Decision Trees
Reminders

• Homework 1: Background
  – Out: Wed, Jan 17 (today)
  – Due: Wed, Jan 24 at 11:59pm
  – Two parts: written part on Canvas, programming part on Autolab
  – unique policy for this assignment: unlimited submissions (i.e. keep submitting until you get 100%)
ML as Function Approximation

Chalkboard

- ML as Function Approximation
  - Problem setting
  - Input space
  - Output space
  - Unknown target function
  - Hypothesis space
  - Training examples
DECISION TREES
Decision Trees

Chalkboard

– Example: Medical Diagnosis
– Does memorization = learning?
– Decision Tree as a hypothesis
– Function approximation for DTs
– Decision Tree Learning
Tree to Predict C-Section Risk

Learned from medical records of 1000 women
Negative examples are C-sections

[S_333+, {S_167-}] .83+ .17-
Fetal_Presentation = 1: [S_822+, {S_116-}] .88+ .12-
| Previous_Csection = 0: [S_767+, {S_81-}] .90+ .10-
| | Primiparous = 0: [S_399+, {S_13-}] .97+ .03-
| | Primiparous = 1: [S_368+, {S_68-}] .84+ .16-
| | | Fetal_Distress = 0: [S_334+, {S_47-}] .88+ .12-
| | | | Birth_Weight < 3349: [S_201+, {S_10.6-}] .95+
| | | | Birth_Weight >= 3349: [S_133+, {S_36.4-}] .78+
| | | | Fetal_Distress = 1: [S_34+, {S_21-}] .62+ .38-
| Previous_Csection = 1: [S_55+, {S_35-}] .61+ .39-
Fetal_Presentation = 2: [S_3+, {S_29-}] .11+ .89-
Fetal_Presentation = 3: [S_8+, {S_22-}] .27+ .73-

(Sims et al., 2000)
Decision Trees

Chalkboard

– Information Theory primer
  • Entropy
  • (Specific) Conditional Entropy
  • Conditional Entropy
  • Information Gain / Mutual Information

– Information Gain as DT splitting criterion
Tennis Example

Dataset:

<table>
<thead>
<tr>
<th>Day</th>
<th>Outlook</th>
<th>Temperature</th>
<th>Humidity</th>
<th>Wind</th>
<th>PlayTennis?</th>
</tr>
</thead>
<tbody>
<tr>
<td>D1</td>
<td>Sunny</td>
<td>Hot</td>
<td>High</td>
<td>Weak</td>
<td>No</td>
</tr>
<tr>
<td>D2</td>
<td>Sunny</td>
<td>Hot</td>
<td>High</td>
<td>Strong</td>
<td>No</td>
</tr>
<tr>
<td>D3</td>
<td>Overcast</td>
<td>Hot</td>
<td>High</td>
<td>Weak</td>
<td>Yes</td>
</tr>
<tr>
<td>D4</td>
<td>Rain</td>
<td>Mild</td>
<td>High</td>
<td>Weak</td>
<td>Yes</td>
</tr>
<tr>
<td>D5</td>
<td>Rain</td>
<td>Cool</td>
<td>Normal</td>
<td>Weak</td>
<td>Yes</td>
</tr>
<tr>
<td>D6</td>
<td>Rain</td>
<td>Cool</td>
<td>Normal</td>
<td>Strong</td>
<td>No</td>
</tr>
<tr>
<td>D7</td>
<td>Overcast</td>
<td>Cool</td>
<td>Normal</td>
<td>Strong</td>
<td>Yes</td>
</tr>
<tr>
<td>D8</td>
<td>Sunny</td>
<td>Mild</td>
<td>High</td>
<td>Weak</td>
<td>No</td>
</tr>
<tr>
<td>D9</td>
<td>Sunny</td>
<td>Cool</td>
<td>Normal</td>
<td>Weak</td>
<td>Yes</td>
</tr>
<tr>
<td>D10</td>
<td>Rain</td>
<td>Mild</td>
<td>Normal</td>
<td>Weak</td>
<td>Yes</td>
</tr>
<tr>
<td>D11</td>
<td>Sunny</td>
<td>Mild</td>
<td>Normal</td>
<td>Strong</td>
<td>Yes</td>
</tr>
<tr>
<td>D12</td>
<td>Overcast</td>
<td>Mild</td>
<td>High</td>
<td>Strong</td>
<td>Yes</td>
</tr>
<tr>
<td>D13</td>
<td>Overcast</td>
<td>Hot</td>
<td>Normal</td>
<td>Weak</td>
<td>Yes</td>
</tr>
<tr>
<td>D14</td>
<td>Rain</td>
<td>Mild</td>
<td>High</td>
<td>Strong</td>
<td>No</td>
</tr>
</tbody>
</table>
Tennis Example

