

Data Mining Techniques

Chapter 6: Decision Trees

What is a classification decision tree?	2
Visualizing decision trees.	3
Classification tree example	4
Regression trees.	5
How to grow a decision tree	6
Finding the splits	7
Growing/pruning the tree	8
Classification purity: gini coefficient	9
Classification purity: entropy reduction	10
Classification purity: chi-square	11
Regression purity: variance reduction	12
Regression purity: F-test.	13
Pruning	14
Extracting rules	15
Further refinements	16

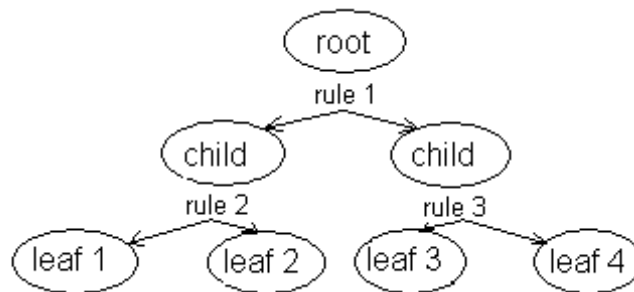
What is a classification decision tree?

- Structure used to divide a collection of records into groups using a sequence of simple decision rules:
 - e.g., classification of all living things.
- Decision rules aim to correctly classify a categorical target variable:
 - e.g., catalog company target variable might be “place an order” = 0 (no) or 1 (yes).
- Rules are based on input variables:
 - e.g., if “recency” < 6 months then predict “order” = 1, otherwise predict “order” = 0.
- Rules are *nested* (imagine branching tree-structure):
 - e.g., if “recency” < 6 months AND “frequency” > 3 purchases per year then predict “order” = 1, otherwise predict “order” = 0.

© Iain Pardoe, 2006

2 / 16

Visualizing decision trees

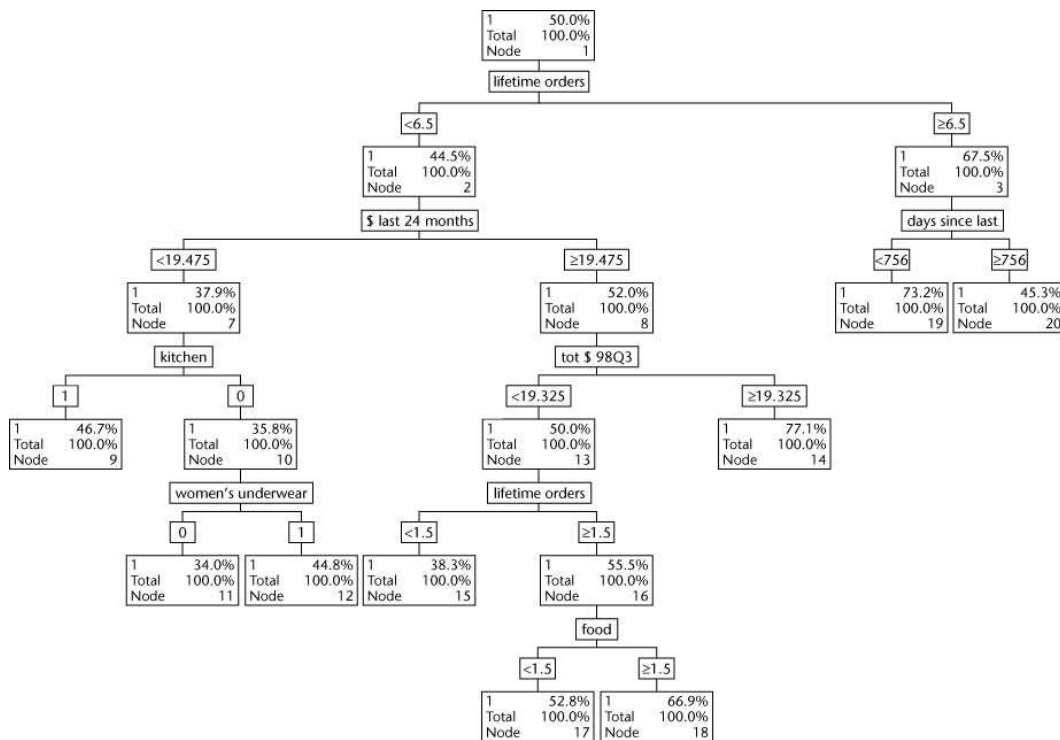


- Start at the “root node” (at the top!).
- Root branches into 2 (or more) “child nodes” (apply rule to decide which one to go to).
- Child nodes also branch based on further rules.
- Keep branching until cannot make any more splits (either too few objects at node or further splits would not improve classification accuracy).
- Proportion of 1’s (score) in terminal node (leaf) suggests likely category:
 - e.g., if proportion of 1’s > 50% then predict target = 1, otherwise predict target = 0.

