C4.5 ALGORITHM (1) HOW TO SPLIT DATA AT A NODE

SYRACUSE UNIVERSITY
School of Information Studies

HOW TO FIND THE BEST DECISION TREE

Too many candidate trees

Manual construction takes too long

Need some machine intelligence to help

DECISION TREE INDUCTION

Many algorithms:

- Hunt's algorithm (one of the earliest)
- CART
- ID3, C4.5
- SLIQ, SPRINT

C4.5 is introduced in this class.

TREE INDUCTION

Key questions to build a decision tree model:

Which attribute to pick as internal node?

How to split the data set at a node?

HOW TO SPLIT DATA AT A NODE

How many branches?

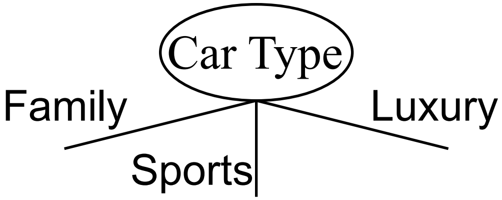
- Splitting can be:
  - Two-way split
  - Multiway split

What are the splitting values?

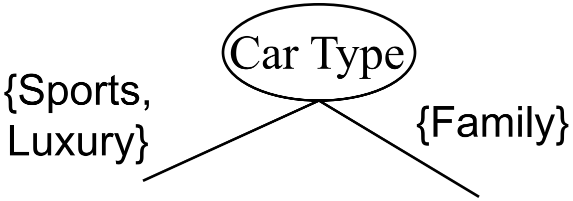
- Splitting conditions depend on attribute type:
  - Nominal or categorical
  - Ordinal
  - Continuous

SPLITTING BASED ON CATEGORICAL ATTRIBUTES

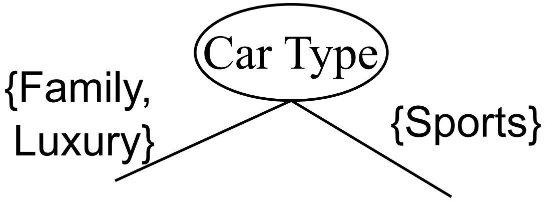
Multiway split: Use as many partitions as distinct values.



Binary split: Divides values into two subsets. Need to find optimal partitioning.



OR



SPLITTING BASED ON CONTINUOUS ATTRIBUTES

Different ways of handling

- Discretization to form an ordinal categorical attribute
  - E.g., age: 1 1 6 7 8 9 9 9 10 10 11 11 12 13 14 15 17 18
- Equal interval: One bin for every six years [0-6][7-12][13-18]
  - 1 1 6 • 7 8 9 9 9 10 10 11 11 12 • 13 14 15 17 18
- Equal frequency: One bin for every six numbers (could have ties)
  - 1 1 6 7 8 • 9 9 9 10 10 11 11 • 12 13 14 15 17 18
- Customized discretization

SPLITTING BASED ON CONTINUOUS ATTRIBUTES

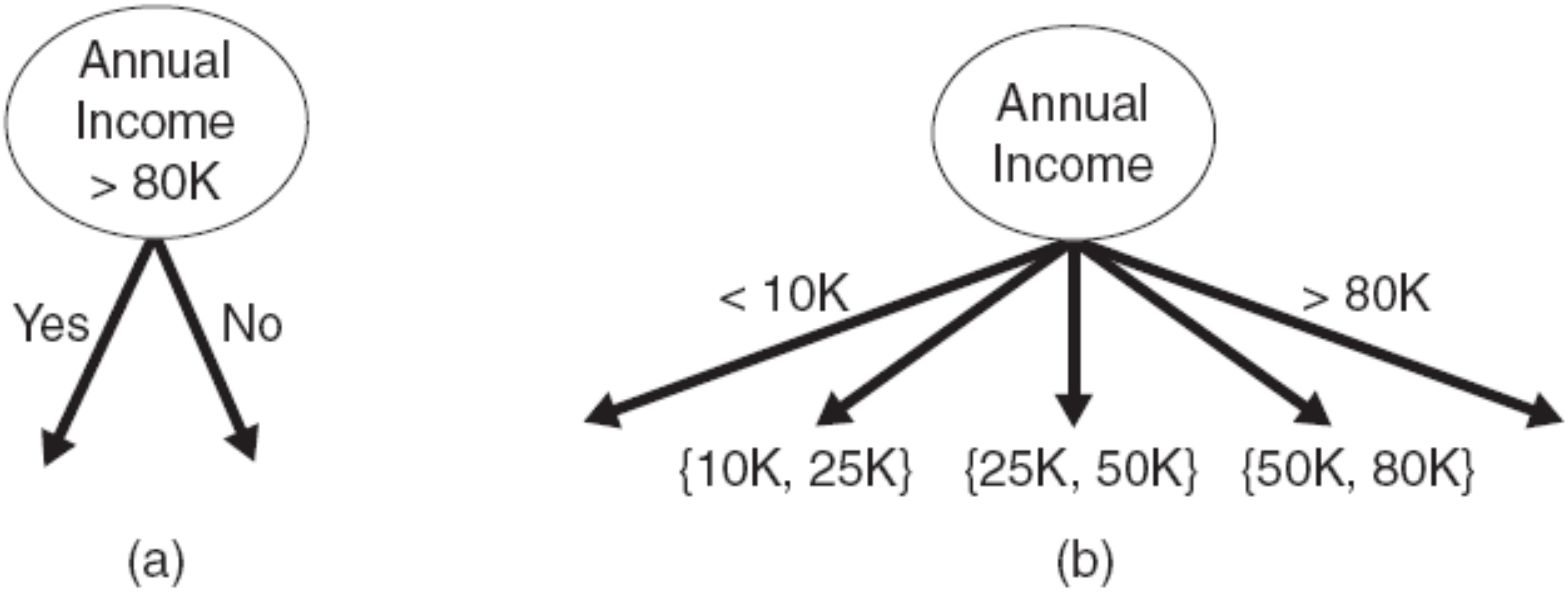


Figure 4.11. Test condition for continuous attributes.