Tree-based methods:

**AID**  Morgan + Sonquist (1963)
Sonquist + Morgan (1964)
Monograph 35, ISR, U. of Michigan

**THAID**  Messenger and Mandell (1972)
Morgan and Messenger (1973)
THAID, SRC-ISIR, U. of Michigan

**CHAID**  Kass, G. V. (1980),
*Applied Statistics* 29, 119–127

**CART**  Brieman, et al. (1984)
*Classification and Regression Trees*, Wadsworth

---

CHAIM

- categorical response variable

Credit Rating:
“bad” “poor” “good” “very good”
- categorical explanatory variables
- create a decision tree

---

Algorithm:

Dividing the cases that reach a certain node in the tree.

(Step 1) Cross tabulate the response variable with each of the explanatory variables.

<table>
<thead>
<tr>
<th>NT=0</th>
<th>NT≥1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bad</td>
<td></td>
</tr>
<tr>
<td>Poor</td>
<td></td>
</tr>
<tr>
<td>Good</td>
<td></td>
</tr>
<tr>
<td>V. Good</td>
<td></td>
</tr>
</tbody>
</table>

When there are more than two columns, find the “best” subtable formed by combining column categories.
(Step 2) This is applied to each table with more than 2 columns.

Compute Pearson $X^2$ tests for independence for each allowable subtable.

<table>
<thead>
<tr>
<th>Fico</th>
<th>&lt; 700</th>
<th>700-750</th>
<th>700-750 &gt; 750</th>
</tr>
</thead>
<tbody>
<tr>
<td>bad</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>poor</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>good</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>v.good</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

$X^2_1$ $X^2_2$

Look for the smallest $X^2$ value. If it is not significant, combine the column categories.

<table>
<thead>
<tr>
<th>&lt; 750</th>
<th>&gt; 750</th>
<th>Repeat step 2 if the new table has more than two columns</th>
</tr>
</thead>
<tbody>
<tr>
<td>bad</td>
<td></td>
<td></td>
</tr>
<tr>
<td>poor</td>
<td></td>
<td></td>
</tr>
<tr>
<td>good</td>
<td></td>
<td></td>
</tr>
<tr>
<td>v.good</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(Step 3) Allows categories combined at step 2 to be broken apart.

For each compound category consisting of at least 3 of the original categories,

- find the “most significant” binary split
- if $X^2$ is significant, implement the split and return to step 2.
- otherwise retain the compound categories for this variable, and move on to the next variable.

(Step 4) You have now completed the “optimal” combining of categories for each explanatory variable.

Find the “most significant of these “optimally” merged explanatory variables.

<table>
<thead>
<tr>
<th>C1+C2</th>
<th>C3</th>
<th>C4+C5+C6</th>
</tr>
</thead>
<tbody>
<tr>
<td>bad</td>
<td></td>
<td></td>
</tr>
<tr>
<td>poor</td>
<td></td>
<td></td>
</tr>
<tr>
<td>good</td>
<td></td>
<td></td>
</tr>
<tr>
<td>v.good</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Compute a “Bonferroni” adjusted chi-squared test of independence for the reduced table for each explanatory variable.

(Step 5) Use the “most significant” variable in step 4 to split the node with respect to the “merged” categories for that variable.

Stop if

- no variable is significant in step 4.
- the number of cases reaching a node is below a specified limit.
Summary:

- CHAID is an algorithm
- Must categorize every variable
  - ordinal variables
  - nominal variables
- At each node it tries to find
  - best explanatory variable
  - best merger of categories

Try to make the distributions of cases across the response categories as different as possible in the "offspring" nodes.

TREEDISC macro is SAS

- modified version of CHAID
- now part of the data mining package
- application to the Wisconsin Driver data
  - response: traffic violations in 1974
    (1) at least one
    (0) none
  - explanatory variables:
    sex
    age
    history of cardiovascular disease
    place of residence
  - missing values are treated as another category

/* This program uses the TREEDISC macro in SAS to apply a modified CHAID algorithm to the Wisconsin driver data. This code is stored in the file chaidwis.sas */

/* First set some graphics options */
/* To print postscript files in UNIX */
/*
goptions cback=white ctext=black
targetdevice=ps300 rotate=landscape;
*/
/* To print postscript files from Windows */
goptions cback=white ctext=black
device=WIN target=ps
rotate=landscape;

DATA SET1;
INFILE 'c:\courses\st557\sas\drivall.dat';
INPUT AGE SEX D V R X;
LABEL AGE = AGE GROUP
      D = DRIVER GROUP
      V = VIOLATION STATUS
      X = COUNT;
run;

proc format; value sex 1 = 'Male'
           2 = 'Female';
    value age 1 = '16-36'
            2 = '36-55'
            3 = 'over 55';
    value d 1 = 'Disease'
            2 = 'Control';
    value v 1 = 'Some'
            2 = 'None';
    value r 1 = '> 150000'
            2 = '39-150000'
            3 = '10-39000'
            4 = '< 10000'
            5 = 'rural';
run;

proc print data=SET1;
run;
/* Load in the xmacros file */
%inc 'c:\courses\st557\sas\xmacro.sas';

