Decision Trees

Data Mining & Analytics

Prof. Zach Pardos

INFO254/154: Spring ’19

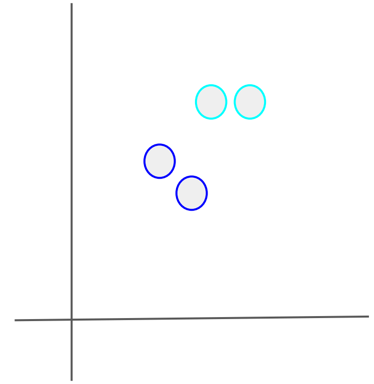
Decision Trees: Terminology

Impurity, uncertainty, entropy, information

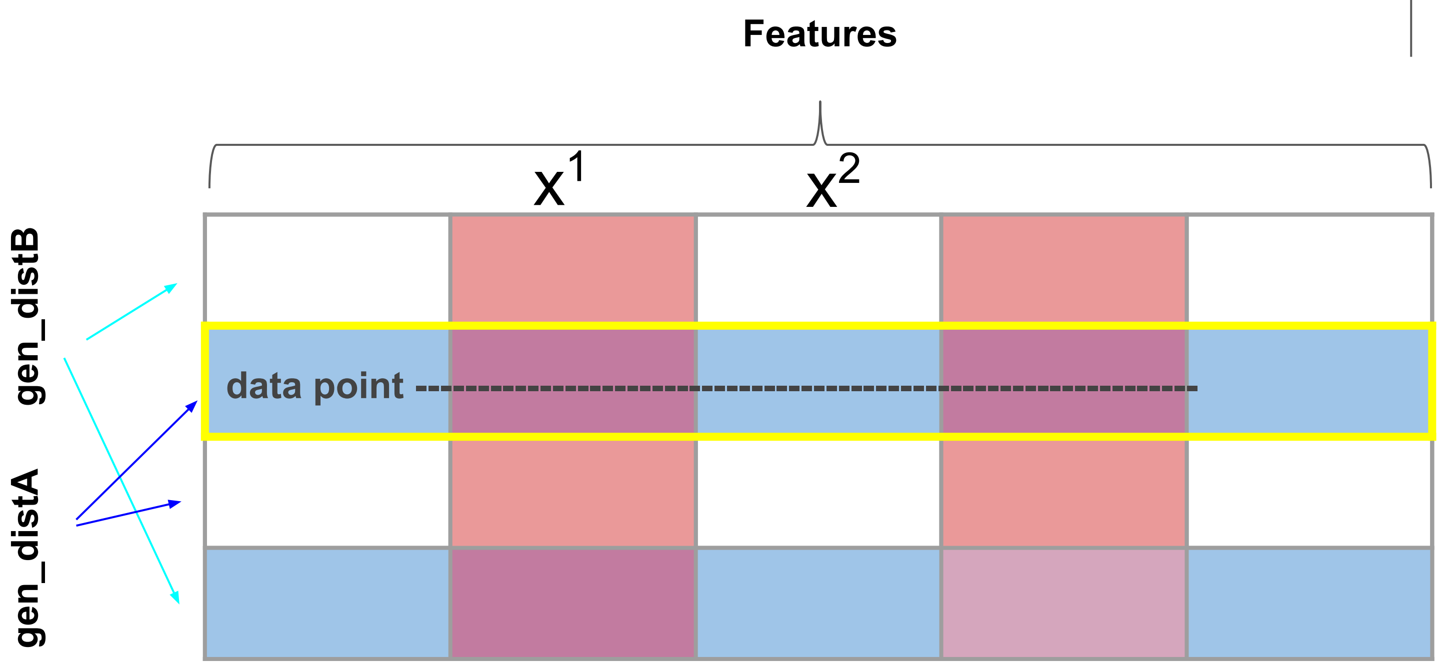
tuples, instances, rows, class, label, target

Review

Clustering: Theory



Instance, row, data point, object, cluster, group, partition

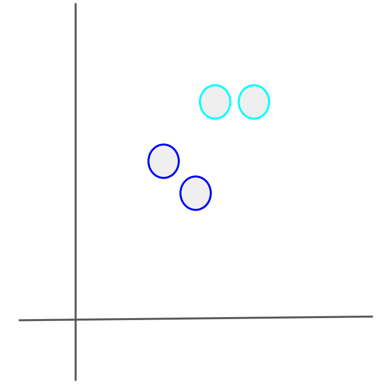


What is the hypothesis behind clustering?

