(a) Use smoothing by bin means to smooth the above data, using a bin depth of 3. Illustrate your steps. Comment on the effect of this technique for the given data.

(b) How might you determine outliers in the data?

(c) What other methods are there for data smoothing?

4. Discuss issues to consider during data integration.

5. Using the data for *age* given in Question 3, answer the following:

(a) Use min-max normalization to transform the value 35 for *age* onto the range [0, 1].

(b) Use z-score normalization to transform the value 35 for *age*, where the standard deviation of *age* is ??.

(c) Use normalization by decimal scaling to transform the value 35 for *age*.

(d) Comment on which method you would prefer to use for the given data, giving reasons as to why.

6. Use a flow-chart to illustrate the following procedures for attribute subset selection:

(a) step-wise forward selection.

(b) step-wise backward elimination

(c) a combination of forward selection and backward elimination.

7. Using the data for *age* given in Question 3:

(a) Plot an equi-width histogram of width 10.

(b) Sketch examples of each of the following sample techniques: SRSWOR, SRSWR, cluster sampling, stratified sampling.

8. Propose a concept hierarchy for the attribute *age* using the 3-4-5 partition rule.

9. Propose an algorithm, in pseudo-code or in your favorite programming language, for

(a) the automatic generation of a concept hierarchy for categorical data based on the number of distinct values of attributes in the given schema,

(b) the automatic generation of a concept hierarchy for numeric data based on the *equi-width* partitioning rule, and

(c) the automatic generation of a concept hierarchy for numeric data based on the *equi-depth* partitioning rule.

Bibliographic Notes

Data preprocessing is discussed in a number of textbooks, including Pyle [28], Kennedy et al. [21], and Weiss and Indurkhya [37]. More specific references to individual preprocessing techniques are given below.

For discussion regarding data quality, see Ballou and Tayi [3], Redman [31], Wand and Wang [35], and Wang, Storey and Firth [36]. The handling of missing attribute values is discussed in Quinlan [29], Breiman et al. [5], and Friedman [11]. A method for the detection of outlier or "garbage" patterns in a handwritten character database is given in Guyon, Matic, and Vapnik [14]. Binning and data normalization are treated in several texts, including [28, 21, 37].

A good survey of data reduction techniques can be found in Barbará et al. [4]. For algorithms on data cubes and their precomputation, see [33, 16, 1, 38, 32]. Greedy methods for attribute subset selection (or *feature subset selection*) are described in several texts, such as Neter et al. [24], and John [18]. A combination forward selection and backward elimination method was proposed in Siedlecki and Sklansky [34]. For a description of wavelets for data compression, see Press et al. [27]. Daubechies transforms are described in Daubechies [6]. The book by Press et al. [27] also contains an introduction to singular value decomposition for principal components analysis.

An introduction to regression and log-linear models can be found in several textbooks, such as [17, 9, 20, 8, 24]. For log-linear models (known as *multiplicative models* in the computer science literature), see Pearl [25]. For a

general introduction to histograms, see [7, 4]. For extensions of single attribute histograms to multiple attributes, see Muralikrishna and DeWitt [23], and Poosala and Ioannidis [26]. Several references to clustering algorithms are given in Chapter 7 of this book, which is devoted to this topic. A survey of multidimensional indexing structures is in Gaede and Günther [12]. The use of multidimensional index trees for data aggregation is discussed in Aoki [2]. Index trees include R-trees (Guttman [13]), quad-trees (Finkel and Bentley [10]), and their variations. For discussion on sampling and data mining, see John and Langley [19], and Kivinen and Mannila [22].

Entropy and information gain are described in Quinlan [30]. Concept hierarchies, and their automatic generation from categorical data are described in Han and Fu [15].

Bibliography

[1] S. Agarwal, R. Agrawal, P. M. Deshpande, A. Gupta, J. F. Naughton, R. Ramakrishnan, and S. Sarawagi. On the computation of multidimensional aggregates. In *Proc. 1996 Int. Conf. Very Large Data Bases*, pages 506–521, Bombay, India, Sept. 1996.

[2] P. M. Aoki. Generalizing "search" in generalized search trees. In *Proc. 1998 Int. Conf. Data Engineering (ICDE'98)*, April 1998.

[3] D. P. Ballou and G. K. Tayi. Enhancing data quality in data warehouse environments. *Communications of ACM*, 42:73–78, 1999.

[4] D. Barbará et al. The new jersey data reduction report. *Bulletin of the Technical Committee on Data Engineering*, 20:3–45, December 1997.

[5] L. Breiman, J. Friedman, R. Olshen, and C. Stone. *Classification and Regression Trees*. Wadsworth International Group, 1984.

[6] I. Daubechies. *Ten Lectures on Wavelets*. Capital City Press, Montpelier, Vermont, 1992.

[7] J. Devore and R. Peck. *Statistics: The Exploration and Analysis of Data*. New York: Duxbury Press, 1997.

[8] J. L. Devore. *Probability and Statistics for Engineering and the Science, 4th ed.* Duxbury Press, 1995.

[9] A. J. Dobson. *An Introduction to Generalized Linear Models*. Chapman and Hall, 1990.

[10] R. A. Finkel and J. L. Bentley. Quad-trees: A data structure for retrieval on composite keys. *ACTA Informatica*, 4:1–9, 1974.

[11] J. H. Friedman. A recursive partitioning decision rule for nonparametric classifiers. *IEEE Trans. on Comp.*, 26:404–408, 1977.

[12] V. Gaede and O. Günther. Multdimensional access methods. *ACM Comput. Surv.*, 30:170–231, 1998.

[13] A. Guttman. R-tree: A dynamic index structure for spatial searching. In *Proc. 1984 ACM-SIGMOD Int. Conf. Management of Data*, June 1984.

[14] I. Guyon, N. Matic, and V. Vapnik. Discoverying informative patterns and data cleaning. In U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, editors, *Advances in Knowledge Discovery and Data Mining*, pages 181–203. AAAI/MIT Press, 1996.

[15] J. Han and Y. Fu. Dynamic generation and refinement of concept hierarchies for knowledge discovery in databases. In *Proc. AAAI'94 Workshop Knowledge Discovery in Databases (KDD'94)*, pages 157–168, Seattle, WA, July 1994.

[16] V. Harinarayan, A. Rajaraman, and J. D. Ullman. Implementing data cubes efficiently. In *Proc. 1996 ACM-SIGMOD Int. Conf. Management of Data*, pages 205–216, Montreal, Canada, June 1996.

[17] M. James. *Classification Algorithms*. John Wiley, 1985.

[18] G. H. John. *Enhancements to the Data Mining Process*. Ph.D. Thesis, Computer Science Dept., Stanford Univeristy, 1997.