- assume each X_j takes values in a set S_j
- let $s_j \subseteq S_j$ be a subset of these values
- example: age classes (0–14, 15–24, ...)
- example: employment status (working full-time, working part-time, seeking work, ...)
- Goal: find $s_1, s_2, \dots s_p$ so that

$$\Pr(X_j \in s_j, j = 1, \dots, p) = \Pr\{\cap_{j=1}^{p}(X_j \in s_j)\}$$

  relatively large
- Note if $s_j = S_j$ then $\Pr(X_j \in s_j) = 1$, i.e. X_j "does not appear"
- Simplification: s_j either S_j or a single value (called v_{0j} on p.440)
- Then want to find subsets $\mathcal{J} \subset \{1, \dots, p\}$ and values $v_{0j}, j \in \mathcal{J}$ so that $\Pr(\cap_{\mathcal{J}} S_j = v_{0j})$ is large

- Special case: each $X_j = 0, 1$ (binary features) then $v_{0j} = 1$ and $\cap_{j \in \mathcal{J}}(X_j = 1) \Rightarrow \prod_{j \in \mathcal{J}} X_j = 1$
- If X_j takes a finite number of values, $v_{j1}, \ldots v_{jn_j}$, say, then create n_j dummy variables $Z_{j1}, Z_{j2}, \ldots, Z_{jn_j}$ that are binary
- Renumber these to $Z_1, \ldots, Z_K$; goal is now to find a subset $\mathcal{K} \subset \{1, \ldots K\}$ to give a large value of

$$\Pr(\prod_{k \in \mathcal{K}} Z_k = 1)$$

- This is estimated by

$$\frac{1}{N} \sum_{i=1}^{N} \prod_{k \in \mathcal{K}} z_{ik} = \widehat{\Pr}(\prod_{\mathcal{K}} Z_k = 1) \equiv T(\mathcal{K})$$

- Implementation: Find all sets $\mathcal{K}_\ell$ so that $T(\mathcal{K}_\ell) > t$: this reduces the number of possible item sets.

- $\mathcal{K}$ is an item set and

$$T(\mathcal{K}) = \frac{1}{N}\sum_{i=1}^{N}\prod_{k\in\mathcal{K}} z_{ik}$$

  is the prevalence of the item set $\mathcal{K}$.
- §14.2.2 describes the APriori algorithm
- The item sets $\mathcal{K}_\ell$ are described by a set of association rules $A \Rightarrow B$
- example $\{\text{peanut butter, jelly}\} \Rightarrow \{\text{bread}\}$
- and summarized by estimates of

$$\begin{array}{lcl} T(A \Rightarrow B) & \Pr(A \cap B) & \text{"support"} \\ C(A \Rightarrow B) & \Pr(B \mid A) & \text{"confidence"} \\ & \dfrac{\Pr(A \cap B)}{\Pr(A)\Pr(B)} & \text{"lift"} \end{array}$$

- See §14.2.3 for an example (that gave 6288 rules!)

- $\mathcal{K}$ is an item set and

$$T(\mathcal{K}) = \frac{1}{N}\sum_{i=1}^{N}\prod_{k\in\mathcal{K}} z_{ik}$$

  is the prevalence of the item set $\mathcal{K}$.
- §14.2.2 describes the APriori algorithm
- The item sets $\mathcal{K}_\ell$ are described by a set of association rules $A \Rightarrow B$
- example $\{\text{peanut butter, jelly}\} \Rightarrow \{\text{bread}\}$
- and summarized by estimates of

$$\begin{array}{lcl} T(A \Rightarrow B) & \Pr(A \cap B) & \text{“support”} \\ C(A \Rightarrow B) & \Pr(B \mid A) & \text{“confidence”} \\ & \dfrac{\Pr(A \cap B)}{\Pr(A)\Pr(B)} & \text{“lift”} \end{array}$$

- See §14.2.3 for an example (that gave 6288 rules!)