© Iain Pardoe, 2006

3 / 16

Classification tree example



© Iain Pardoe, 2006

4 / 16

Regression trees

- Decision trees can also be used for prediction problems with a quantitative target variable:
 - e.g., target = amount spent.
- Such trees are called regression trees.
- Average value of the target variable in a terminal node (leaf) used as prediction for objects in that node.

© Iain Pardoe, 2006

5 / 16

How to grow a decision tree

- Classification trees:
 - choose rules to make *purity* of child nodes as high as possible (i.e., child nodes as close to 0% or 100% for the target variable categories as possible).
- Regression trees:
 - choose rules to make variance of child nodes as small as possible.

© Iain Pardoe, 2006

6 / 16

Finding the splits

- Decide which input variable makes the best split (highest purity/lowest variance in child nodes).
- Splitting criteria include:
 - classification—gini coefficient, entropy reduction (information gain), chi-square;
 - regression—variance reduction, F-test.
- Splitting on a quantitative input:
 - if $X_1 < k_1$ go left, if $X_1 \geq k_1$ go right;
 - not sensitive to outliers or skewed data.
- Splitting on a qualitative input:
 - if $X_2 \in \{A, B, C\}$ go left, if $X_2 \in \{D, E\}$ go right.
- No problem with missing data (“missing” is its own category).

© Iain Pardoe, 2006

7 / 16

Growing/pruning the tree

- Can use input variables multiple times for different nodes to fine-tune classifications/predictions.
- Tree is complete (full) when no more splits are possible:
 - each leaf is completely pure/has zero variance;
 - however, while this fits the training sample perfectly, it won't work well on the validation sample—overfitting!
- Use validation sample to *prune* tree (cut off lower down branches) using fit measures:
 - leaf error rates (proportion of misclassifications or RMSE);
 - leaf lift (proportion of leaf responders / proportion of population responders).
- Use test sample to assess fit of final selected model.

© Iain Pardoe, 2006

8 / 16

Classification purity: gini coefficient

- The gini coefficient (aka population diversity) is the sum of squares of category proportions, e.g.:
 - 50%/50%, $g = 0.5^2 + 0.5^2 = 0.5$ (lowest purity);
 - 90%/10%, $g = 0.9^2 + 0.1^2 = 0.82$;
 - 100%/0%, $g = 1^2 + 0^2 = 1.0$ (highest purity).
- Total impact of split =
proportion reaching node 1 $\times g_1$ +
proportion reaching node 2 $\times g_2$ + ...
- Select the split that results in the largest *increase* in this weighted average.
- Examples:
 - example from class 9 on credit risk;
 - homework 5 question 6 (based on book example p181–2).

© Iain Pardoe, 2006

9 / 16

Classification purity: entropy reduction

- Entropy (measures “chaos”) calculated as follows:
 - 50%/50%, $e = -1[.5 \log_2(.5) + .5 \log_2(.5)] = 1$;
 - 90%/10%, $e = -1[.9 \log_2(.9) + .1 \log_2(.1)] = .47$;
 - 100%/0%, $e = -1[1 \log_2(1) + 0 \log_2(0)] = 0$.
- Total impact of split =
proportion reaching node 1 $\times e_1$ +
proportion reaching node 2 $\times e_2$ + ...
- Select split that results in largest *reduction* in this weighted average (entropy loss \equiv information gain).
- Examples:
 - example from class 9 on credit risk;
 - homework 5 question 6 (based on book example p181–2).
- Refinement: *information gain ratio* to prevent input vars with many categories leading to “bushy trees.”

© Iain Pardoe, 2006

10 / 16

Classification purity: chi-square

- Based on chi-square test from chapter 5.

- Class 9 example on credit risk, history split:

	loan						
	observed		total	expected		difference	
	bad	good		bad	good	bad	good
good hist	2	14	16	7.5	8.5	-5.5	5.5
bad hist	13	3	16	7.5	8.5	5.5	-5.5
total	15	17	32	15	17		

- Test statistic:

$$\chi^2 = \frac{-5.5^2}{7.5} + \frac{5.5^2}{8.5} + \frac{5.5^2}{7.5} + \frac{-5.5^2}{8.5} = 15.2.$$

- Significant differences since p-value = 0.0001:

- Excel: =CHIDIST(15.2,1);
- degrees of freedom = $(r - 1)(c - 1)$.

