

Oracle Database 11g: Data Mining Techniques

Student Guide

D73528GC10

Edition 1.0

February 2012

D76223

ORACLE®

Author

Brian Pottle

**Technical Contributors
and Reviewers**

Charlie Berger

Mark Kelly

Dan Vlamis

Tim Vlamis

Karl Rexer

Editors

Richard Wallis

Daniel Milne

Graphic Designer

Rajiv Chandrabhanu

Publishers

Veena Narasimhan

Joseph Fernandez

Nita Brozowski

Copyright © 2012, Oracle and/or its affiliates. All rights reserved.

Disclaimer

This document contains proprietary information and is protected by copyright and other intellectual property laws. You may copy and print this document solely for your own use in an Oracle training course. The document may not be modified or altered in any way. Except where your use constitutes "fair use" under copyright law, you may not use, share, download, upload, copy, print, display, perform, reproduce, publish, license, post, transmit, or distribute this document in whole or in part without the express authorization of Oracle.

The information contained in this document is subject to change without notice. If you find any problems in the document, please report them in writing to: Oracle University, 500 Oracle Parkway, Redwood Shores, California 94065 USA. This document is not warranted to be error-free.

Restricted Rights Notice

If this documentation is delivered to the United States Government or anyone using the documentation on behalf of the United States Government, the following notice is applicable:

U.S. GOVERNMENT RIGHTS

The U.S. Government's rights to use, modify, reproduce, release, perform, display, or disclose these training materials are restricted by the terms of the applicable Oracle license agreement and/or the applicable U.S. Government contract.

Trademark Notice

Oracle and Java are registered trademarks of Oracle and/or its affiliates. Other names may be trademarks of their respective owners.

Contents

Preface

1 Introduction

- Objectives 1-2
- Course Goals 1-3
- Related Courses and Resources 1-4
- Course Agenda 1-5
- Summary 1-6

2 Data Mining Concepts and Terminology

- Objectives 2-2
- What Is Data Mining? 2-3
- Who Uses Data Mining? 2-4
- Defining Key Data Mining Properties 2-5
- Data Mining Categories 2-7
- Examining Supervised Data Mining 2-8
- Supervised Data Mining Steps 2-9
- Supervised Data Mining Techniques 2-11
- Supervised Data Mining Algorithms 2-12
- Examining Unsupervised Data Mining 2-13
- Unsupervised Data Mining Techniques 2-14
- Unsupervised Data Mining Algorithms 2-15
- What Data Mining Can and Cannot Do 2-16
- Summary 2-17

3 The Data Mining Process

- Objectives 3-2
- Steps in the Data Mining Process 3-3
- Step 1: Problem Definition 3-4
- Step 2: Data Selection, Preparation, and Preprocessing 3-5
- Data Preparation Tips 3-6
- Data Preparation: Aggregation 3-8
- Data Preparation: Mining Attribute Types 3-9
- Automatic Data Preparation (ADP) 3-10
- Step 3: Building and Evaluating Models 3-11
- Model-Building Tasks 3-12

Train and Test Models: Supervised Learning 3-13
Evaluating Model Results 3-14
Applying the Selected Model(s) 3-15
Step 4: Deploying Results 3-16
Summary 3-17

4 Introducing Oracle Data Miner 11g Release 2

Objectives 4-2
Oracle Data Mining Architecture 4-3
Oracle Data Mining Related Uses 4-5
Oracle Data Miner 11g Release 2 4-6
Setting Up Oracle Data Miner 11g Release 2 (Using SQL Developer 3.1) 4-7
Installing SQL Developer 4-8
Setting Up Data Miner Step 1: Create SQL Developer Connection for SYS 4-9
Setting Up Data Miner Step 2: Create Database Account for Data Mining User 4-10
Setting Up Data Miner Step 3: Create a Connection for Data Mining User 4-13
Setting Up Data Miner Step 4: Enable the Data Miner GUI and User 4-14
Setting Up Data Miner Step 5: Install the Oracle Data Miner Repository 4-15
Data Miner Repository Installation Process 4-16
Creating a Data Miner Project and Workflow 4-17
Introducing the Data Miner Interface 4-18
Examining Oracle Data Miner Nodes 4-19
Previewing a Data Miner Workflow 4-21
Summary 4-22

5 Using Classification Models

Objectives 5-2
Revisiting Supervised Learning 5-3
Supervised Learning: Examining Classification Techniques 5-4
Using Classification Models 5-5
Building a Workflow 5-6
Starting Point for Any Workflow: A Data Source 5-7
Adding a Data Source to the Workflow 5-8
Defining a Data Source: Step 1 5-9
Viewing Source Data 5-10
Defining a Data Source: Step 2 5-11
Completed Data Source Node 5-12
Previewing the Source Data 5-13
Adding an Explore Data Node 5-14
Connecting Workflow Nodes 5-15
Specifying a “Group By” for the Explore Data Node 5-16

