of algorithms. Not only does IBM have embedded analytics in DB2, but following its acquisition of SPSS, IBM has incorporated Clementine and SPSS into IBM Modeler.

There are still a large number of stand-alone data mining tools based on a single algorithm or on a collection of algorithms called a suite. Target users include both statisticians and business intelligence analysts. The leading suites include SAS Enterprise Miner, SAS JMP, IBM Modeler, Salford Systems SPM, Statistica, XLMiner, and RapidMiner. Suites are characterized by providing a wide range of functionality, frequently accessed via a graphical user interface designed to enhance model-building productivity. A popular approach for many of these GUIs is to provide a workflow interface in which the data mining steps and analysis are linked together.

Many suites have outstanding visualization tools and links to statistical packages that extend the range of tasks they can perform. They provide interactive data transformation tools as well as a procedural scripting language for more complex data transformations. The suite vendors are working to link their tools more closely to underlying DBMSs; for example, data transformations might be handled by the DBMS. Data mining models can be exported to be incorporated into the DBMS through generating SQL, procedural language code (e.g., C++ or Java), or a standardized data mining model language called Predictive Model Markup Language (PMML).

In contrast to the general-purpose suites, application-specific tools are intended for particular analytic applications such as credit scoring, customer retention, and product marketing. Their focus may be further sharpened to address the needs of specialized markets such as mortgage lending or financial services. The target user is an analyst with expertise in the application domain. Therefore the interfaces, the algorithms, and even the terminology are customized for that particular industry, application, or customer. While less flexible than general-purpose tools, they offer the advantage of already incorporating domain knowledge into the product design, and can provide very good solutions with less effort. Data mining companies including SAS, IBM, and RapidMiner offer vertical market tools, as do industry specialists such as Fair Isaac. Other companies, such as Domo, are focusing on creating dashboards with analytics and visualizations for business intelligence.

Another technological shift has occurred with the spread of open source model building tools and open core tools. A somewhat simplified view of open source software is that the source code for the tool is available at no charge to the community of users and can be modified or enhanced by them. These enhancements are submitted to the originator or copyright holder, who can add them to the base package. Open core is a more recent approach in which a core set of functionality remains open and free, but there are proprietary extensions that are not free.

The most important open source statistical analysis software is R. R is descended from a Bell Labs program called S, which was commercialized as S+. Many data mining algorithms have been added to R, along with a plethora of statistics, data management tools, and visualization tools. Because it is essentially a programming language, R has enormous flexibility but a steeper learning curve than many of the GUI-based tools. Although there are some GUIs for R, the overwhelming majority of use is through programming.

Some vendors, as well as the open source community, are adding statistical and data mining tools to Python, a popular programming language that is generally easier to use than C++ or Java, and faster than R.

As mentioned above, the cloud-computing vendors have moved into the data mining/predictive analytics business by offering AaaS (Analytics as a Service) and pricing their products on a transaction basis. These products are oriented more toward application developers than business intelligence analysts. A big part of the attraction of mining data in the cloud is the ability to store and manage enormous amounts of data without requiring the expense and complexity of building an in-house capability. This can also enable a more rapid implementation of large distributed multi-user applications. Cloud based data can be used with non-cloud-based analytics if the vendors analytics do not meet the users needs.

Amazon has added Amazon Machine Learning to its Amazon Web Services (AWS), taking advantage of predictive modeling tools developed for Amazons internal use. AWS supports both relational databases and Hadoop data management. Models cannot be exported, because they are intended to be applied to data stored on the Amazon cloud.

Google is very active in cloud analytics with its BigQuery and Prediction API. BigQuery allows the use of Google infrastructure to access large amounts of data using a SQL-like interface. The Prediction API can be accessed from a variety of languages including R and Python. It uses a variety of machine learning algorithms and automatically selects the best results. Unfortunately, this is not a transparent process. Furthermore, as with Amazon, models cannot be exported.