If we are interested in a particular consequence, $P(B \mid A)$, we could create a ‘response’ variable $y = 1\{x \in B\}$ and use methods for supervised learning such as logistic regression, classification, etc.

A more clever use of supervised learning for association rules is described in §14.2.4 and §14.2.5, suggestion in §14.2.6 to use CART

- §14.3.7: Mixture models for clustering:
  $f(x) = \sum_{k=1}^{K} \pi_k f_k(x; \theta_k)$
- (π_k, θ_k) to be estimated by maximum likelihood (EM algorithm)
- Methods related to principal components:
- $X = UDV^T$, V, U are orthogonal, D is diagonal
- $z_1 = Xv_1, z_2 = Xv_2, \ldots$ are the first, second principal components
- principal curves do this construction locally
- self-organizing maps use a binned version of data (`batchSOM`)
- independent component analysis seeks vectors with slightly different properties
- all of these methods form the basis for various graphical displays

- Regression: linear, ridge, lasso, logistic, polynomial splines, smoothing splines, kernel methods, additive models, regression trees, MARS, projection pursuit, neural networks Chapters 3, 5, 9, 11
- Classification: logistic regression, linear discriminant analysis, generalized additive models, kernel methods, naive Bayes, classification trees, support vector machines, neural networks Chapters 4, 6, 9, 11, 12
- Model Selection: AIC, cross-validation, test error and training error Chapter 7
- Unsupervised learning: Kmeans clustering, hierarchical clustering, assocation rules Chapter 14
- Left out: , k-nearest neighbours, boosting and bagging, flexible discriminant analysis, mixture discriminant analysis, prototype methods, self-organizing maps, independent components analysis Chapters 10, 12, 13, 14
- Suggestion: Read Chapter 2

- use the unsupervised method K-means for supervised learning on the K-class problem
- K-means: choose a number (R) of cluster centers, for each center identify training points closer to it than to any other center, compute the means of the new clusters to use as cluster centers for the next iteration
- for classification: do this on the training data separately for each of the K classes
- the cluster centers are now called prototypes
- assign a class label to each of the R prototypes in each of the K classes
- classify a new point with feature vector x to the class of the closest prototype
- Figure 13.1: 3 classes, 5 prototypes per class
- Figure 13.2: 2 classes, 5 prototypes per class

- model-free, memory based
- given a query point, x_0, find the k training points closest in (Euclidean) distance
- classify using majority vote (break ties at random)
- 1-nearest neighbours has low bias, high variance, k large has high bias, low variance
- possible to get bounds on the best possible error rate, see p.417,8
- Figures 13.3–13.6

References

- Pattern Recognition and Neural Networks. B.D. Ripley (1996), Cambridge University Press. *Good discussion of many machine learning methods.*
- Classification (2nd ed.), A. D. Gordon (1999), Chapman & Hall/CRC Press. *Unsupervised learning/clustering; see Ch. 2 for good description of dissimilarity measures.*
- Finding Groups in Data: An Introduction to Cluster Analysis, L. Kaufman and P.J. Rousseeuw, (1990) Wiley. *Learn all about daisy, agnes, and many other of R's clustering methods.*
- Modern Applied Statistics with S (4th Ed.), W.N. Venables and B.D. Ripley (2002), Springer-Verlag. *The bible for computing with Splus and R; Ch. 11 covers unsupervised learning, Chs. 8,9 and 12 cover supervised learning.*
- Principles of Data Mining. D. Hand, H. Mannila, P. Smyth (2001) MIT Press. *Nice blend of computer science and statistical methods. Clustering covered in Ch. 9*

Pattern Recognition and Neural Networks. B.D. Ripley (1996), Cambridge University Press.