- Credit card split has $\chi^2 = 0.536$, p-value = 0.464.
- Prefer history split as differences more significant.

© Iain Pardoe, 2006

11 / 16

Regression purity: variance reduction

- Designed for quantitative target variables, but works for qualitative target with 2 categories also.
- Class 9 example on credit risk, history split:
- Parent (root) node has 15 bad loans (target=1) and 17 good loans (target=2):
 - mean: $(15/32)1 + (17/32)2 = 1.53125$;
 - var: $[15(-0.53125)^2 + 17(0.46875)^2]/32 = 0.249$.
- History split first child node, 2 bad, 14 good:
 - mean: $(2/16)1 + (14/16)2 = 1.875$;
 - var: $[2(-0.875)^2 + 14(0.125)^2]/16 = 0.109375$.
- History split second child node, 13 bad, 3 good:
 - mean: $(13/16)1 + (3/16)2 = 1.1875$;
 - var: $[13(-0.1875)^2 + 3(0.8125)^2]/16 = 0.152344$.
- Reduction in variance = $0.249 - [0.5(0.109375) + 0.5(0.152344)] = 0.118$.
- Better than credit card split with 0.004 reduction.

© Iain Pardoe, 2006

12 / 16

Regression purity: F-test

- Designed for quantitative target variables, but works for qualitative target with 2 categories also.
- Class 9 example on credit risk, history split:
- Parent 15/17 bad/good (1/2), mean = 1.53125.
- History split:
 - first child, 2/14, mean = 1.875;
 - second child, 13/3, mean = 1.1875.
- “Between” mean square error = $[16(0.34375)^2 + 16(-0.34375)^2] / (2 - 1) = 3.78125$.
- “Within” mean square error = $[2(-0.875)^2 + 14(0.125)^2 + 13(-0.1875)^2 + 3(0.8125)^2] / (32 - 2) = 0.139583$.
- $F = 3.78125 / 0.139583 = 27.09$, p-value = 0.00001 (=FDIST(27.09, 1, 30)).
- Better than credit card split with $F = 0.133 / 0.261 = 0.511$, p-value = 0.480.

© Iain Pardoe, 2006

13 / 16

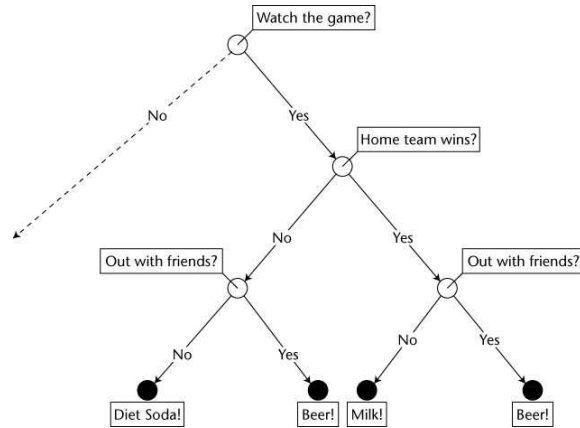
Pruning

- Remove lower branches to prevent overfitting.
- XLMiner: uses overall error rate in validation sample:
 - minimum error tree;
 - best pruned tree (smallest tree within one standard error of minimum).
- CART: adjusts error rate in training sample to penalize trees with too many branches (*cf.* adjusted R^2), then assesses results with validation sample.
- C5: calculates a confidence interval for true error rate, and uses high end of interval as an estimate.
- Stability-based pruning: uses validation sample directly to prune unstable branches.

© Iain Pardoe, 2006

14 / 16

Extracting rules



How many distinct rules are there for this tree?

© Iain Pardoe, 2006

15 / 16

Further refinements

- Taking asymmetric costs into account.
- Using more than one input variable at a time.
- Allowing linear combinations of quantitative input variables.
- Neural network trees.
- Piecewise regression using trees.
- Alternate tree representations:
 - box diagrams;
 - tree ring diagrams.
- Decision trees in practice:
 - as a data exploration tool (picking important input variables);
 - application to sequential events;
 - simulating the future.

© Iain Pardoe, 2006

16 / 16