Decision Trees & Overfitting
Summary of Decision Trees (so far)

- Decision tree induction \(\rightarrow\) choose the best attribute
  - Choose split via information gain
  - Build tree greedily, recursing on children of split
  - Stop when we achieve homogeny
    - i.e., when all instance in a child have the same class
Summary of Decision Trees (so far)

Information Gain: Mutual information of attribute $A$ and the class variable of data set $X$

$$\text{InfoGain}(X, A) = H(X) - H(X \mid A)$$

$$= H(X) - \sum_{v \in \text{values}(A)} \frac{|\{x \in X \mid x_A = v\}|}{|X|} \times H(\{x \in X \mid x_A = v\})$$

Entropy:

$$H(X) = -\sum_{c \in \text{Classes}} \frac{|\{x \in X \mid \text{class}(x) = c\}|}{|X|} \log_2 \frac{|\{x \in X \mid \text{class}(x) = c\}|}{|X|}$$
## Restaurant Example

Random: Patrons or Wait-time; Least-values: Patrons; Most-values: Type; Max-gain: ???

<table>
<thead>
<tr>
<th>Type</th>
<th>Patrons variable</th>
<th>Wait-time variable</th>
</tr>
</thead>
<tbody>
<tr>
<td>French</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>Italian</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>Thai</td>
<td>N</td>
<td>Y</td>
</tr>
<tr>
<td>Burger</td>
<td>N</td>
<td>Y</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Patrons variable</th>
<th>Empty</th>
<th>Some</th>
<th>Full</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td></td>
<td></td>
<td>NY Y</td>
</tr>
</tbody>
</table>
Computing information gain

\[ I(X) = ? \]

\[ I(\text{Pat, } X) = ? \]

\[ I(\text{Type, } X) = ? \]

Gain (Pat, X) = ?
Gain (Type, X) = ?

Based on Slide from M. desJardins & T. Finin
### Computing information gain

#### I(X) =

\[-(0.5 \log 0.5 + 0.5 \log 0.5)\]

\[= 0.5 + 0.5 = 1\]

#### I (Pat, X) = ?

<table>
<thead>
<tr>
<th></th>
<th>French</th>
<th></th>
<th>Italian</th>
<th></th>
<th>Thai</th>
<th></th>
<th>Burger</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Y</td>
<td></td>
<td>Y</td>
<td></td>
<td>Y</td>
<td></td>
<td>N</td>
<td></td>
</tr>
<tr>
<td></td>
<td>N</td>
<td></td>
<td>N</td>
<td></td>
<td>N Y</td>
<td></td>
<td>N Y</td>
<td></td>
</tr>
<tr>
<td></td>
<td>N</td>
<td></td>
<td>Y</td>
<td></td>
<td>Y</td>
<td></td>
<td>N Y</td>
<td></td>
</tr>
<tr>
<td></td>
<td>N</td>
<td></td>
<td>Y</td>
<td></td>
<td>Y</td>
<td></td>
<td>N Y</td>
<td></td>
</tr>
</tbody>
</table>

#### I (Type, X) = ?

Gain (Pat, X) = ?

Gain (Type, X) = ?

---

Based on Slide from M. desJardins & T. Finin
Computing information gain

I(X) = 
- (.5 log .5 + .5 log .5)  
= .5 + .5 = 1

I (Pat, X) = 
2/12 (0) + 4/12 (0) +  
6/12 (- (4/6 log 4/6 + 2/6 log 2/6))  
= 1/2 (2/3*.6 + 1/3*1.6)  
= .47

Gain (Pat, X) = ?  
Gain (Type, X) = ?

Based on Slide from M. desJardins & T. Finin
Computing information gain

$I(X) =\
- (.5 \log .5 + .5 \log .5)\
= .5 + .5 = 1$

$I (\text{Pat}, X) =\
\frac{2}{12} (0) + \frac{4}{12} (0) +\
\frac{6}{12} (- (\frac{4}{6} \log \frac{4}{6} +\
\frac{2}{6} \log \frac{2}{6})))\
= \frac{1}{2} (\frac{2}{3}* .6 +\
\frac{1}{3}*1.6)\
= .47$

$I (\text{Type}, X) =\
\frac{2}{12} (1) + \frac{2}{12} (1) +\
\frac{4}{12} (1) + \frac{4}{12} (1) = 1$

Gain (Pat, X) = ?
Gain (Type, X) = ?

