DMML, 14 Jan 2020

Items $I = \{i_1, \ldots, i_N\}$

Transactions $T = \{t_1, \ldots, t_m\}$

→ A-priori algorithm

Frequent itemsets

↓

Association Rules

$X \rightarrow Y$

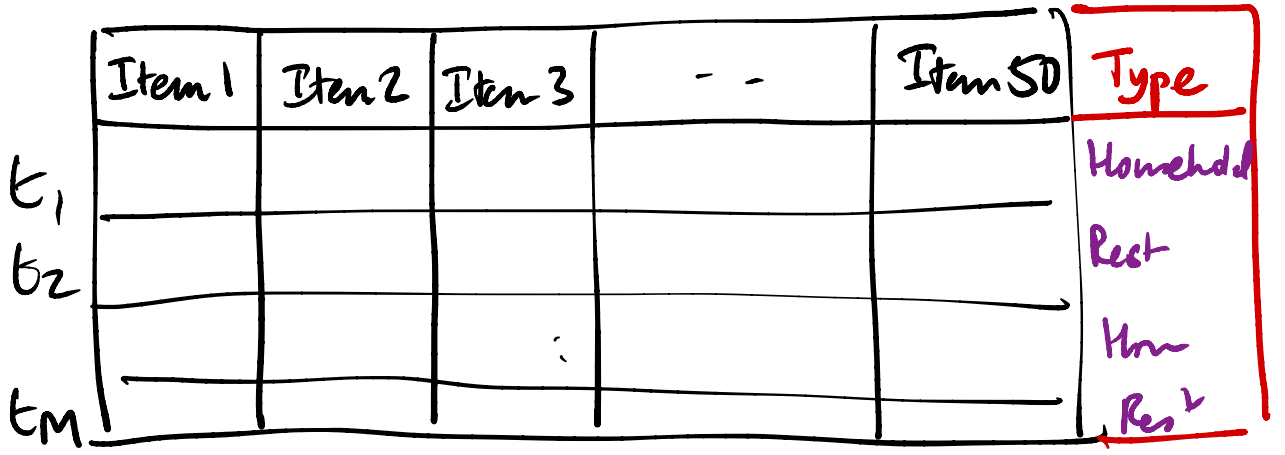
1. Variable thresholds

   Frequency of interest may depend on items

2. Sequential information

   Buy X now, later buy Y

Think of transactions as a table



Suppose one column has a different status?

Specialized rules : Set of Items → Type

Classification of transactions

Class Association Rules

Another example

Transactions are document - short news items

Type = topic

Document - set of words

{ball, pitch, catch} → Sport

{award, box office} → Entertainment

Supervised learning

Classification → Predict a number

↳ Predict a category

Marks in mock exams → Board exam marks

Symptoms + Test results → Disease?

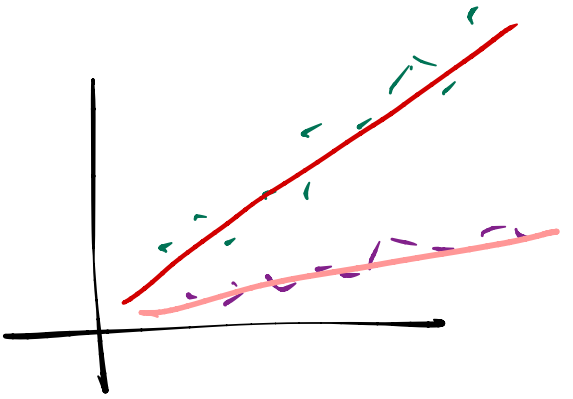
Historical labelled data → Build a model to predict labels

