Oracle Data Miner decision trees classify and analyze data.

Have you ever wondered which of your prospects are most likely to buy, or who will provide the most-lucrative deals? Who should you target in your marketing campaign, and what will be important to them when they call? Which products and versions of products provide your customers with what they need, and which ones fall short and reflect negatively on your company? Oracle Data Miner helps bridge the gap between business questions such as these and the technical and statistical tasks related to mining the data for answers.

Oracle Data Miner can analyze an existing data set from your data warehouse and classify the data that influences or directly relates to some desired outcome or target. Specifically, data classification deals with finding patterns and relationships in order to group like records for easier and more efficient analysis. This article focuses on a specific type of classification called decision trees.

Decision Trees

Although many data mining tasks are reserved for data analysts, business users tend to feel comfortable with decision trees. They are logical and visual, and the output can be explained in typical business terms.

Using a decision tree is a way to classify existing data, determine the factors or rules that relate to the target result, and apply them to predict an outcome, which means that

- Business users can determine the factors that most influence buying decisions
- Marketing departments can target the correct group of prospects, excluding those with very little probability of buying
- Data and financial analysts can forecast sales by analyzing attributes of current prospects
- Business analysts can adjust goals and strategies as trends change
- Companies can structure support, enhancements, and desupport to ensure maximum customer satisfaction

So, no PhD in mathematics is required for using and understanding decision trees. To illustrate this, I’ll analyze the following business problem:

A product manufacturer has two products, A and B. Overall customer feedback has been positive, but the owner wants to find out if there is anything in terms of support, enhancements, and desupport that can improve the satisfaction rate. The company maintains limited information about its customers, including only the product they are using, the product version, and the last time they received an upgrade.

By using this information, feedback solicited from a sample customer population, and Oracle Data Miner, the company can produce the decision tree model shown in Figure 1.

Each box in the tree in Figure 1 is called a node, and each line is called a branch. The top box in the tree, or the root, includes all cases in the sample.

A decision tree splits data on attributes in an attempt to identify the best predictors of a target value. These predictors form a rule or set of rules that, when applied to a node, will predict the outcome. You can think of them as IF-THEN statements for making decisions.

Oracle Data Miner makes generating decision trees easy. To illustrate this, I’ll generate the decision tree shown in Figure 1. To follow along, download and install Oracle Data Miner and perform the following steps to set up your datasource for this example:

1. Download the sample data file.
2. Unzip the contents, open a prompt, and change (cd) to the directory containing the create_user.sql script.
3. Log in to SQL*Plus as SYSDBA, and run create_user.sql.

The script creates a user, SURVEYS, with the password SURVEYS, and grants all necessary permissions. It connects to SURVEYS, creates a CUSTOMER_SATISFACTION table, and seeds it with 4,920 records. The table has the structure shown in Listing 1.

Code Listing 1: Description of CUSTOMER_SATISFACTION table

```
SQL> desc CUSTOMER_SATISFACTION
CUSTOMER_SATISFACTION_ID NOT NULL NUMBER(10)
PRODUCT VARCHAR2(100)
VERSION NUMBER(12)
LAST_UPGRADE_YEAR VARCHAR2(4)
FEEDBACK VARCHAR2(10)
```

Data is stored in this table as follows:
CUSTOMER_SATISFACTION_ID is the primary key and therefore a unique identifier.

PRODUCT can be either A or B.

VERSION is 1, 2, or 3.

LAST_UPGRADE_YEAR ranges from 1999 to 2006.

FEEDBACK is either POSITIVE or NEGATIVE.

To launch Oracle Data Miner, locate the bin directory, where you unzipped the Oracle Data Miner download, and double-click odminerw.exe. Provide connection information to the SURVEYS schema (SURVEYS/SURVEYS) when prompted, and click OK to save the connection details. Click OK again to connect and display the Oracle Data Miner interface shown in Figure 2.

Note that the CUSTOMER_SATISFACTION survey table is shown in the right pane. You can later display this survey table by clicking SURVEYS -> Tables -> CUSTOMER_SATISFACTION.

Follow these steps to create the decision tree:

1. From the Activity menu, select Build.
2. The first wizard screen is informational. Click Next.
3. In Step 1 of 5: Model Type, leave the Function Type as Classification and Algorithm as Decision Tree. Click Next. In Oracle Data Miner, classification is a function type, or category of algorithms, related to a classification, and a decision tree is an algorithm that belongs to that function type.
4. In Step 2 of 5: Data, make sure the selected schema is SURVEYS and the table/view is CUSTOMER_SATISFACTION. If they are not selected, select them from the list. The unique identifier is a single key. Make sure the primary key, CUSTOMER_SATISFACTION_ID, is selected. In the Select Columns section, select all columns. Click Next.
5. In Step 3 of 5: Data Usage, select all columns except CUSTOMER_SATISFACTION_ID as input and select FEEDBACK as the target. Click Next. The FEEDBACK column in the table lists a response as either POSITIVE or NEGATIVE. Oracle Data Miner will analyze this binary response against all attributes to determine which ones affect customer feedback the most.
6. In Step 4 of 5: Select Preferred Target Value, select Negative. Click Next. (This value can be changed at any time.) Selecting Negative instructs Oracle Data Miner to look for decisions that result in a negative response, and the decision tree will be structured with this in mind.
7. In Step 5 of 5: Activity Name, change the name to CUSTOMER_SAT_1 (increment the last number if you rerun this). Click Next.
8. In Finish leave the Run upon finish check box checked and click Finish to start the process.