That there is a set (K) of generating distributions from which the data were created

Review

Classification



Instance, row, data point, object, cluster, group, partition

Target **Features**

		x^1	x^2		
B					
A	**data point**				
A					
B					

Classification: $X_m \rightarrow Y^n$ (the target)

Decision Trees: Exercise

Impurity, uncertainty, entropy, information

tuples, instances, rows, class, label, target

Classify: Glasses? using this dataset made last week

Use only 1 rule and one attribute
(e.g., if value X is > N then Glasses = True, else False)

Given the following candidate attributes:

- **Height**
- **Hair color**

How good was your classification?

Decision Trees: Exercise

Impurity, uncertainty, entropy, information

tuples, instances, rows, class, label, target

Classify: Glasses? using this dataset made last week

Use only 2 rules and 1 or both attributes

Given the following candidate attributes:

- **Height**
- **Hair color**

How good was your classification?

Decision Trees: Exercise

Impurity, uncertainty, entropy, information

tuples, instances, rows, class, label, target

How did you treat categoricals?

What was your goodness metric?

Given enough rules, could you always get 100% accuracy?

If there is no real relationship between the features and the label, how would you expect your rules to perform on newly collected data?

Decision Trees: Exercise

Impurity, uncertainty, entropy, information

tuples, instances, rows, class, label, target

How many rules are too many? (Overfit?)

Decision Trees: Terminology

Impurity, uncertainty, entropy, information

tuples, instances, rows, class, label, target

Major types of machine learning:

- **Unsupervised Learning:** grouping instances by a notion of similarity

- **Supervised Learning:** grouping instances by a notion of purity (wrt the label)
  - "Regression" (predicting a continuous value)
  - Classification (predicting a categorical)
    - Decision Trees (C4.5 / CART) - has been generalized to regression

Decision Trees: Example dataset

Table 8.1 Class-Labeled Training Tuples from the *AllElectronics* Customer Database

RID	*age*	*income*	*student*	*credit_rating*	*Class: buys_computer*
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

Han, Camber, Pei (2011), Sec 8.2

Decision Trees: Measures of Impurity

Information
(Shannon's entropy)

$$Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i)$$

- Used in ID3 / C4.5 (Quinlan, 1979-1986)

Gini Index
(impurity)

$$Gini(D) = 1 - \sum_{i=1}^{m} p_i^2$$

- Used in CART (Breiman ~1984)
- Requires binary decision splits

p_i is the percentage of instances in *D* with *i* as the label

- Both are "greedy" algorithms
- Both produce "inspectable" models
- Neither have any distributional assumptions (non-parametric)

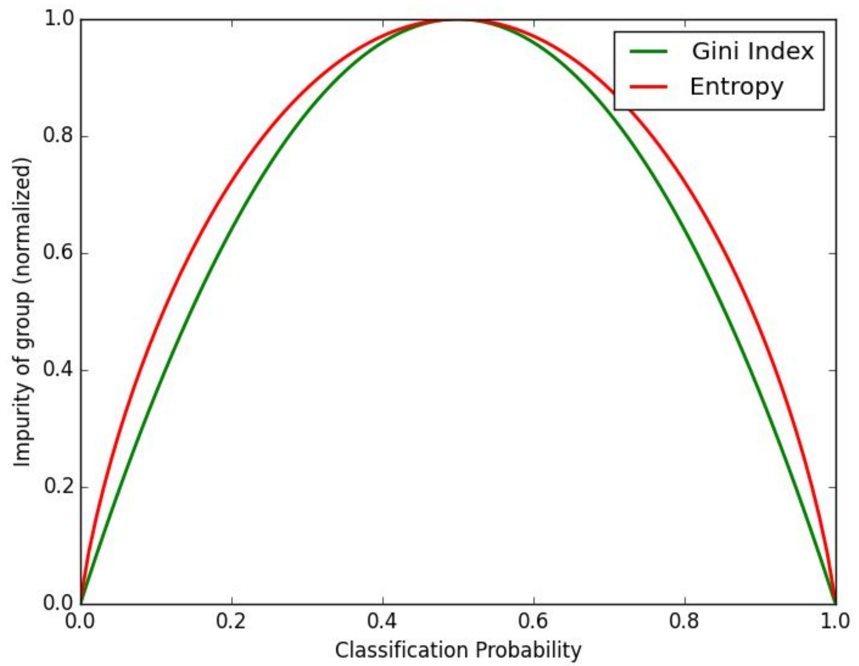
Decision Trees: Measures of Impurity

Information (Shannon's entropy)

