Data Mining Techniques

Course Summary

Description

This course provides the students with the skills necessary to set up, execute, and interpret the output from data mining analysis tools. This course is based on the following book: Data Mining Techniques, 3rd Edition, by Gordon S. Linoff and Michael J. A. Berry, published April, 2011 by Wiley Publishing, Inc, ISBN: 978-0-470-65093-6. The course is usually taught in one, two, or three days. Roughly 7 chapters are taught per day. The first 7 chapters are taught in 1 day, 14 chapters are taught in 2 days, and all 21 chapters are taught in 3 days. If the students wish, the instructor and the students may decide to skip around the chapters.

Topics

- What Is Data Mining and Why Do It?
- Data Mining Applications in Marketing and Customer Relationship Management
- The Data Mining Process
- What You Should Know About Data
- Descriptions and Prediction: Profiling and Predictive Modeling
- Data Mining Using Classic Statistical Techniques
- Decision Trees
- Artificial Neural Networks
- Nearest Neighbor Approaches: Memory-Based Reasoning and Collaborative Filtering
- Knowing When to Worry: Using Survival Analysis to Understand Customers
- Genetic Algorithms and Swarm Intelligence
- Tell Me Something New: Pattern Discovery and Data Mining
- Alternative Approaches to Cluster Detection
- Market Basket Analysis and Association Rules
- Link Analysis
- Data Warehousing, OLAP, Analytic Sandboxes, and Data Mining
- Building Customer Signatures
- Derived Variables: Making the Data Mean More
- Too Much of a Good Thing? Techniques for Reducing the Number of Variables
- Listen Carefully to What Your Customers Say: Text Mining

Audience

This course is intended for users, power users, programmers, analysts, DBAs, Data Modelers, or anyone else who needs to do data mining.

Prerequisites

Students should have at least some experience with coding SQL for any relational database management system plus at least a conceptual understanding of Data Warehousing.

Duration

One to three days

Due to the nature of this material, this document refers to numerous hardware and software products by their trade names. References to other companies and their products are for informational purposes only, and all trademarks are the properties of their respective companies. It is not the intent of ProTech Professional Technical Services, Inc. to use any of these names generically

ProTech
protechtraining.com

"Charting the Course ...

... to Your Success!"

Data Mining Techniques

Course Outline

I. What Is Data Mining and Why Do It?
- A. What Is Data Mining?
- B. Why Now?
- C. Skills for the Data Miner
- D. The Virtuous Cycle of Data Mining
- E. A Case Study in Business Data Mining
- F. Steps of the Virtuous Cycle
- G. Data Mining in the Context of the Virtuous Cycle

II. Data Mining Applications in Marketing and Customer Relationship Management
- A. Two Customer Lifecycles
- B. Organize Business Processes Around the Customer Lifecycle
- C. Data Mining Applications for Customer Acquisition
- D. A Data Mining Example: Choosing the Right Place to Advertise
- E. Data Mining to Improve Direct Marketing Campaigns
- F. Using Current Customers to Learn About Prospects
- G. Data Mining Applications for Customer Relationship Management
- H. Retention
- I. Beyond the Customer Lifecycle

III. The Data Mining Process
- A. What Can Go Wrong?
- B. Data Mining Styles
- C. Goals, Tasks, and Techniques
- D. Formulating Data Mining Problems: From Goals to Tasks to Techniques
- E. What Techniques for Which Tasks?

IV. What You Should Know About Data
- A. Occam's Razor
- B. Looking At and Measuring Data
- C. Measuring Response
- D. Multiple Comparisons
- E. Chi-Square Test
- F. An Example: Chi-Square for Regions and Starts
- G. Case Study: Comparing Two Recommendation Systems with an A/B Test
- H. Data Mining and Statistics

V. Descriptions and Prediction: Profiling and Predictive Modeling
- A. Directed Data Mining Models
- B. Directed Data Mining Methodology
- C. Step 1: Translate the Business Problem into a Data Mining Problem
- D. Step 2: Select Appropriate Data
- E. Step 3: Get to Know the Data
- F. Step 4: Create a Model Set
- G. Step 5: Fix Problems with the Data
- H. Step 6: Transform Data to Bring Information to the Surface
- I. Step 7: Build Models
- J. Step 8: Assess Models
- K. Step 9: Deploy Models
- L. Step 10: Assess Results
- M. Step 11: Begin Again

VI. Data Mining Using Classic Statistical Techniques
- A. Similarity Models
- B. Table Lookup Models
- C. RFM: A Widely Used Lookup Model
- D. Naïve Bayesian Models
- E. Linear Regression
- F. Multiple Regression
- G. Logistic Regression
- H. Fixed Effects and Hierarchical Effects

VII. Decision Trees
- A. What Is a Decision Tree and How Is It Used?
- B. Decision Trees Are Local Models
- C. Growing Decision Trees
- D. Finding the Best Split
- E. Pruning
- F. Decision Tree Variations
- G. Assessing the Quality of a Decision Tree
- H. When Are Decision Trees Appropriate?
- I. Case Study: Process Control in a Coffee Roasting Plant

Due to the nature of this material, this document refers to numerous hardware and software products by their trade names. References to other companies and their products are for informational purposes only, and all trademarks are the properties of their respective companies. It is not the intent of ProTech Professional Technical Services, Inc. to use any of these names generically

Data Mining Techniques

Course Outline (cont'd)

VIII. Artificial Neural Networks
- A. A Bit of History
- B. The Biological Model
- C. Artificial Neural Networks
- D. A Sample Application: Real Estate Appraisal
- E. Training Neural Networks
- F. Radial Basis Function Networks
- G. Neural Networks in Practice
- H. Choosing the Training Set
- I. Preparing the Data
- J. Interpreting the Output from a Neural Network
- K. Neural Networks for Time Series
- L. Can Neural Network Models Be Explained?

