Subject: Chi-square decision trees
Posted by James Kuyper on Fri, 19 Apr 2002 14:59:44 GMT
View Forum Message <> Reply to Message

Theres's a standard dataset characterization technique I used a couple of decades ago, and I want to use it again, and I can't remember the name of the technique.

The context is that you have a discrete dependent variable, and a large number of discrete independent variables. The basic idea was to choose a particular independent variable, and a dividing value for that variable (if there were only two different possible values, the dividing value would be between them). The program would then cumulate a 2-way table counting the number of data points with each possible value of the dependent value on one dimension, and for the second dimension dividing the data points according to whether they are above or below the dividing value for the chosen independent variable. For this table, you can compute a chi-squared statistic which indicates how statistically significant the correlation between those two variables is. For any given variable, if the variable had more than two different values, the dividing point was chosen to maximize that significance.

Each basic step of the process involved choosing the particular variable that had the most significant chi-squared value. Then, the process would repeat in a hierarchial fashion on each subset determined by that variable. Obviously, subdivision can't go any further if all of the elements in a subset had the same value for the dependent variable, or had exactly the same combination of values for all of the independent variables. Less obviously, it would halt as soon as there were too few elements in a subset to support further subdivision.

Does anyone recognise the technique I'm describing? Do you remember what the name is? Is there an IDL routine that implements it?