/* Load in the TREEDISC macro */
%inc 'c:\courses\st557\sas\treedisc.sas';

/* Compute a tree for predicting violation status (V) from age, sex, disease status(D) and residence(R) */
%treedisc(data=set1, depvar=v, freq=x, 
   ordinal=age: r:, nominal=d: sex:, 
   outtree=trd, options=noformat, 
   trace=long);

/* Draw the tree on one page */
%treedisc(intree=trd, draw=graphics);

/* Draw a larger tree on several pages */
goptions cback=white ctext=black 
device=WIN target=ps rotate=portrait;
%treedisc(intree=trd, 
   draw=graphics, pos=90 120);

TREEDISC Analysis

Values of AGE : 1 2 3
Values of R : 1 2 3 4 5
Values of D : 1 2
Values of SEX : 1 2

Dependent variable (DV): V
DV values: 1 2

Splits Considered for Node 1

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Type</th>
<th>Chi-Square</th>
<th>Adjusted p</th>
</tr>
</thead>
<tbody>
<tr>
<td>AGE</td>
<td>Ordinal</td>
<td>57.39</td>
<td>0.0001</td>
</tr>
<tr>
<td>SEX</td>
<td>Nominal</td>
<td>36.80</td>
<td>0.0001</td>
</tr>
<tr>
<td>D</td>
<td>Nominal</td>
<td>4.40</td>
<td>0.0359</td>
</tr>
<tr>
<td>R</td>
<td>Ordinal</td>
<td>2.53</td>
<td>0.4458</td>
</tr>
</tbody>
</table>

Best split: AGE Ordinal with p = 0.0000

New node: 3 AGE = 2 3
DV count: 147 1864

New node: 2 AGE = 1
DV count: 133 656
Splits Considered for Node 2

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Type</th>
<th>Chi-Square Adjusted p</th>
</tr>
</thead>
<tbody>
<tr>
<td>SEX</td>
<td>Nominal</td>
<td>41.59</td>
</tr>
<tr>
<td>D</td>
<td>Nominal</td>
<td>0.01</td>
</tr>
<tr>
<td>R</td>
<td>Ordinal</td>
<td>0.15</td>
</tr>
</tbody>
</table>

Best split: SEX Nominal with p = 0.0000

New node: 5 SEX = 1
DV count: 102 302

New node: 4 SEX = 2
DV count: 31 354

Splits Considered for Node 20

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Type</th>
<th>Chi-Square Adjusted p</th>
</tr>
</thead>
<tbody>
<tr>
<td>R</td>
<td>Ordinal</td>
<td>1.41</td>
</tr>
<tr>
<td>D</td>
<td>Nominal</td>
<td>0.06</td>
</tr>
</tbody>
</table>

Best split: R Ordinal with p = 0.7031
*** Reject split

TREEDISC Analysis of Dependent Variable (DV) V

V value(s): 1 2
DV counts: 280 2520
Best p-value(s): 0.0001 0.0001

AGE value(s): 1
DV counts: 133 666
Best p-value(s): 0.0001 0.9193

SEX value(s): 2
DV counts: 31 354
Best p-value(s): 0.6064 0.8571

SEX value(s): 1
DV counts: 102 302
Best p-value(s): 0.7334 0.9703

AGE value(s): 2 3
DV counts: 147 1864
Best p-value(s): 0.0001 0.0221

SEX value(s): 2
DV counts: 20 563
Best p-value(s): 0.0856 0.5368

D value(s): 2
DV counts: 0 127
<table>
<thead>
<tr>
<th>Variable</th>
<th>Value(s)</th>
<th>DV counts</th>
<th>Best p-value(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>D</td>
<td>1</td>
<td>6 152</td>
<td>0.0592</td>
</tr>
<tr>
<td>R</td>
<td>1</td>
<td>3 22</td>
<td></td>
</tr>
<tr>
<td>R</td>
<td>2 3 4</td>
<td>1 111</td>
<td>0.5928</td>
</tr>
<tr>
<td>R</td>
<td>5</td>
<td>2 19</td>
<td></td>
</tr>
<tr>
<td>SEX</td>
<td>1</td>
<td>127 1301</td>
<td>0.0232 0.1940</td>
</tr>
<tr>
<td>AGE</td>
<td>2</td>
<td>58 462</td>
<td>0.0215 0.7310</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Variable</th>
<th>Value(s)</th>
<th>DV counts</th>
<th>Best p-value(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>D</td>
<td>2</td>
<td>18 217</td>
<td>0.1317</td>
</tr>
<tr>
<td>R</td>
<td>1</td>
<td>40 245</td>
<td>0.3814</td>
</tr>
<tr>
<td>AGE</td>
<td>3</td>
<td>69 839</td>
<td>0.0363 0.8254</td>
</tr>
<tr>
<td>R</td>
<td>1</td>
<td>20 139</td>
<td>0.8899</td>
</tr>
<tr>
<td>R</td>
<td>2 3 4 5</td>
<td>49 700</td>
<td>0.7031 0.8101</td>
</tr>
</tbody>
</table>