- Running the Explore Data Node 5-17
- Viewing the Explore Data Results 5-18
- Examining the Results 5-19
- Creating Classification Models 5-20
- Adding a Classification Node 5-21
- Classification Node Default Properties 5-22
- Connecting Data Source and Class Build Nodes 5-23
- Selecting Target and Case ID Attributes 5-24
- Viewing Advanced Settings 5-25
- Accessing Help 5-26
- Specifying Advanced Settings 5-27
- Specifying Test Properties 5-28
- Building the Models 5-29
- Successful Build Indicators 5-30
- Viewing the Event Log 5-31
- Comparing the Models 5-32
- Comparing Test Results: Performance 5-33
- Comparing Test Results: Lift 5-34
- Comparing Test Results: Performance Matrix 5-35
- Comparing Test Results: Profit 5-36
- Tuning a Model 5-37
- Selecting a Specific Model 5-38
- Navigating the Decision Tree Model Interface 5-39
- Examining the Decision Tree Test Results 5-40
- Viewing the Leaf Node in the Decision Tree 5-41
- Scoring: Applying the Model to Your Data 5-42
- Step 1: Specify the Desired Model 5-43
- Step 2: Add a New Data Source Node 5-44
- Modifying a Node Name 5-45
- Step 3: Add an Apply Node 5-46
- Step 4: Connect the Nodes 5-47
- Step 5: Creating Output Format 5-48
- Step 6: Run the Apply Node 5-49
- Viewing Predictive Results 5-50
- Sorting Results 5-51
- Examining Predictions 5-52
- Summary 5-53

6 Using Regression Models

- Objectives 6-2
- Supervised Learning: Examining Regression Techniques 6-3

Using Regression Models	6-4
Creating a New Workflow	6-5
Adding the Data Source	6-6
Understanding Target and Input Variables	6-7
Add, Connect, and Run an Explore Data Node	6-8
Exploring Source Data Statistics	6-9
Identifying Non-Input Variables	6-10
Filtering the Source Data	6-11
Deselecting Columns	6-12
Specifying Other Column Filter Settings	6-13
Running the Source Data Filter	6-14
Creating Regression Models	6-15
Adding a Regression Node	6-16
Regression Node Default Properties	6-17
Connecting and Editing the Regression Node	6-18
Viewing Advanced Settings: GLM Regression	6-19
Viewing Advanced Settings: SVM Regression	6-20
Building the Models	6-21
Successful Build Indicators	6-22
Viewing Models and Results	6-23
Viewing GLM Model Details	6-24
Viewing GLM Model Settings	6-25
Viewing GLM Model Coefficients	6-26
Sorting Model Coefficients	6-27
Viewing SVM Model Information	6-28
Viewing Test Results for the GLM Model	6-29
GLM Model Test Results: Residual Tab	6-30
GLM Model Test Results: Predicted and Actual	6-31
Viewing Test Results for the SVM Model	6-32
SVM Model Test Results: Predicted and Actual	6-33
Comparing Test Results	6-34
Comparing GLM and SVM Test Results	6-35
Scoring Data with the Selected Model	6-36
Step 1: Specify the Chosen Model	6-37
Step 2: Add a New Data Source Node	6-38
Remove Non-Input and Target Variables	6-39
Modifying the Node Name	6-40
Step 3: Add an Apply Node	6-41
Step 4: Connect the Nodes	6-42
Step 5a: Modify Output Format Remove Apply Columns	6-43
Step 5b: Modify Output Format Add Data Columns	6-44

Step 5c: Modify Output Format Set Column Order 6-45
Step 6: Run the Apply Node 6-46
Viewing Predictive Results 6-47
Sorting Results 6-48
Examining Predictions 6-49
Summary 6-50