At this point, Oracle Data Miner determines the best attribute splits, builds the model, and tests the rules that are generated against the data set. Figure 3 shows the activity steps.

Click the Result link in the Build activity to see the decision tree, shown in Figure 4. Note that you can control the number of node levels to show. To see the splits and branches shown in Figure 1, show only two levels.

The results in Figure 4 are the same as those shown in Figure 1. Note that the first split, or branch, is on the Last Upgrade Year attribute. If Last Upgrade Year is greater than or equal to 2003, the response is positive with 90 percent confidence. Note that the number of cases is 1,604 of the 2,951 shown in the root. This is roughly 54 percent of the total, which is referred to as support in the node. If Last Upgrade Year is less than or equal to 2002, the response is predicted to be negative with 91 percent confidence. This node has 1,347 cases, which is 46 percent of the total.

Of the four level-2 nodes, one shows that customers using versions 2 or 3 of either product A or B who upgraded between 2003 and 2006 are predicted with 92 percent confidence to have a positive opinion of the product, and another node shows that customers using product A who upgraded between 1999 and 2001 are predicted with 100 percent confidence to have a negative opinion of the product. Surely this information has value in upcoming support, enhancement, and desupport planning.

Note that there may be occasions when an attribute value is null for a record. What then? Oracle Data Miner determines not only attributes’ relationships to the target but also their relationship to each other. Notice the Split Rules section at the bottom of Figure 4. It shows a surrogate value for Node ID 1. If no Last Upgrade Year is available for a record, Oracle Data Miner still includes the record in the node if the version is 3. Oracle Data Miner determined this automatically, without requiring anything to be explicitly defined.

Applying the Build

To see how this decision tree build will be applied to individual records, open the build results again, by navigating to Mining Activities -> Classification, selecting CUSTOMER_SAT_1, clicking the Result link in the Build activity, and clicking Show Leaves Only at the top of the Result Viewer window. Each leaf has an associated split rule that appears at the bottom of the window if the node is clicked.

For example, the last leaf (Node ID = 20 in Figure 4) predicts a negative response with 100 percent confidence. Click the node, and the following split rule appears at the bottom:

\[
\text{PRODUCT is in A AND} \\
\text{LAST_UPGRADE_YEAR is in \{1999 2001 2002\}}
\]

Any record satisfying this rule will be predicted to have a negative response with 100 percent confidence.

Other customers that were not part of the survey are recorded in the CUSTOMER_SATISFACTION_NEW table. All columns are populated except for FEEDBACK (the target). To find out whether they are likely to respond positively or negatively, apply the CUSTOMER_SAT_1 build rules, as follows:

1. Open the main Oracle Data Miner window, and select Activity -> Apply from the main menu.
2. The first page is a welcome page. Click Next.
3. In Step 1 of 5: Build Activity, select the CUSTOMER_SAT_1 build activity just completed. Click Next.
4. In Step 2 of 5: Apply Data, click the Select link to the right of the table and navigate to SURVEYS -> CUSTOMER_SATISFACTION_NEW. Click Next.
5. In Step 3 of 5: Supplemental Attributes, choose only the CUSTOMER_SATISFACTION_ID column. Click Next.
6. In Step 4 of 5: Apply Option, leave the default settings. Click Next.
7. In Step 5 of 5: Activity Name, call it CUSTOMER_SAT_APPLY_1. Click Next.
8. Click Finish.

When the activity is finished, click the Result link in the activity window to see the predicted value; the probability associated with the prediction; and the cost, which is similar to the probability ranking, except that a lower cost means a better prediction.

Oracle Data Miner also makes it easy to see which rule was used for each record. Click one of the records, and then click the Rule button to the right of the results. The Rule Viewer appears, as shown in Figure 5.

Stats Disclaimer
Figure 1: Decision tree

Figure 2: Oracle Data Miner interface with CUSTOMER_SATISFACTION table selected

Figure 3: CUSTOMER_SAT_1 classification activity steps

One of the first business/statistics books I remember reading was called How to Lie With Statistics, by Darrell Huff. No, it isn’t an instruction manual on how to get away with it. On the contrary, it is a small book that highlights how common it is to derive the wrong conclusions from statistical analysis not because of faulty calculations but, rather, because of faulty assumptions, incorrect problem definition, and inaccurate or incomplete data. Finding meaning where there is none, missing meaning where there is some, and finding the wrong meaning in the wrong statistic are all discussed. The following completely fabricated example illustrates this type of analysis:

People more than 2.25 meters (almost 7.5 feet) tall are far less likely to get hurt in a car accident than those who are shorter... and news headlines across the country read, “Study confirms that really tall people are much stronger and more resilient than shorter people.”

Or could it be that people that tall are likely to be riding in a much larger vehicle? Or could it be that instead of a percentage, raw numbers are used in reporting the statistics? How many people are in this extremely tall population, versus the shorter group, to begin with?

Oracle Data Miner does a great job of making the math and programming tasks of data analysis easy, but problem definition, data selection, and correct application of the results are entirely in the hands of the user. As long as the problem is defined correctly in the first place, you will be happy with the results.

Conclusion

Decision trees are great for business users, providing logical output that can be discussed in business terms. Although they are simple to create with Oracle Data Miner, they are extremely powerful and accurate when provided with a good data set. They can offer predictions with associated probabilities and a single location for viewing the predictions, probabilities, and rules behind the predictions.

Ron Hardman works with Academy District 20 schools in Colorado Springs, Colorado, and is the founder of 5 Mile Software. He is coauthor of Oracle Database 10 g PL/SQL Programming and Expert PL/SQL, both from Oracle Press, and is an Oracle ACE.