Gini Index (impurity)

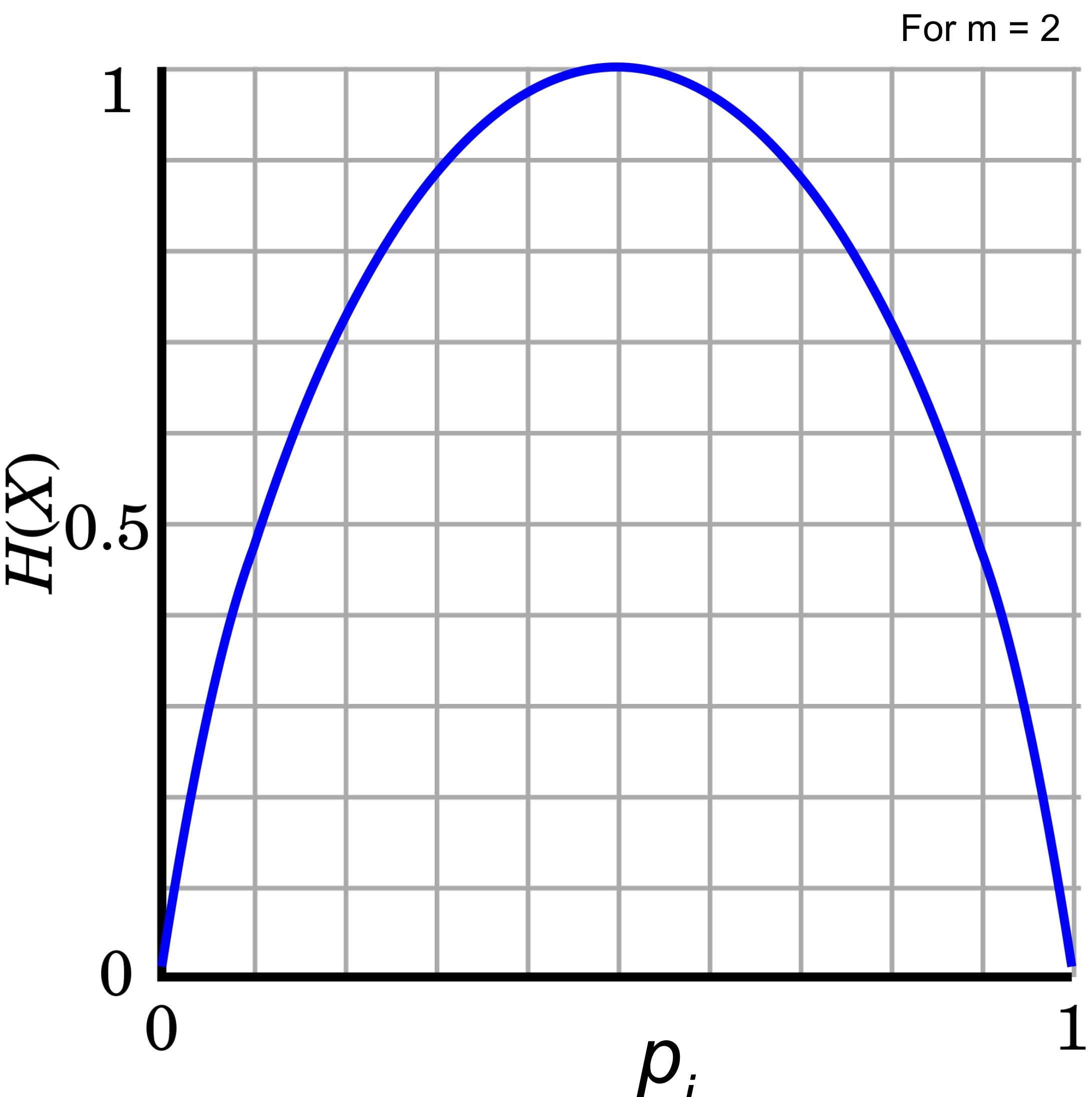
Note that this is *normalizing* Gini index to a range of 0 and 1.