IX. Nearest Neighbor Approaches: Memory-Based Reasoning and Collaborative Filtering
- A. Memory-Based Reasoning
- B. Challenges of MBR
- C. Case Study: Using MBR for Classifying Anomalies in Mammograms
- D. Measuring Distance and Similarity
- E. The Combination Function: Asking the Neighbors for Advice
- F. Case Study: Shazam — Finding Nearest Neighbors for Audio Files
- G. Collaborative Filtering: A Nearest-Neighbor Approach to Making Recommendations

X. Knowing When to Worry: Using Survival Analysis to Understand Customers
- A. Customer Survival
- B. Hazard Probabilities
- C. From Hazards to Survival
- D. Proportional Hazards
- E. Survival Analysis in Practice

XI. Genetic Algorithms and Swarm Intelligence
- A. Optimization
- B. Genetic Algorithms
- C. The Traveling Salesman Problem
- D. Case Study: Using Genetic Algorithms for Resource Optimization
- E. Case Study: Evolving a Solution for Classifying Complaints

XII. Tell Me Something New: Pattern Discovery and Data Mining
- A. Undirected Techniques, Undirected Data Mining
- B. What is Undirected Data Mining?
- C. Methodology for Undirected Data Mining

XIII. Finding Islands of Similarity: Automatic Cluster Detection
- A. Searching for Islands of Simplicity
- B. Customer Segmentation and Clustering
- C. The K-Means Clustering Algorithm
- D. Interpreting Clusters
- E. Evaluating Clusters
- F. Case Study: Clustering Towns
- G. Variations on K-Means
- H. Data Preparation for Clustering

XIV. Alternative Approaches to Cluster Detection
- A. Shortcomings of K-Means
- B. Gaussian Mixture Models
- C. Divisive Clustering
- D. Agglomerative (Hierarchical) Clustering
- E. Self-Organizing Maps
- F. The Search Continues for Islands of Simplicity

XV. Market Basket Analysis and Association Rules
- A. Defining Market Basket Analysis
- B. Case Study: Spanish or English
- C. Association Analysis
- D. Building Association Rules
- E. Extending the Ideas
- F. Association Rules and Cross-Selling
- G. Sequential Pattern Analysis

XVI. Link Analysis
- A. Basic Graph Theory
- B. Social Network Analysis
- C. Mining Call Graphs
- D. Case Study: Tracking Down the Leader of the Pack
- E. Case Study: Who Is Using Fax Machines from Home?
- F. How Google Came to Rule the World

Due to the nature of this material, this document refers to numerous hardware and software products by their trade names. References to other companies and their products are for informational purposes only, and all trademarks are the properties of their respective companies. It is not the intent of ProTech Professional Technical Services, Inc. to use any of these names generically

Data Mining Techniques

Course Outline (cont'd)

XVII. Data Warehousing, OLAP, Analytic Sandboxes, and Data Mining
- A. The Architecture of Data
- B. A General Architecture for Data Warehousing
- C. Analytic Sandboxes
- D. Where Does OLAP Fit In?
- E. Where Data Mining Fits in with Data Warehousing

XVIII. Building Customer Signatures
- A. Finding Customers in Data
- B. Designing Signatures
- C. What a Signature Looks Like
- D. Process for Creating Signatures
- E. Dealing with Missing Values

XIX. Derived Variables: Making the Data Mean More
- A. Handset Churn Rate as a Predictor of Churn
- B. Single-Variable Transformations
- C. Combining Variables
- D. Extracting Features from Time Series
- E. Extracting Features from Geography
- F. Using Model Scores as Inputs
- G. Handling Sparse Data
- H. Capturing Customer Behavior from Transactions

XX. Too Much of a Good Thing? Techniques for Reducing the Number of Variables
- A. Problems with Too Many Variables
- B. The Sparse Data Problem
- C. Flavors of Variable Reduction Techniques
- D. Sequential Selection of Features
- E. Other Directed Variable Selection Methods
- F. Principal Components
- G. Variable Clustering

XXI. Listen Carefully to What Your Customers Say: Text Mining
- A. What Is Text Mining?
- B. Working with Text Data
- C. Case Study: Ad Hoc Text Mining
- D. Classifying News Stories Using MBR
- E. From Text to Numbers
- F. Text Mining and Naïve Bayesian Models
- G. DIRECTV: A Case Study in Customer Service

Due to the nature of this material, this document refers to numerous hardware and software products by their trade names. References to other companies and their products are for informational purposes only, and all trademarks are the properties of their respective companies. It is not the intent of ProTech Professional Technical Services, Inc. to use any of these names generically