Decision Tree Data Mining Example from Larson Text

By Susan Miertschin
Problem

The Maximum Miniatures Marketing Department wants to do a targeted mailing promoting the Mythic World line of figurines. Previous research has shown that the buyers of the Mythic World line of product do not have any children living at home. Unfortunately, the list purchased for the Mythic World mailing does not include the statistic on the number of children at home for each household. It does include other facts about the household, namely number of cars owned, marital status, and whether the address is a rental property or occupied by the owner.

The marketing department would like to find a way, using the three facts included in the mailing list data and the information known about current customer, to predict which households have no children living at home. The mailing will then be sent only to those households likely to have no children living at home.

Wholesale customers have customer numbers below 5000. Retail customers have customer numbers of 5000 and above. Of course, we should only use retail customers as our data for this mining operation.
The Problem is One of Classification

- MaxMin wants to classify the households in the mailing list as either having no children living at home or having children living at home
  - Even though this data point is not part of the mailing list data set
Recall - Data Mining - Task Types

- **Classification**
  - Fit items into slots – *Larson*
  - Assign items in a collection to target categories or classes – *Oracle book* ([http://download.oracle.com/docs/cd/B28359_01/datamine.111/b28129/classify.htm](http://download.oracle.com/docs/cd/B28359_01/datamine.111/b28129/classify.htm))
  - Applies to categorical attribute data
    - Categorical – the attribute takes on two or more discrete values

- **Clustering**

- **Discovering Association Rules**

- **Discovering Sequential Patterns – Sequence Analysis**

- **Regression**

- **Detective Deviations from Normal**
Classification

- Here is data about loan applicants. Which are good risks and which are poor risks?
- Which of our current customers are likely to increase their current rate of spending if given an affinity card?
- Which consumers are likely to buy a new cell-phone product if I send them a direct mailing?
- *Which potential customers in the new mailing list are likely to have no children at home?*
Classification

Predictive or Descriptive?
Supervised or Unsupervised?
Classification

Predictive
Supervised
Classification – Predictive - Supervised

- Predict which loan applicants are good risks and which are poor risks
- Predict which of our current customers are likely to increase their current rate of spending if given an affinity card
- Predict which consumers are likely to buy a new cell-phone product if I send them a direct mailing
- Predict which patients will respond better to treatment A or treatment B
- Predict which potential customers in the new mailing list have no children at home.
The Learn-by-Doing Section in Larson

- Creates a data mining model to determine whether the data included with the mailing list can predict whether the household has children or not, and thus, determine a more targeted subset of data for the mailing list.
Choose Classification Algorithms Available through MS Analysis Services

- MS Decision Trees
- MS Naïve Bayes
- MS Clustering
- MS Neural Network
- Other tools implement these (perhaps) and other algorithms
- Example: WEKA (Open Source) has
  - 13 (or so) different algorithms based on Bayesian probabilities
  - 16 (or so) different algorithms based on trees
  - 12 (or so) different algorithms for clustering
WEKA (Waikato Environment for Knowledge Analysis)

Classifiers Based on Bayesian Probabilities

Classifiers Based on Trees
WEKA (Waikato Environment for Knowledge Analysis)

Clusterers

Classification Using a Neural Network
What Data to Mine?

- Customer data in the OLAP Cube created earlier
- Open MaxMinSalesDM in Visual Studio
- An archive file is available from Blackboard (ITEC4397MaxMinSalesDM.zip)
- The cube defined here has been deployed to the class server
- DO NOT re-deploy, *please*
  - I am not sure that each of you will be able to access the cube simultaneously – we will have to see
The Data Mining Model is Already Built

- The model is to predict whether number of children at home is 0 or more than 0
- Number of children at home is the output variable
- What are the input variables?
- Expand Data Source Views in Solution Explorer
  - double click MaxMinSalesDM.dsv to see the structure of the cube and see what data is available
  - Is the output variable there?
  - What input variables are also in the email list?
Data Mining Model

- Customer dimension is the main source of data
- Homeowner attribute – what data is in there?
Data Mining Model

Marital Status

Num Cars Owned
Data Mining Model

- Number of children at home
- So – can the other three data points predict the value of this attribute?
- That’s the question we are to mine for
Data Mining Model

- We also want to use only data for retail customers
- Excluding wholesale customers
- Retail customer have a customer number of 5000 and above
- This query pulls the data that the mining model will examine

```
SELECT Homeowner, MaritalStatus, NumCarsOwned, NumChildrenAtHome
FROM Customer
WHERE PK_Customer>=5000
```
What Do the Models Tell Us?
Decision Tree Algorithm Results

- In Solution Explorer, expand the Mining Structures nodes and double click on the Children at Home data mining model.
- Use the Model Mining View and select the Decision Tree model to examine the results.
- Derived Rules: no cars + unmarried + not homeowner OR no cars + married + not homeowner.
- Also look at the dependency network viewer.
Decision Trees – Most Popular Data Mining Tool

- Easy to understand
- Easy to implement
- Easy to use
- Computationally thrifty
- Predicts categorical output from categorical and/or real input values
Naïve Bayes Algorithm Results

- select the Naïve Bayes model to examine the results
- Dependency network viewer is a primary tool
- Attribute Profiles, Attribute Characteristics, Attribute Discrimination lets you refine
Clustering Algorithm Results

- Cluster diagram shows data that clusters on number of children at home = 0 based on other characteristics
- Rename the two darkest clusters (No Children At Home 1 and 2, e.g.)
- Cluster Profiles
- Cluster Characteristics
- Cluster Discrimination
Decision Trees

By Susan Miertschin