1. Introduction
2. Statistical Decision Theory
3. Linear Discriminant Analysis
4. Flexible Discriminants
5. Feed-forward Neural Networks
6. Nonparametric Methods
7. Tree-structured classifiers
8. Belief Networks
9. Unsupervised Methods
10. Finding Good Pattern Features

Principles of Data Mining. D. Hand, H. Mannila, P. Smyth (2001) MIT Press.

1. Introduction
2. Measurement and Data
3. Visualizing and Exploring Data
4. Data Analysis and Uncertainty
5. A Systematic Overview of Data Mining Algorithms
6. Models and Patterns
7. Score Functions for Data Mining Algorithms
8. Search and Optimization Methods
9. Descriptive Modeling
10. Predictive Modeling for Classification
11. Predictive Modeling for Regression
12. Data Organization and databases
13. Finding Patterns and Rules
14. Retrieval by Content

Pattern Recognition

Unsupervised

Statistical Theory

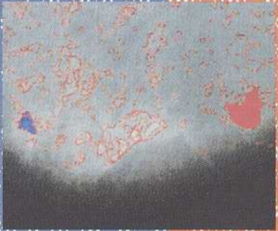
Feature Selection

Linear Methods

Flexible Discriminants

Neural Networks

Non-parametric Methods



A Review of Software Packages for Data Mining

Dominique HAUGHTON, Joel DEICHMANN, Abdolreza ESHGHI,
Selin SAYEK, Nicholas TEEBAGY, and Heikki TOPI

We present to the statistical community an overview of five data mining packages with the intent of leaving the reader with a sense of the different capabilities, the ease or difficulty of use, and the user interface of each package. We are not attempting to perform a controlled comparison of the algorithms in each package to decide which has the strongest predictive power, but instead hope to give an idea of the approach to predictive modeling used in each of them. The packages are compared in the areas of descriptive statistics and graphics, predictive models, and association (market basket) analysis.

As expected, the packages affiliated with the most popular statistical software packages (SAS and SPSS) provide the broadest range of features with remarkably similar modeling and interface approaches, whereas the other packages all have their special sets of features and specific target audiences whom we believe each of the packages will serve well. It is essential that an organization considering the purchase of a data mining package carefully evaluate the available options and choose the one that provides the best fit with its particular needs.

KEY WORDS: Clementine; Ghostminer; Quadstone; SAS Enterprise Miner; XLMiner.

1. INTRODUCTION

The term "data mining" has come to refer to a set of techniques that originated in statistics, computer science, and related areas that are typically used in the context of large datasets. The purpose of data mining is to reveal previously hidden associations between variables that are potentially relevant for managerial decision making. The exploratory and modeling techrithms in each package to decide which has the strongest predictive power, but instead aim to give an idea of the approach to predictive modeling used in each of them.

The article is structured as follows: we first outline the methodology we used to evaluate the packages and give a summary of key characteristics of each package. We continue by focusing on descriptive statistics and exploratory graphs. The section that follows is devoted to predictive modeling, covering model building and assessment. A section on association (market basket) analysis is then provided, followed by a conclusion.

2. METHODOLOGY

The list of packages we have selected for this review is by no means exhaustive. We have chosen to cover the data mining packages associated with the two leading statistical packages, SAS and SPSS. We also decided to review two "stand-alone" packages, GhostMiner and Quadstone, and an Excel add-on, XLMiner.

We compare the packages in the areas of descriptive statistics and graphics, predictive models, and association (market basket) analysis. Predictive modeling is one of the main applications of data mining, and exploratory descriptive analyses always precede modeling efforts. Association analysis, in which "baskets" of goods purchased together are identified, is also very commonly used.

For the descriptive and modeling analysis, we used the Direct Marketing Educational Foundation dataset 2, merged with Census geo-demographic variables from dataset 6 (www.the-dma.org/dmef). The dataset contains 19,185 observations and concerns a business with multiple divisions, each mailing different catalogs to a unified customer database. The target variable, BUY10, equals unity if a customer made a purchase from