Microsoft is an active player in cloud analytics with its Azure Machine Learning Studio and Stream Analytics. Azure works with Hadoop clusters as well as with traditional relational databases. Azure ML offers a broad range of

works with Hadoop clusters as well as with traditional relational databases. Azure ML offers a broad range of algorithms such as boosted trees and support vector machines as well as supporting R scripts and Python. Azure ML also supports a workflow interface making it more suitable for the nonprogrammer data scientist. The real-time analytics component is designed to allow streaming data from a variety of sources to be analyzed on the fly. XLMiner's cloud version is based on Microsoft Azure. Microsoft also acquired Revolution Analytics, a major player in the R analytics business, with a view to integrating Revolution's "R Enterprise" with SQL Server and Azure ML. R Enterprise includes extensions to R that eliminate memory limitations and take advantage of parallel processing.

One drawback of the cloud-based analytics tools is a relative lack of transparency and user control over the algorithms and their parameters. In some cases, the service will simply select a single model that is a black box to the user. Another drawback is that for the most part cloud-based tools are aimed at more sophisticated data scientists who are systems savvy.

Data science is playing a central role in enabling many organizations to optimize everything from production to marketing. New storage options and analytical tools promise even greater capabilities. The key is to select technology that's appropriate for an organization's unique goals and constraints. As always, human judgment is the most important component of a data mining solution.

This book's focus is on a comprehensive understanding of the different techniques and algorithms used in data mining, and less on the data management requirements of real-time deployment of data mining models. R makes it ideal for this purpose, and for exploration, prototyping, and piloting of solutions.

***

*Herb Edelstein is president of Two Crows Consulting (www.twocrows.com), a leading data mining consulting firm near Washington, DC. He is an internationally recognized expert in data mining and data warehousing, a widely published author on these topics, and a popular speaker.

Problems

1. Assuming that data mining techniques are to be used in the following cases, identify whether the task required is supervised or unsupervised learning.
   a. Deciding whether to issue a loan to an applicant based on demographic and financial data (with reference to a database of similar data on prior customers).
   b. In an online bookstore, making recommendations to customers concerning additional items to buy based on the buying patterns in prior transactions.
   c. Identifying a network data packet as dangerous (virus, hacker attack) based on comparison to other packets whose threat status is known.
   d. Identifying segments of similar customers.
   e. Predicting whether a company will go bankrupt based on comparing its financial data to those of similar bankrupt and nonbankrupt firms.
   f. Estimating the repair time required for an aircraft based on a trouble ticket.
   g. Automated sorting of mail by zip code scanning.
   h. Printing of custom discount coupons at the conclusion of a grocery store checkout based on what you just bought and what others have bought previously.
2. Describe the difference in roles assumed by the validation partition and the test partition.
3. Consider the sample from a database of credit applicants in Table 2.15. Comment on the likelihood that it was sampled randomly, and whether it is likely to be a useful sample.

Table 2.15 Sample from a database of credit applications

OBS	CHECK ACCT	DURATION	HISTORY	NEW CAR	USED CAR	FURNITURE	RADIO TV	EDUC	RETRAIN	AMOUNT	S A
1	0	6	4	0	0	0	1	0	0	1169	
8	1	36	2	0	1	0	0	0	0	6948	

8	1	36	2	0	1	0	0	0	0	6948
16	0	24	2	0	0	0	1	0	0	1282
24	1	12	4	0	1	0	0	0	0	1804
32	0	24	2	0	0	1	0	0	0	4020
40	1	9	2	0	0	0	1	0	0	458
48	0	6	2	0	1	0	0	0	0	1352
56	3	6	1	1	0	0	0	0	0	783
64	1	48	0	0	0	0	0	0	1	14421
72	3	7	4	0	0	0	1	0	0	730
80	1	30	2	0	0	1	0	0	0	3832
88	1	36	2	0	0	0	0	1	0	12612
96	1	54	0	0	0	0	0	0	1	15945
104	1	9	4	0	0	1	0	0	0	1919
112	2	15	2	0	0	0	0	1	0	392

4. Consider the sample from a bank database shown in Table 2.16; it was selected randomly from a larger database to be the training set. *Personal Loan* indicates whether a solicitation for a personal loan was accepted and is the response variable. A campaign is planned for a similar solicitation in the future and the bank is looking for a model that will identify likely responders. Examine the data carefully and indicate what your next step would be.