With a binary label, Gini index ranges from 0 to 0.50.



Information Gain

Entropy curve (A measure of impurity and uncertainty)



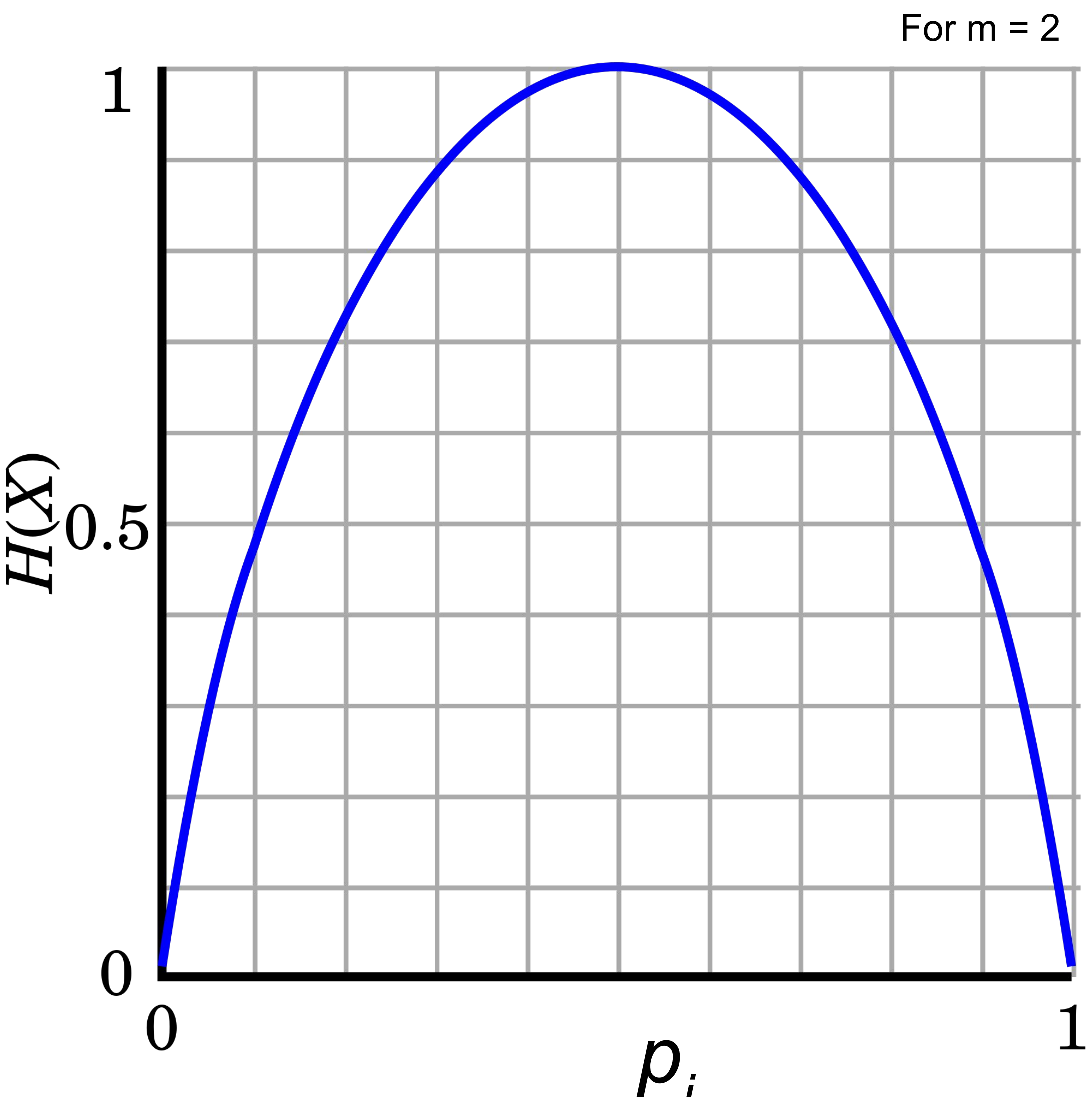
$$Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i)$$

What is the value of *Info(D)* when we have a completely pure set of tuples in D?

What is the value of *Info(D)* when our tuples' binary labels are split 50/50?

