1, 42, 10) and (20, 0, 36, 8):

   (a) Compute the *Euclidean distance* between the two objects.

   (b) Compute the *Manhattan distance* between the two objects.

   (c) Compute the *Minkowski distance* between the two objects, using $h = 3$.

   (d) Compute the *supremum distance* between the two objects.

   Answer:

   (a) Compute the *Euclidean distance* between the two objects.
   The Euclidean distance is computed using Equation (2.6).
   Therefore, we have $\sqrt{(22-20)^2 + (1-0)^2 + (42-36)^2 + (10-8)^2} = \sqrt{45} = 6.7082$.

   (b) Compute the *Manhattan distance* between the two objects.
   The Manhattan distance is computed using Equation (2.7). Therefore, we have $|22-20| + |1-0| + |42-36| + |10-8| = 11$.

   (c) Compute the *Minkowski distance* between the two objects, using $h = 3$.
   The Minkowski disance is

   $$d(i,j) = \sqrt[h]{|x_{i1} - x_{j1}|^h + |x_{i2} - x_{j2}|^h + \cdot + |x_{ip} - x_{jp}|^h} \tag{2.10}$$

   where h is a real number such that $h \geq 1$.
   Therefore, with $h = 3$, we have $\sqrt[3]{|22-20|^3 + |1-0|^3 + |42-36|^3 + |10-8|^3} = \sqrt[3]{233} = 6.1534$.

   (d) Compute the *supremum distance* between the two objects.
   The supremum distance is computed using Equation (2.8). Therefore, we have a supremum distance of 6.

■

7. The *median* is one of the most important holistic measures in data analysis. Propose several methods for median approximation. Analyze their respective complexity under different parameter settings and decide to what extent the real value can be approximated. Moreover, suggest a heuristic strategy to balance between accuracy and complexity and then apply it to all methods you have given.

   Answer:

   This question can be dealt with either theoretically or empirically, but doing some experiments to get the result is perhaps more interesting.

   We can give students some data sets sampled from different distributions, e.g., uniform, Gaussian (both two are symmetric) and exponential, gamma (both two are skewed). For example, if we use Equation

[1] The Euclidean normal of vector $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ is defined as $\sqrt{x_1^2 + x_2^2 + \ldots + x_n^2}$. Conceptually, it is the length of the vector.

(2.11) to do approximation as proposed in the chapter, the most straightforward way is to divide all data into k equal length intervals.

$$median = L_1 + (\frac{N/2 - (\sum freq)_l}{freq_{median}})width, \tag{2.11}$$

where L_1 is the lower boundary of the median interval, N is the number of values in the entire data set, $(\sum freq)_l$ is the sum of the frequencies of all of the intervals that are lower than the median interval, $freq_{median}$ is the frequency of the median interval, and $width$ is the width of the median interval.

Obviously, the error incurred will be decreased as k becomes larger and larger; however, the time used in the whole procedure will also increase. Let's analyze this kind of relationship more formally. It seems the product of error made and time used is a good optimality measure. From this point, we can do many tests for each type of distributions (so that the result won't be dominated by randomness) and find the k giving the best trade-off. In practice, this parameter value can be chosen to improve system performance.

There are also some other approaches to approximate the median, students can propose them, analyze the best trade-off point, and compare the results among different approaches. A possible way is as following: Hierarchically divide the whole data set into intervals: at first, divide it into k regions, find the region in which the median resides; then secondly, divide this particular region into k sub-regions, find the sub-region in which the median resides; .... We will iteratively doing this, until the width of the sub-region reaches a predefined threshold, and then the median approximation formula as above stated is applied. By doing this, we can confine the median to a smaller area without globally partitioning all data into shorter intervals, which is expensive (the cost is proportional to the number of intervals).

■

8. It is important to define or select similarity measures in data analysis. However, there is no commonly-accepted subjective similarity measure. Results can vary depending on the similarity measures used. Nonetheless, seemingly different similarity measures may be equivalent after some transformation.

   Suppose we have the following two-dimensional data set:

	A_1	A_2
$\boldsymbol{x_1}$	1.5	1.7
$\boldsymbol{x_2}$	2	1.9
$\boldsymbol{x_3}$	1.6	1.8
$\boldsymbol{x_4}$	1.2	1.5
$\boldsymbol{x_5}$	1.5	1.0

   (a) Consider the data as two-dimensional data points. Given a new data point, $\boldsymbol{x} = (1.4, 1.6)$ as a query, rank the database points based on similarity with the query using Euclidean distance, Manhattan distance, supremum distance, and cosine similarity.

   (b) Normalize the data set to make the norm of each data point equal to 1. Use Euclidean distance on the transformed data to rank the data points.

   Answer:

   (a) Use Equation (2.6) to compute the Euclidean distance, Equation (2.7) to compute the Manhattan distance, Equation (2.8) to compute the supremum distance, and Equation (2.9) to compute the cosine similarity between the input data point and each of the data points in the data set. Doing so yields the following table

	Euclidean dist.	Manhattan dist.	supremum dist.	cosine sim.
$\boldsymbol{x_1}$	0.1414	0.2	0.1	0.99999
$\boldsymbol{x_2}$	0.6708	0.9	0.6	0.99575
$\boldsymbol{x_3}$	0.2828	0.4	0.2	0.99997
$\boldsymbol{x_4}$	0.2236	0.3	0.2	0.99903
$\boldsymbol{x_5}$	0.6083	0.7	0.6	0.96536

These values produce the following rankings of the data points based on similarity:

Euclidean distance: x_1, x_4, x_3, x_5, x_2

Manhattan distance: x_1, x_4, x_3, x_5, x_2

Supremum distance: x_1, x_4, x_3, x_5, x_2

Cosine similarity: x_1, x_3, x_4, x_2, x_5

(b) The normalized query is (0.65850, 0.75258). The normalized data set is given by the following table

	A_1	A_2
$\boldsymbol{x_1}$	0.66162	0.74984
$\boldsymbol{x_2}$	0.72500	0.68875
$\boldsymbol{x_3}$	0.66436	0.74741
$\boldsymbol{x_4}$	0.62470	0.78087
$\boldsymbol{x_5}$	0.83205	0.55470

Recomputing the Euclidean distances as before yields

	Euclidean dist.
$\boldsymbol{x_1}$	0.00415
$\boldsymbol{x_2}$	0.09217
$\boldsymbol{x_3}$	0.00781
$\boldsymbol{x_4}$	0.04409
$\boldsymbol{x_5}$	0.26320

which results in the final ranking of the transformed data points: x_1, x_3, x_4, x_2, x_5

■

2.2 Supplementary Exercises

1. Briefly outline how to compute the dissimilarity between objects described by *ratio-scaled variables*.

   Answer:

   Three methods include:

   - Treat ratio-scaled variables as interval-scaled variables, so that the Minkowski, Manhattan, or Euclidean distance can be used to compute the dissimilarity.
   - Apply a logarithmic transformation to a ratio-scaled variable f having value x_{if} for object i by using the formula $y_{if} = \log(x_{if})$. The y_{if} values can be treated as interval-valued,
   - Treat x_{if} as continuous ordinal data, and treat their ranks as interval-scaled variables.

   ■