B.Sc. in Computing
Data Mining

Lab sheet #4 = classification with Decision Trees

rapid-i.com

Overview

Objective:

- Using a Decision tree to classify a dataset

- Agenda:
  - Creating a decision tree model
  - Changing parameters

- Applying a model to a new dataset

Exercises

There are a number of exercises through out the slides, and repeated again on the last slide.

At the end of the lab, you must complete an MCQ test based on your answers to these exercises.

Building a Decision tree

Load the risk dataset (risk500.csv) into your RM repository (available on moodle & studentshare). Make the attribute risk the class label, and set ID to be an ID attribute.

Attributes:

ID: customer ID
AGE
INCOME
MARITAL: marital status of single; married; or widowed / separated / divorced;
NUMKIDS: number of children
NUMCARDS: number of credit cards
HOWPAID: if you are paid monthly or weekly
MORTGAGE: do you have a mortgage or not
STORECAR: Number of store cords for shops
LOANS: number of loans
RISK: class label, what type of risk you are. There are three classes: Good risk; Bad profit; Bad loss

Building a Decision tree

Complete and initial Exploratory Data Analysis (EDA) on the risk dataset.

Exercise 1:

Does the risk dataset have missing values?

Are any attributes skewed?

Are there outliers

From a scatter matrix with plot set to RISK, do you think it will be easy to distinguish between the three classes?

From a quartile Color Matrix with COLOR set to RISK, which attributes are likely to be the most useful?

Building a Decision tree

Start a new process called Lab4-risk-DT. Drag the RISK dataset onto the new process to create a retrieve operator.

Add an X-Validation block and connect it to the dataset. Output the tree and model performance. The default is to create a pruned tree.

Exercise 2:

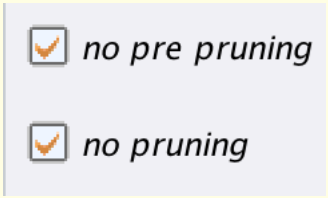
How many branches on the pruned tree?

How accurate is performance when the tree is fully pruned?

Building a Decision tree

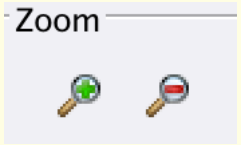
Drill down to the Decision Tree parameters.

Click the two check boxes at the end to turn off pre and post pruning



Run the process again.

This time the tree is too big to capture on one window. Click on the zoom out button to get an idea of size:



Take a look at the text view in the tree window to look at a text version of the if..then.. rules. Each leaf shows how many rows match it:

```
INCOME > 43613.500: good risk {good risk=1, bad profit=0, bad loss=0}
```

Building a Decision Tree

Exercise 3:

Typically how many rows match each branch in a fully gown tree?
Can you find any leaf node that matches more the 10 rows of data?

Has the accuracy improved compared to the pruned tree.

Building a Decision Tree

The un-pruned tree is over fitting the data, but the pruned tree is under-fitting the data.

The get the better sized tree, we are going to change some of the other parameters, which are:

Minimum size for split: Only nodes the match at least this number of rows will be considered for splitting

Minimum leaf size: only use an attribute if all branches produced match at least this many rows

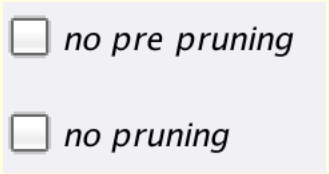
Maximum depth: the longest branch allowed

Minimal gain: The information gain must be a least this amount before a branch will be added (pre-pruning).

Confidence: the improvement in error rate before a branch is deleted (post pruning)

Building a Decision tree

Turn pruning back on:



Experiment 1. Reduce minimum gain to 0.05.

Exercise 4:

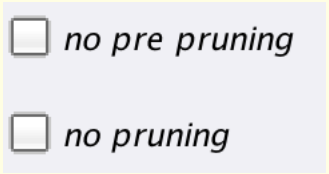
What is the effect on the tree size?

Looking at the text view for the tree, are there still branches matching less than 10 rows in the dataset?

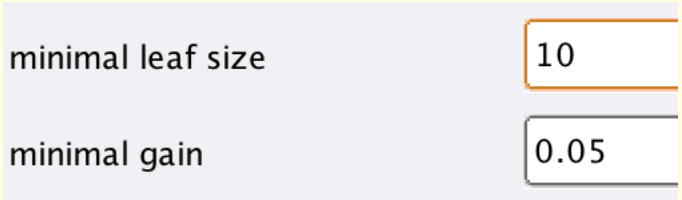
What is the effect on overall accuracy?

Building a Decision tree

Leaving pruning on:



Experiment 2. Leave minimum gain at 0.05 but increase minimum leaf size to 10.



Exercise 5:

What is the effect on the tree size?

Looking at the text view for the tree, are there any branches matching less than 10 rows in the dataset?

What is the effect on overall accuracy?

Building a Decision tree

Return to the lab3-Titanic process from last week.

Run the process to recap on the tree size and and overall performance.

Turn off all pruning and run the process again?

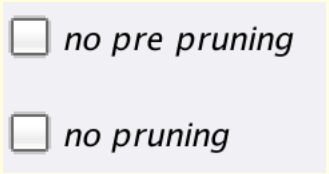
Exercise 6:

What is the main attribute used by the fully grown tree?

Does it make sense to use this attribute?

Building a Decision tree

Turn pruning back on:



And run experiment 1 and 2 as you did with the risk dataset, i.e.

Experiment 1. Reduce minimum gain to 0.05.
Experiment 2. Leave minimum gain at 0.05 but increase minimum leaf size to 10.

Exercise 7:

Does partial pruning improve the overall accuracy of the titanic dataset?
What parameter settings gave the best accuracy?