Based on Slide from M. desJardins & T. Finin
**Computing information gain**

\[ I(X) = - (.5 \log .5 + .5 \log .5) = .5 + .5 = 1 \]

\[ I(Pat, X) = \frac{2}{12} (0) + \frac{4}{12} (0) + \frac{6}{12} (- (\frac{4}{6} \log 4/6 + \frac{2}{6} \log 2/6)) = \frac{1}{2} (2/3 \times 0.6 + 1/3 \times 1.6) = 0.47 \]

\[ I(\text{Type}, X) = \frac{2}{12} (1) + \frac{2}{12} (1) + \frac{4}{12} (1) + \frac{4}{12} (1) = 1 \]

\[ \text{Gain (Pat, X)} = 1 - 0.47 = 0.53 \]

\[ \text{Gain (Type, X)} = 1 - 1 = 0 \]
Attributes with Many Values

• Problem
  – If attribute has many values, InfoGain() will select it
  – e.g., imagine using \texttt{date = Jan\_28\_2011} as an attribute

• Alternative approach: use GainRatio() instead

\[
GainRatio(X, A) = \frac{InfoGain(X, A)}{SplitInformation(X, A)}
\]

\[
SplitInformation(X, A) = - \sum_{v \in values(A)} \frac{|X_v|}{|X|} \log_2 \frac{|X_v|}{|X|}
\]

where \( X_v \) is a subset of \( X \) for which \( A \) has value \( v \)

Based on slide by Pedro Domingos
### Computing Gain Ratio

Already computed:
- $I(X) = 1$
- $I(Pat, X) = 0.47$
- $I(Type, X) = 1$
- $Gain(Pat, X) = 0.53$
- $Gain(Type, X) = 0$

<table>
<thead>
<tr>
<th></th>
<th>French</th>
<th>Italian</th>
<th>Thai</th>
<th>Burger</th>
<th>Empty</th>
<th>Some</th>
<th>Full</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>N Y</td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>N Y</td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>N Y</td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>N Y</td>
</tr>
</tbody>
</table>

SplitInfo $(Pat, X) = - \left( \frac{1}{6} \log \frac{1}{6} + \frac{1}{3} \log \frac{1}{3} + \frac{1}{2} \log \frac{1}{2} \right)$

$$= \frac{1}{6} \times 2.6 + \frac{1}{3} \times 1.6 + \frac{1}{2} \times 1 = 1.47$$

SplitInfo $(Type, X) = \frac{1}{6} \log \frac{1}{6} + \frac{1}{6} \log \frac{1}{6} + \frac{1}{3} \log \frac{1}{3} + \frac{1}{3} \log \frac{1}{3}$

$$= \frac{1}{6} \times 2.6 + \frac{1}{6} \times 2.6 + \frac{1}{3} \times 1.6 + \frac{1}{3} \times 1.6 = 1.93$$

Based on Slide from M. desJardins & T. Finin
## Computing Gain Ratio

Already computed:

- \( I(X) = 1 \)
- \( I(Pat, X) = 0.47 \)
- \( I(Type, X) = 1 \)
- \( \text{Gain}(Pat, X) = 0.53 \)
- \( \text{Gain}(Type, X) = 0 \)

<table>
<thead>
<tr>
<th></th>
<th>French</th>
<th>Y</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Italian</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Thai</td>
<td>N</td>
<td>Y</td>
<td>N Y</td>
</tr>
<tr>
<td>Burger</td>
<td>N</td>
<td>Y</td>
<td>N Y</td>
</tr>
<tr>
<td></td>
<td>Empty</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Some</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Full</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\[ \text{SplitInfo}(Pat, X) = - \left( \frac{1}{6} \log_2 \frac{1}{6} + \frac{1}{3} \log_2 \frac{1}{3} + \frac{1}{2} \log_2 \frac{1}{2} \right) \]

\[ = \frac{1}{6} \times 2.6 + \frac{1}{3} \times 1.6 + \frac{1}{2} \times 1 = 1.47 \]

\[ \text{SplitInfo}(Type, X) = \frac{1}{6} \log_2 \frac{1}{6} + \frac{1}{6} \log_2 \frac{1}{6} + \frac{1}{3} \log_2 \frac{1}{3} + \frac{1}{3} \log_2 \frac{1}{3} \]

\[ = \frac{1}{6} \times 2.6 + \frac{1}{6} \times 2.6 + \frac{1}{3} \times 1.6 + \frac{1}{3} \times 1.6 = 1.93 \]

\[ \text{GainRatio}(Pat, X) = \frac{\text{Gain}(Pat, X)}{\text{SplitInfo}(Pat, X)} = \frac{0.53}{1.47} = 0.36 \]

\[ \text{GainRatio}(Type, X) = \frac{\text{Gain}(Type, X)}{\text{SplitInfo}(Type, X)} = \frac{0}{1.93} = 0 \]
Extensions of ID3