Table 2.16 Sample from a bank database

OBS	AGE	EXPERIENCE	INCOME	ZIP CODE	FAMILY	CC AVG	EDUC	MORTGAGE	PERSONAL LOAN	SECURITI ACCT
1	25	1	49	91107	4	1.6	1	0	0	1
4	35	9	100	94112	1	2.7	2	0	0	0
5	35	8	45	91330	4	1	2	0	0	0
9	35	10	81	90089	3	0.6	2	104	0	0
10	34	9	180	93023	1	8.9	3	0	1	0
12	29	5	45	90277	3	0.1	2	0	0	0
17	38	14	130	95010	4	4.7	3	134	1	0
18	42	18	81	94305	4	2.4	1	0	0	0
21	56	31	25	94015	4	0.9	2	111	0	0
26	43	19	29	94305	3	0.5	1	97	0	0
29	56	30	48	94539	1	2.2	3	0	0	0
30	38	13	119	94104	1	3.3	2	0	1	0
35	31	5	50	94035	4	1.8	3	0	0	0
36	48	24	81	92647	3	0.7	1	0	0	0
37	59	35	121	94720	1	2.9	1	0	0	0
38	51	25	71	95814	1	1.4	3	198	0	0
39	42	18	141	94114	3	5	3	0	1	1
41	57	32	84	92672	3	1.6	3	0	0	1

5. Using the concept of overfitting, explain why when a model is fit to training data, zero error with those data is not necessarily good.

6. In fitting a model to classify prospects as purchasers or nonpurchasers, a certain company drew the training data from internal data that include demographic and purchase information. Future data to be classified will be lists

purchased from other sources, with demographic (but not purchase) data included. It was found that "refund issued" was a useful predictor in the training data. Why is this not an appropriate variable to include in the model?

7. A dataset has 1000 records and 50 variables with 5% of the values missing, spread randomly throughout the records and variables. An analyst decides to remove records with missing values. About how many records would you expect to be removed?
8. Normalize the data in Table 2.17, showing calculations.

Table 2.17

Age	Income ($)
25	49,000
56	156,000
65	99,000
32	192,000
41	39,000
49	57,000

9. Statistical distance between records can be measured in several ways. Consider Euclidean distance, measured as the square root of the sum of the squared differences. For the first two records in Table 2.17, it is

$$\sqrt{(25-56)^2+(49,000-156,000)^2}.$$

Can normalizing the data change which two records are farthest from each other in terms of Euclidean distance?

10. Two models are applied to a dataset that has been partitioned. Model A is considerably more accurate than model B on the training data, but slightly less accurate than model B on the validation data. Which model are you more likely to consider for final deployment?
11. The dataset *ToyotaCorolla.csv* contains data on used cars on sale during the late summer of 2004 in the Netherlands. It has 1436 records containing details on 38 attributes, including *Price*, *Age*, *Kilometers*, *HP*, and other specifications.
    a. Explore the data using the data visualization capabilities of R. Which of the pairs among the variables seem to be correlated?
    b. We plan to analyze the data using various data mining techniques described in future chapters. Prepare the data for use as follows:
        i. The dataset has two categorical attributes, *Fuel Type* and *Metallic*. Describe how you would convert these to binary variables. Confirm this using R's functions to transform categorical data into dummies.
        ii. Prepare the dataset (as factored into dummies) for data mining techniques of supervised learning by creating partitions in R. Select all the variables and use default values for the random seed and partitioning percentages for training (50%), validation (30%), and test (20%) sets. Describe the roles that these partitions will play in modeling.

Notes

[1] Harney, K., "Zestimates may not be as right as you'd like", *Washington Post*, Feb. 7, 2015, p. T10.

[2] The data are a slightly cleaned version of the Property Assessment FY2014 data at https://data.boston.gov/dataset/property-assessment (accessed December 2017).

[3] The full data dictionary provided by the City of Boston is available at https://data.boston.gov/dataset/property-assessment; we have modified a few variable names.