Information Gain

Entropy curve (A measure of impurity and uncertainty)



Entropy of a candidate attribute:

$$Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times Info(D_j)$$

v refers to the number of unique values (or split points) in attribute *A*.

D_j contains only the instances in *D* for which attribute *A* has the value *j*

Information Gain (entropy reducing) of candidate attribute to split on is:

$$Gain(A) = Info(D) - Info_A(D).$$

Decision Trees: Attribute selection methods

Gini Index: a measure of impurity

- Alternative measure of impurity of class labels among a particular set of tuples (used by CART alg)

For parent node:
$$Gini(D) = 1 - \sum_{i=1}^{m} p_i^2$$

For an attribute/split:
$$Gini_A(D) = \frac{|D_1|}{|D|} Gini(D_1) + \frac{|D_2|}{|D|} Gini(D_2)$$

Change in impurity (maximize this when choosing attribute/split):

$$\Delta Gini(A) = Gini(D) - Gini_A(D).$$

p_i = probability that a tuple in D belongs to class C_i
m = number of classes

Decision Trees: Example dataset

Table 8.1 Class-Labeled Training Tuples from the *AllElectronics* Customer Database

RID	*age*	*income*	*student*	*credit_rating*	*Class: buys_computer*
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

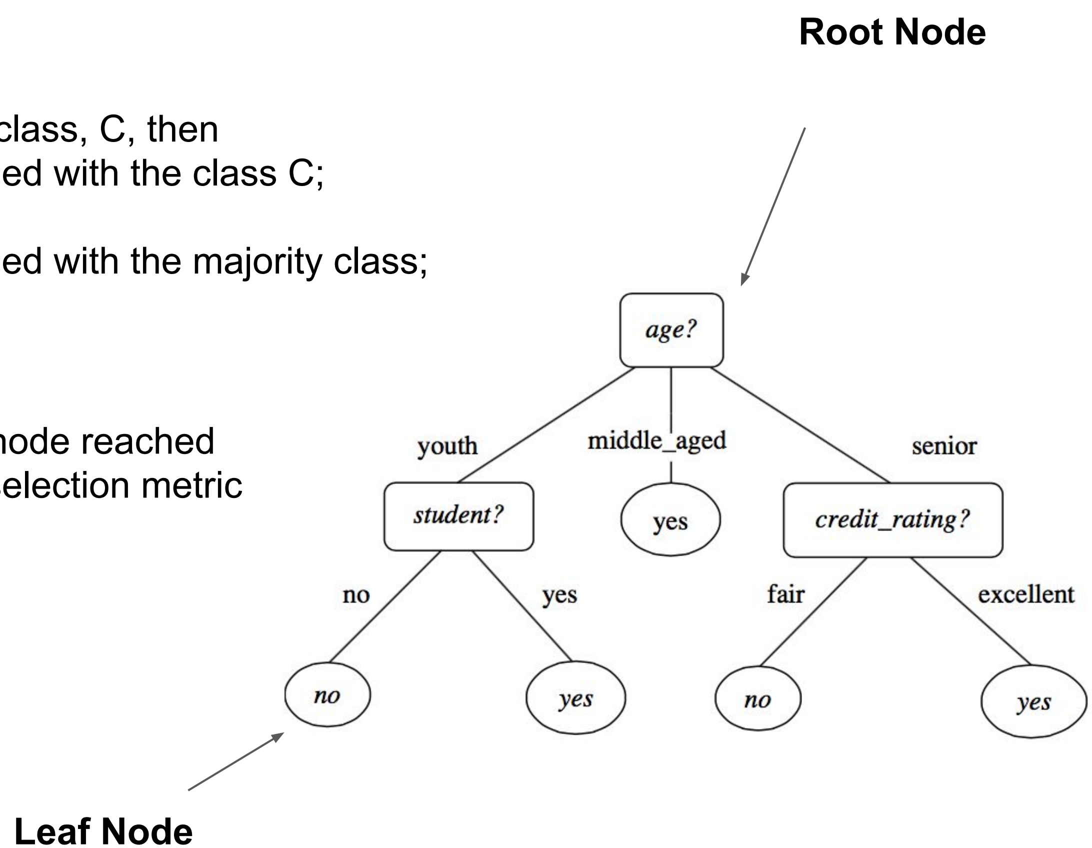
Han, Camber, Pei (2011), Sec 8.2

Decision Trees: Algorithm

(1) Create a node N;
(2) if tuples in D are all of the same class, C, then
(3)   return N as a leaf node labeled with the class C;
(4) if attribute_list is empty then
(5)   return N as a leaf node labeled with the majority class;