- Using gain ratios
- Real-valued data
- Noisy data and overfitting
- Generation of rules
- Setting parameters
- Cross-validation for experimental validation of performance
- C4.5 is an extension of ID3 that accounts for unavailable values, continuous attribute value ranges, pruning of decision trees, rule derivation, and so on
Real-Valued Features

- Change to binary splits by choosing a threshold
- One method:
  - Sort instances by value, identify adjacencies with different classes
    Temperature: \[40 \ 48 \ 60 \ 72 \ 80 \ 90\]
    PlayTennis: \[No \ No \ Yes \ Yes \ Yes \ No\]
  - Choose among splits by InfoGain()
Unknown Attribute Values

What if some examples are missing values of $A$?

Use training example anyway, sort through tree:

- If node $n$ tests $A$, assign most common value of $A$ among other examples sorted to node $n$
- Assign most common value of $A$ among other examples with same class label
- Assign probability $p_i$ to each possible value $v_i$ of $A$.
  
  Assign fraction $p_i$ of example to each descendent of tree

Classify new examples in same fashion
Noisy Data

• Many kinds of “noise” can occur in the examples:
  – Two examples have same attribute/value pairs, but different classifications
  – Some values of attributes are incorrect because of errors in the data acquisition process or the preprocessing phase
  – The instance was labeled incorrectly (+ instead of -)

• Also, some attributes are irrelevant to the decision-making process
  – e.g., color of a die is irrelevant to its outcome
Overfitting

• Irrelevant attributes can result in overfitting the training example data
  – If hypothesis space has many dimensions (large number of attributes), we may find meaningless regularity in the data that is irrelevant to the true, important, distinguishing features

• If we have too little training data, even a reasonable hypothesis space will ‘overfit’
Overfitting in Decision Trees

Consider adding a noisy training example to the following tree:

```
  Outlook
     /   \
  Sunny  Overcast  Rain
     |      |       |
  Yes   Yes   Yes

  Humidity
  /   \
High  Normal
     /   \
No   Yes

  Wind
  /   \
Strong  Weak
     /   \
No   Yes
```

What would be the effect of adding:

\(<\text{outlook}=\text{sunny}, \text{temperature}=\text{hot}, \text{humidity}=\text{normal}, \text{wind}=\text{strong}, \text{playTennis}=\text{No}>\)?
Overfitting

Consider error of hypothesis $h$ over

- training data: $\text{error}_{\text{train}}(h)$
- entire distribution $\mathcal{D}$ of data: $\text{error}_{\mathcal{D}}(h)$

Hypothesis $h \in H$ overfits training data if there is an alternative hypothesis $h' \in H$ such that

$$\text{error}_{\text{train}}(h) < \text{error}_{\text{train}}(h')$$

and

$$\text{error}_{\mathcal{D}}(h) > \text{error}_{\mathcal{D}}(h')$$
Overfitting in Decision Tree Learning

![Graph showing accuracy vs size of tree (number of nodes). The graph compares accuracy on training data (solid line) and accuracy on test data (dashed line). As the size of the tree increases, the accuracy on the training data generally increases, while the accuracy on the test data starts to decrease after a certain point.]
Avoiding Overfitting

How can we avoid overfitting?
• Stop growing when data split is not statistically significant
• Acquire more training data
• Remove irrelevant attributes (manual process – not always possible)
• Grow full tree, then post-prune

How to select “best” tree:
• Measure performance over training data
• Measure performance over separate validation data set
• Add complexity penalty to performance measure
Reduced-Error Pruning

Split data into *training* and *validation* sets

Grow tree based on *training set*

Do until further pruning is harmful:

1. Evaluate impact on validation set of pruning each possible node (plus those below it)
2. Greedily remove the node that most improves *validation set* accuracy
Pruning Decision Trees

• Pruning of the decision tree is done by replacing a whole subtree by a leaf node.

• The replacement takes place if a decision rule establishes that the expected error rate in the subtree is greater than in the single leaf.

• For example,

Training

<table>
<thead>
<tr>
<th>Color</th>
<th>1 positive</th>
<th>0 negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>red</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>blue</td>
<td>2</td>
<td>0</td>
</tr>
</tbody>
</table>

Validation

<table>
<thead>
<tr>
<th>Color</th>
<th>1 positive</th>
<th>1 negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>red</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>blue</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

If we had simply predicted the majority class (negative), we make 2 errors instead of 4.