Which attribute yields the best classifier?
Tennis Example

Which attribute should be tested here?

\[ S_{\text{Sunny}} = \{D_1, D_2, D_8, D_9, D_{11}\} \]

\[ \text{Gain} \left( S_{\text{Sunny}}, \text{Humidity} \right) = 0.970 - \frac{3}{5} 0.0 - \frac{2}{5} 0.0 = 0.970 \]

\[ \text{Gain} \left( S_{\text{Sunny}}, \text{Temperature} \right) = 0.970 - \frac{2}{5} 0.0 - \frac{2}{5} 1.0 - \frac{1}{5} 0.0 = 0.570 \]

\[ \text{Gain} \left( S_{\text{Sunny}}, \text{Wind} \right) = 0.970 - \frac{2}{5} 1.0 - \frac{3}{5} 0.918 = 0.019 \]
Decision Tree Learning Example

In-Class Exercise

1. Which attribute would misclassification rate select for the next split?
2. Which attribute would information gain select for the next split?

Dataset:
Output Y, Attributes A and B

<table>
<thead>
<tr>
<th>Y</th>
<th>A</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
Decision Trees

Chalkboard

– ID3 as Search
– Inductive Bias of Decision Trees
– Occam’s Razor
Overfitting

Consider a hypothesis \( h \) and its

- Error rate over training data: \( \text{error}_{\text{train}}(h) \)
- True error rate over all data: \( \text{error}_{\text{true}}(h) \)

We say \( h \) **overfits** the training data if

\[
\text{error}_{\text{true}}(h) > \text{error}_{\text{train}}(h)
\]

Amount of overfitting =

\[
\text{error}_{\text{true}}(h) - \text{error}_{\text{train}}(h)
\]
Overfitting in Decision Tree Learning

Figure from Tom Mitchell
How to Avoid Overfitting?

For Decision Trees...

1. Do not grow tree beyond some **maximum depth**
2. Do not split if splitting criterion (e.g. Info. Gain) is **below some threshold**
3. Stop growing when the split is **not statistically significant**
4. Grow the entire tree, then prune
Reduced-Error Pruning

Split data into *training* and *validation* set

Create tree that classifies *training* set correctly

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 one that most improves *validation* set accuracy

- produces smallest version of most accurate subtree

- What if data is limited?
Effect of Reduced-Error Pruning

Slide from Tom Mitchell
Questions

• Will ID3 always include all the attributes in the tree?
• What if some attributes are real-valued? Can learning still be done efficiently?
• What if some attributes are missing?
Learning Objectives

You should be able to...

1. Implement Decision Tree training and prediction
2. Use effective splitting criteria for Decision Trees and be able to define entropy, conditional entropy, and mutual information / information gain
3. Explain the difference between memorization and generalization [CIML]
4. Describe the inductive bias of a decision tree
5. Formalize a learning problem by identifying the input space, output space, hypothesis space, and target function
6. Explain the difference between true error and training error
7. Judge whether a decision tree is "underfitting" or "overfitting"
8. Implement a pruning or early stopping method to combat overfitting in Decision Tree learning