The CHAID Algorithm

note: As far as I'm aware this won't be in any exam but is good to help make informed decisions in relation to the course material.

CHAID: Chi-square automatic Interaction Detector

A tool used to discover the relationship between variables

For categorical data, the chi-square test is used.

Chi-Square

χ^2

A way of taking the difference between the actual and expected value and translating that into a number.

So think of χ^2 as a way of seeing whether the difference between the actual value and the expected value is significant enough (making use of the χ^2 distribution)

Look at the difference between the actual and expected and square. Divide by the expected

$\chi^2 = \sum_{i=0}^{n} \frac{Actual_i - Expected_i}{Expected_i}$

Where n is the number of catagories you have

After calculating the value look at a chi-squared distribution for the degrees of freedom

degrees of freedom (DF) = n - 1

So if you have 4 categories your degrees of freedom are 3

From the distribution you can get the probability of getting the chi-square value of the one you got. If that probability if greater than the significance value you have chosen than you fail to reject the null hypothesis. i.e the extreme case you got from your random sample is not rare enough to reject the null hypothesis.

Steps in CHAID

1. Calculate the distribution of the dependent variable
2. For each explanatory/independent variable, X, find the pair of categories of X that are the least significantly different. (largest p-value) with respect to dependent variable, Y. (if Y is categorical use the χ^2 test to calculate the p-value)
3. Now check if the p-value from the pair is greater that the significance value for merging, if it is then merge the pair to create a composite variable and repeat the process.
    - Stop the merging when left with only two categories or the significance value for merging can't be met.

4. Now calculate the adjusted p-value for the new categories and the dependent variables categories using the Bonferroni adjustment.
5. repeat for the next categorical independent variable
6. Now split the root node by the independent variable which has the lowest adjusted p-value
7. Grow tree until stopping criteria reached