Input - labelled data — Simplify, assume Yes/No

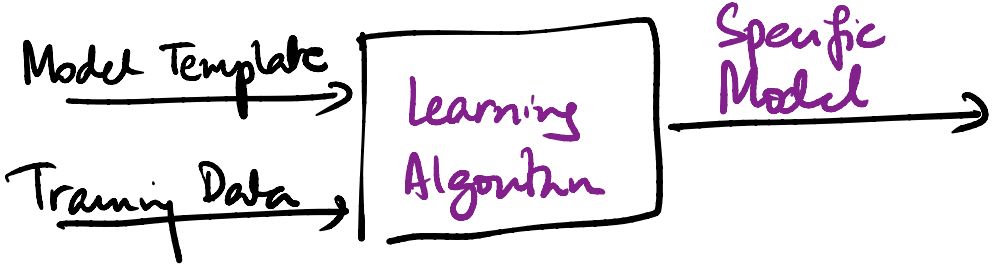
Individual data items have attributes

$\{a_1, a_2, \ldots, a_k\}, c$

Training Data



$mx+c$



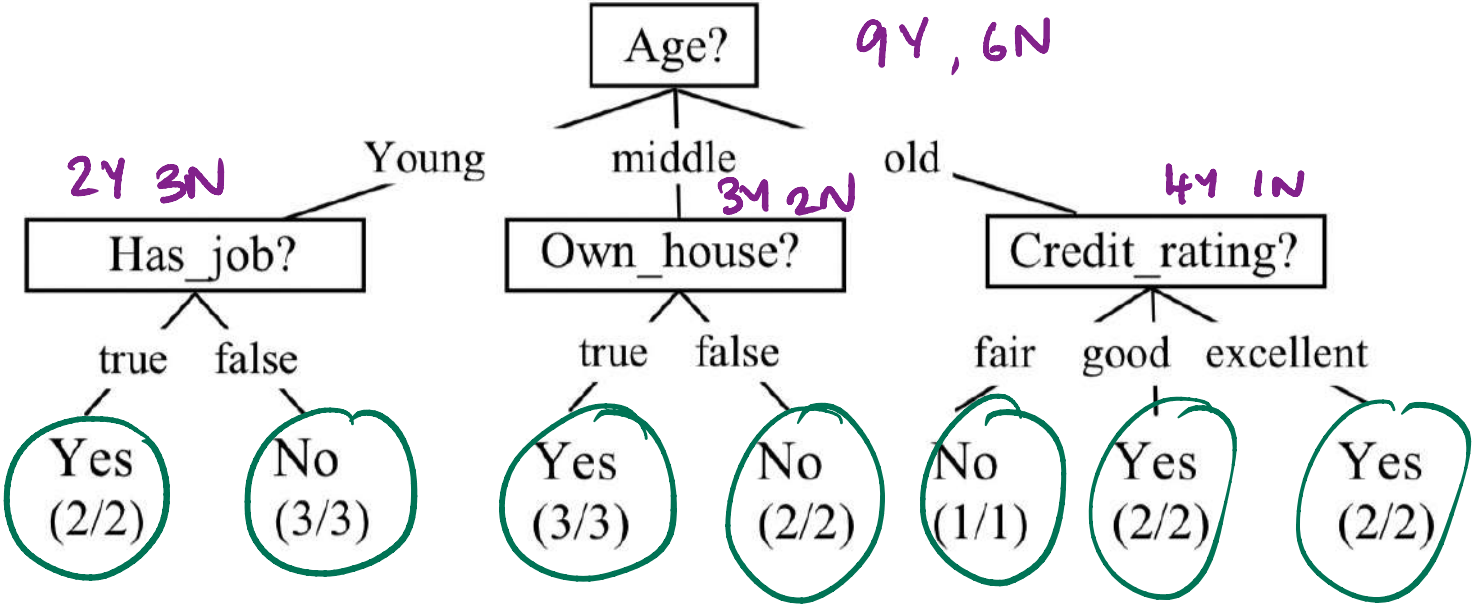
Implicit assumption

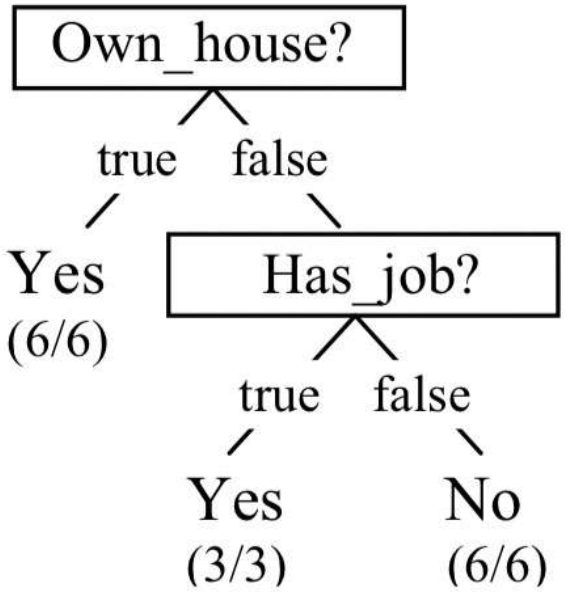
Training data has same "distribution" as the unseen data

ID	Age	Has_job	Own_house	Credit_rating	Class
1	young	false	false	fair	**No**
2	young	false	false	good	**No**
3	young	true	false	good	**Yes**
4	young	true	true	fair	**Yes**
5	young	false	false	fair	**No**
6	middle	false	false	fair	**No**
7	middle	false	false	good	**No**
8	middle	true	true	good	**Yes**
9	middle	false	true	excellent	**Yes**
10	middle	false	true	excellent	**Yes**
11	old	false	true	excellent	**Yes**
12	old	false	true	good	**Yes**
13	old	true	false	good	**Yes**
14	old	true	false	excellent	**Yes**
15	old	false	false	fair	**No**