7 Using Clustering Models

Objectives 7-2
Revisiting Unsupervised Data Mining 7-3
Unsupervised Data Mining: Clustering 7-4
Using a Clustering Model 7-5
Creating the New Workflow 7-6
Using Existing Workflow Objects 7-7
Connecting and Editing the Explore Data Node 7-8
Specifying a “Group By” for the Explore Data Node 7-9
Running and Viewing the Explore Data Node 7-10
Exploring Source Data Statistics 7-11
Creating Clustering Models 7-12
Adding a Clustering Node 7-13
Default Properties of Clustering Nodes 7-14
Connecting the Nodes and Modifying Node Properties 7-15
Examining Cluster Model Settings 7-16
K-Means Advanced Algorithm Settings 7-18
O-Cluster Advanced Algorithm Settings 7-20
Building the Models 7-21
Viewing the Models 7-22
Viewing K-Means Model Results 7-23
Navigating the K-Means Tree 7-24
Viewing Results with Cluster Rule 7-25
Viewing Other Cluster Definitions 7-26
Viewing Cluster Detail 7-27
Comparing Clusters 7-28
Multiple Row Selection with Cluster Compare 7-29
Renaming Clusters 7-30
Selecting Clusters by Name 7-31
Viewing the O-Cluster Model Results 7-32
Applying a Clustering Model 7-33
Creating Output Format: Apply Columns 7-34
Creating Output Format: Data Columns 7-35
Running the Model and Viewing Data 7-36

Viewing Results 7-37

Summary 7-38

8 Performing Market Basket Analysis

Objectives 8-2

Unsupervised Data Mining: Association Rules 8-3

Using an Association Rules Model 8-4

Creating a New Project and Workflow 8-5

Adding a Data Source from a New Schema 8-6

Adding New Schemas 8-7

Selecting the New Table or View 8-8

Previewing the Data Source 8-9

Examining Sorted Data 8-10

Creating an Association Rules Model 8-11

Adding an Association Node 8-12

Connecting Data Source and Assoc Build Nodes 8-13

Specifying Required Input Items 8-14

Specifying Advanced Settings 8-16

Building the Model 8-18

Viewing the Model 8-19

Viewing Model Results 8-20

Examining Model Rules 8-22

Creating Model Rule Filters 8-23

Applying Model Rule Filters 8-24

Viewing Model Itemsets 8-25

Creating Model Details Output 8-26

Viewing Model Details 8-27

Using Descriptive Columns in the Output 8-28

Adding a Join Node 8-29

Defining the Join 8-30

Specifying Output Columns for the Join 8-31

Defining a New Assoc Build Node 8-32

Viewing Model Results 8-33

Viewing Descriptive Model Rules 8-34

Creating New Model Details Output 8-35

Viewing Descriptive Model Details 8-36

Summary 8-37

9 Performing Anomaly Detection

Objectives 9-2

Unsupervised Data Mining: Anomaly Detection 9-3

- Using an Anomaly Detection Model 9-4
- Creating the New Workflow 9-5
- Adding a Data Source 9-6
- Browsing Source Data Attributes 9-7
- Previewing the Source Data 9-8
- Creating an Anomaly Detection Model 9-9
- Adding an Anomaly Detection Node 9-10
- Connecting the Nodes and Selecting the Case ID 9-11
- Specifying the SVM Algorithm Kernel Function 9-12
- Building the Model 9-13
- Viewing Model Results 9-14
- Viewing “Normal” Results 9-15
- Viewing “Anomalous” Results 9-16
- Comparing Model Results 9-17
- Applying the Model 9-18
- Creating Output Format: Apply Columns 9-19
- Creating Output Format: Data Columns 9-20
- Running the Model and Viewing Data 9-21
- Sorting Output Results 9-22
- Examining Results 9-23
- Summary 9-24

10 Deploying Data Mining Results

- Objectives 10-2
- Agenda 10-3
- Creating a Table or View of Results 10-4
- Modifying Output Format 10-5
- Generating the Output and Viewing Results 10-6
- Viewing Output Results 10-7
- Using Output Results 10-8
- Agenda 10-9
- SQL Code Automatically Generated by Data Miner 10-10
- Deploying SQL Code to the Clipboard 10-11
- Accessing Deployed Code 10-12
- Executing Deployed Code 10-13
- Modifying Deployed Code 10-14
- Summary 10-15

A Using SQL Developer

- Objectives A-2
- What Is Oracle SQL Developer? A-3

Specifications of SQL Developer	A-4
SQL Developer 3.1 Interface	A-5
Creating a Database Connection	A-7
Browsing Database Objects	A-10
Displaying the Table Structure	A-11
Browsing Files	A-12
Creating a Schema Object	A-13
Creating a New Table: Example	A-14
Using the SQL Worksheet	A-15
Executing SQL Statements	A-19
Saving SQL Scripts	A-20
Executing Saved Script Files: Method 1	A-21
Executing Saved Script Files: Method 2	A-22
Formatting the SQL Code	A-23
Using Snippets	A-24
Using Snippets: Example	A-25
Debugging Procedures and Functions	A-26
Database Reporting	A-27
Creating a User-Defined Report	A-28
Search Engines and External Tools	A-29
Setting Preferences	A-30
Resetting the SQL Developer Layout	A-32
Data Modeler in SQL Developer	A-33
Summary	A-34