Other Stopping Criteria

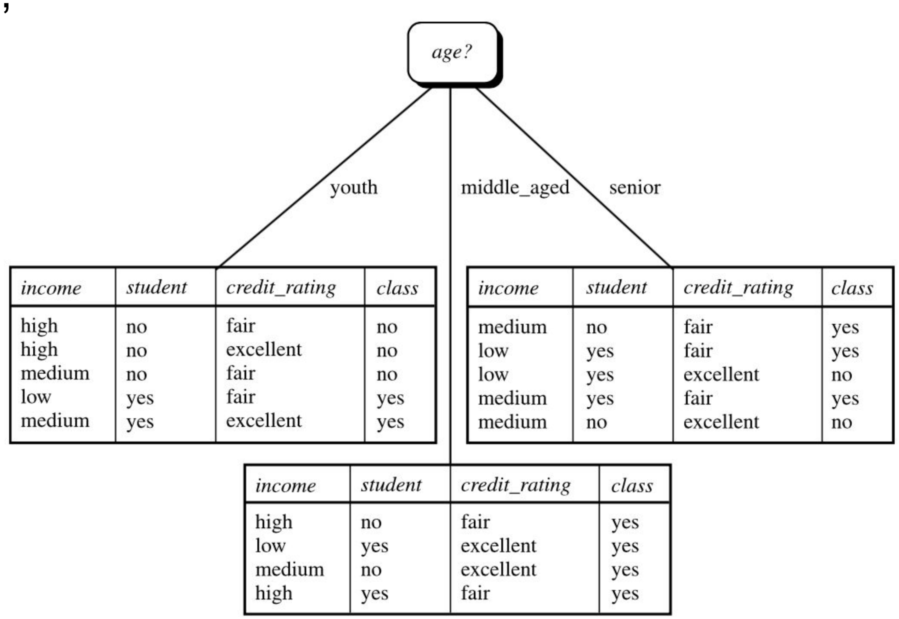
- Max depth reached
- Minimum number of tuples in a leaf node reached
- Some minimum change in attribute selection metric reached



Han, Camber, Pei (2011), Sec 8.2

Decision Trees: Algorithm (training)

(6) apply Attribute selection method to find best splitting criterion (Gini or Info)
(7) label node with splitting criterion
(8) if splitting attribute is nominal/categorical and multiway splits are allowed then
(9) remove attribute from attribute list
(10) for each outcome of j of splitting criterion
(11) let D_j be the set of data tuples in D satisfying outcome j;
(12) if D_j is empty then
(13) attach a leaf labeled with the majority class in D to node N;
(14) else attach the node returned by generate_decision_tree to node N;
(15) return N;



youth:

income	*student*	*credit_rating*	*class*
high	no	fair	no
high	no	excellent	no
medium	no	fair	no
low	yes	fair	yes
medium	yes	excellent	yes

middle_aged:

income	*student*	*credit_rating*	*class*
high	no	fair	yes
low	yes	excellent	yes
medium	no	excellent	yes
high	yes	fair	yes

senior:

income	*student*	*credit_rating*	*class*
medium	no	fair	yes
low	yes	fair	yes
low	yes	excellent	no
medium	yes	fair	yes
medium	no	excellent	no

Han, Camber, Pei (2011), Sec 8.2

Decision Trees: Other parameters

- Max Depth
- Minimum Leaf Size
- Post-pruning (cost complexity ratio)
  - Minimizes increase in error from pruning while also minimizing the number of rules used.

$$\frac{err(prune(T,t),S) - err(T,S)}{|leaves(T)| - |leaves(prune(T,t))|}$$

- err(Tree, Data) returns the accuracy of classification of Data using Tree
- prune(T,t) returns a tree which is the rules in tree T minus the pruned rules in t
- |leaves(Tree)| returns the number of leaves in Tree

Pruning is called “pessimistic” when using the training set (S = D) to conduct the pruning. It is simply called “pruning” otherwise, when held out data is used (S - D’) to conduct the pruning.

Data Mining & Analytics

[addendum to late turn-in forgiveness policy]

Prof. Zach Pardos

INFO254/154: Spring ’19