2 correct
4 incorrect
Effect of Reduced-Error Pruning

Accuracy vs. Size of tree (number of nodes)

- On training data
- On test data
- On test data (during pruning)
Effect of Reduced-Error Pruning

The tree is pruned back to the red line where it gives more accurate results on the test data.

On training data it looks great

But that's not the case for the test data

On training data
On test data
On test data (during pruning)
Converting a Tree to Rules

IF \((\text{Outlook} = \text{Sunny}) \text{ AND } (\text{Humidity} = \text{High})\)
THEN \(\text{PlayTennis} = \text{No}\)

IF \((\text{Outlook} = \text{Sunny}) \text{ AND } (\text{Humidity} = \text{Normal})\)
THEN \(\text{PlayTennis} = \text{Yes}\)

\ldots
Converting Decision Trees to Rules

• It is easy to derive rules from a decision tree: write a rule for each path from the root to a leaf

  \((\text{Outlook} = \text{Sunny}) \text{ AND } (\text{Humidity} = \text{High}) \rightarrow \text{PlayTennis} = \text{No}\)

• To simplify the resulting rule set:
  – Let LHS be the left-hand side of a rule
  – LHS’ obtained from LHS by eliminating some conditions
  – Replace LHS by LHS' in this rule if the subsets of the training set satisfying LHS and LHS' are equal
  – A rule may be eliminated by using meta-conditions such as “if no other rule applies”
Rule Post-Pruning

1. Convert tree to equivalent set of rules
2. Prune each rule independently of others
3. Sort final rules into desired sequence for use

Perhaps most frequently used method (e.g., C4.5)
Scaling Up

• ID3, C4.5, etc.: assumes that data fits in memory (OK for up to hundreds of thousands of examples)

• SPRINT, SLIQ: multiple sequential scans of data (OK for up to millions of examples)

• VFDT: at most one sequential scan (OK for up to billions of examples)
Comparison of Learning Methods

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Neural Nets</th>
<th>SVM</th>
<th>Trees</th>
<th>MARS</th>
<th>k-NN, Kernels</th>
</tr>
</thead>
<tbody>
<tr>
<td>Natural handling of data of “mixed” type</td>
<td>▼</td>
<td>▼</td>
<td>▲</td>
<td>▲</td>
<td>▼</td>
</tr>
<tr>
<td>Handling of missing values</td>
<td>▼</td>
<td>▼</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
</tr>
<tr>
<td>Robustness to outliers in input space</td>
<td>▼</td>
<td>▼</td>
<td>▲</td>
<td>▼</td>
<td>▲</td>
</tr>
<tr>
<td>Insensitive to monotone transformations of inputs</td>
<td>▼</td>
<td>▼</td>
<td>▲</td>
<td>▼</td>
<td>▼</td>
</tr>
<tr>
<td>Computational scalability (large $N$)</td>
<td>▼</td>
<td>▼</td>
<td>▲</td>
<td>▲</td>
<td>▼</td>
</tr>
<tr>
<td>Ability to deal with irrelevant inputs</td>
<td>▼</td>
<td>▼</td>
<td>▲</td>
<td>▲</td>
<td>▼</td>
</tr>
<tr>
<td>Ability to extract linear combinations of features</td>
<td>▲</td>
<td>▲</td>
<td>▼</td>
<td>▼</td>
<td></td>
</tr>
<tr>
<td>Interpretability</td>
<td>▼</td>
<td>▼</td>
<td>▲</td>
<td>▲</td>
<td>▼</td>
</tr>
<tr>
<td>Predictive power</td>
<td>▲</td>
<td>▲</td>
<td>▼</td>
<td>▼</td>
<td>▲</td>
</tr>
</tbody>
</table>

[Table 10.3 from Hastie, et al. Elements of Statistical Learning, 2nd Edition]
Summary: Decision Tree Learning

• Representation: decision trees
• Bias: prefer small decision trees
• Search algorithm: greedy
• Heuristic function: information gain or information content or others
• Overfitting / pruning
Summary: Decision Tree Learning

• Widely used in practice

• Strengths include
  – Fast and simple to implement
  – Can convert to rules
  – Handles noisy data

• Weaknesses include
  – Univariante splits/partitioning using only one attribute at a time --- limits types of possible trees
  – Large decision trees may be hard to understand
  – Requires fixed-length feature vectors
  – Non-incremental (i.e., batch method)
  – Sacrifices predictive power