Decision Tree

20 questions

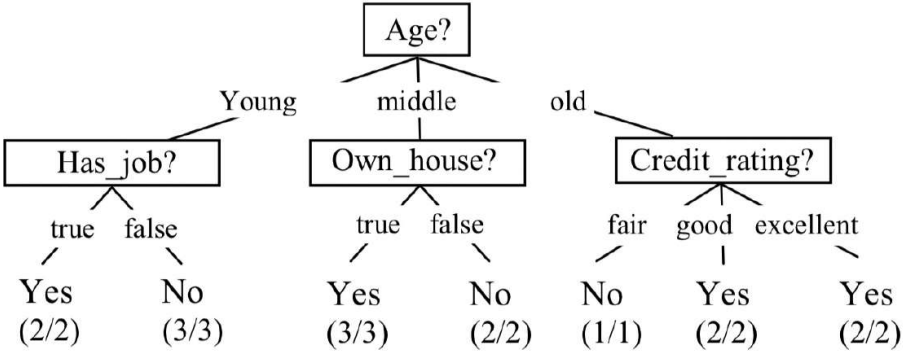


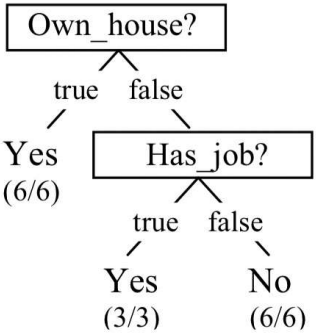


Another tree for the same training data

Starting question determines the tree

At each level, what to ask?





Which is better?

Smaller is better

Occam's Razor

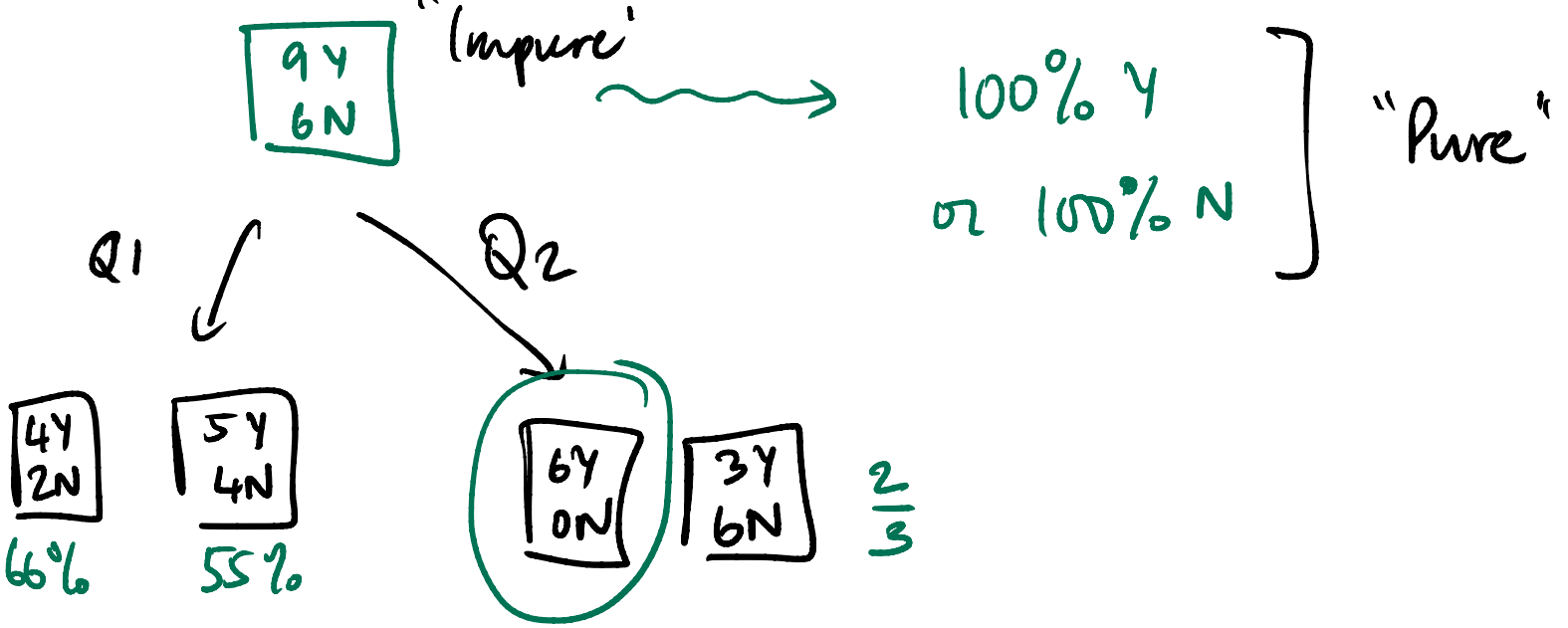
William of Ockham

Computationally - no efficient way to find smallest tree

NP-complete

Apply some approximate solution

Greedy – next question is the one that "improves" our prediction best



Define a measure of purity

Ask questions to reduce impurity (or increase purity) most at one step