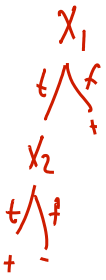
Real-Valued inputs

- What should we do if some of the inputs are real-valued?

mpg	cylinders	displacemen	horsepower	weight	acceleration	modelyear	maker
good	4	97	75	2265	18.2	77	asia
bad	6	199	90	2648	15	70	america
bad	4	121	110	2600	12.8	77	europe
bad	8	350	175	4100	13	73	america
bad	6	198	95	3102	16.5	74	america
bad	4	108	94	2379	16.5	73	asia
bad	4	113	95	2228	14	71	asia
bad	8	302	139	3570	12.8	78	america
:	:	:	:	:	:	:	:
:	:	:	:	:	:	:	:
:	:	:	:	:	:	:	:
good	4	120	79	2625	18.6	82	america
bad	8	455	225	4425	10	70	america
good	4	107	86	2464	15.5	76	europe
bad	5	131	103	2830	15.9	78	europe

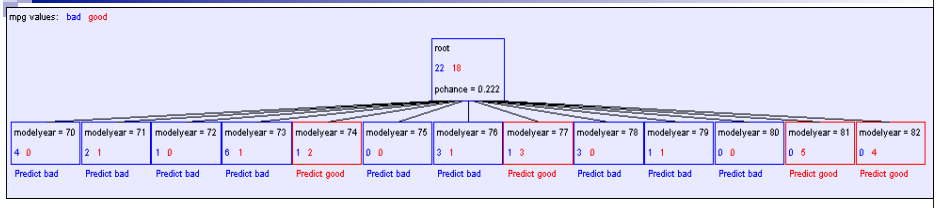


Infinite number of possible split values!!!

Finite dataset, only finite number of relevant splits!

Idea One: Branch on each possible real value

"One branch for each numeric value" idea:

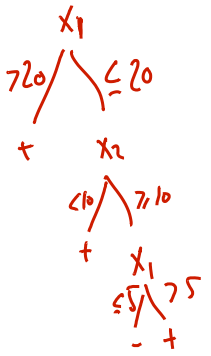


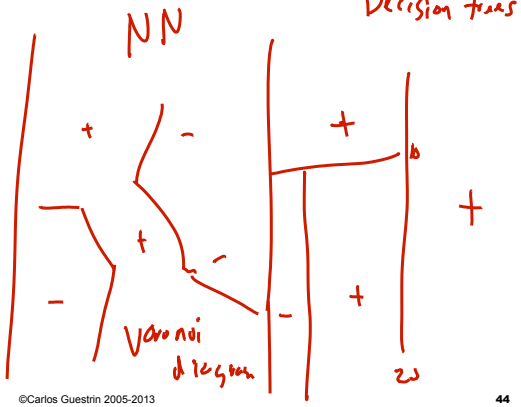
Hopeless: with such high branching factor will shatter the dataset and overfit like crazy

Threshold splits

- Binary tree, split on attribute X_i
  - One branch: $X_i < t$
  - Other branch: $X_i \geq t$

Split on same feature again? in discrete case: NO continuous, yes!





Choosing threshold split

- Binary tree, split on attribute X_i
  - One branch: $X_i < t$ ← pick threshold
  - Other branch: $X_i \geq t$
- Search through possible values of *t*
  - Seems hard!!!
- But only finite number of *t*'s are important
  - Sort data according to X_i into $\{x_1,\ldots,x_m\}$
  - Consider split points of the form $x_a + (x_{a+1} - x_a)/2$

20 ; 21 ; 25 ; 30..

A better idea: thresholded splits

- Suppose X_i is real valued
- Define $IG(Y|X_i{:}t)$ as $H(Y) - H(Y|X_i{:}t)$
- Define $H(Y|X_i{:}t)$ =

$$H(Y|X_i < t)\, P(X_i < t) + H(Y|X_i >= t)\, P(X_i >= t)$$

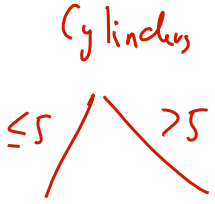
  - $IG(Y|X_i{:}t)$ is the information gain for predicting Y if all you know is whether X_i is greater than or less than t
- Then define $IG^*(Y|X_i) = max_t\, IG(Y|X_i{:}t)$
- For each real-valued attribute, use $IG^*(Y|X_i)$ for assessing its suitability as a split
- Note, may split on an attribute multiple times, with different thresholds

Example with MPG

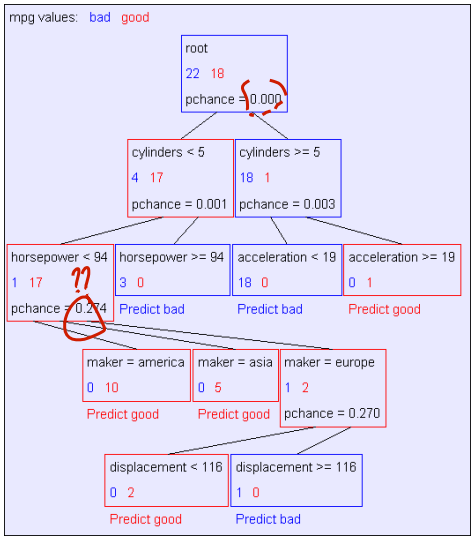
Information gains using the training set (40 records)

mpg values: bad good

Input	Value	Distribution	Info Gain
cylinders	< 5		0.48268
	>= 5		
displacement	< 198		0.428205
	>= 198		
horsepower	< 94		0.48268
	>= 94		
weight	< 2789		0.379471
	>= 2789		
acceleration	< 18.2		0.159982
	>= 18.2		
modelyear	< 81		0.319193
	>= 81		
maker	america		0.0437265
	asia		
	europe		



Example tree using reals



What you need to know about decision trees

- Decision trees are one of the most popular data mining tools
  - Easy to understand
  - Easy to implement
  - Easy to use
  - Computationally cheap (to solve heuristically)
- Information gain to select attributes (ID3, C4.5,…)
- Presented for classification, can be used for regression and density estimation too
- Decision trees will overfit!!!
  - Zero bias classifier ! Lots of variance
  - Must use tricks to find “simple trees”, e.g.,
    - Fixed depth/Early stopping
    - Pruning
    - Hypothesis testing