Lecture 2: Decision Trees

Machine Learning
Queens College
Today

• Decision Trees
  – Entropy and Information Theory
  – Information Gain
  – “Purity” or “Classification Accuracy”
Decision Trees

- Trees that define a decision process
  - Internal nodes: questions associated with a specific feature
  - Leaves: Decisions

- Want a fast meal?
  - Yes
  - How about coffee?
    - Yes: Starbucks
    - No: McDonald’s
  - No
    - On expense account?
      - Yes: 21 Club
      - No: T.G.I. Friday’s
• Very easy to evaluate.
• Nested if statements
More formal Definition of a Decision Tree

• A **Tree** data structure
• Each **internal node** corresponds to a feature
• **Leaves** are associated with target values.
• Nodes with **nominal features** have \( N \) children, where \( N \) is the number of nominal values
• Nodes with **continuous features** have two children for values less than and greater than or equal to a **break point**.
Training a Decision Tree

• How do you decide what feature to use?
• For continuous features how do you decide what break point to use?

• Goal: Optimize Classification Accuracy.
## Example Data Set

<table>
<thead>
<tr>
<th>Height</th>
<th>Weight</th>
<th>Eye Color</th>
<th>Gender</th>
</tr>
</thead>
<tbody>
<tr>
<td>66</td>
<td>170</td>
<td>Blue</td>
<td>Male</td>
</tr>
<tr>
<td>73</td>
<td>210</td>
<td>Brown</td>
<td>Male</td>
</tr>
<tr>
<td>72</td>
<td>165</td>
<td>Green</td>
<td>Male</td>
</tr>
<tr>
<td>70</td>
<td>180</td>
<td>Blue</td>
<td>Male</td>
</tr>
<tr>
<td>74</td>
<td>185</td>
<td>Brown</td>
<td>Male</td>
</tr>
<tr>
<td>68</td>
<td>155</td>
<td>Green</td>
<td>Male</td>
</tr>
<tr>
<td>65</td>
<td>150</td>
<td>Blue</td>
<td>Female</td>
</tr>
<tr>
<td>64</td>
<td>120</td>
<td>Brown</td>
<td>Female</td>
</tr>
<tr>
<td>63</td>
<td>125</td>
<td>Green</td>
<td>Female</td>
</tr>
<tr>
<td>67</td>
<td>140</td>
<td>Blue</td>
<td>Female</td>
</tr>
<tr>
<td>68</td>
<td>165</td>
<td>Brown</td>
<td>Female</td>
</tr>
<tr>
<td>66</td>
<td>130</td>
<td>Green</td>
<td>Female</td>
</tr>
</tbody>
</table>
Baseline Classification Accuracy

• Select the majority class.
  – Here 6/12 Male, 6/12 Female.
  – Baseline Accuracy: 50%

• How good is each branch?
  – The improvement to classification accuracy
Training Example

• Possible branches

```
<table>
<thead>
<tr>
<th>color</th>
<th>blue</th>
<th>brown</th>
<th>green</th>
</tr>
</thead>
<tbody>
<tr>
<td>M/F</td>
<td>2/2</td>
<td>2/2</td>
<td>2/2</td>
</tr>
</tbody>
</table>
```
<table>
<thead>
<tr>
<th>Height</th>
<th>Weight</th>
<th>Eye Color</th>
<th>Gender</th>
</tr>
</thead>
<tbody>
<tr>
<td>63</td>
<td>125</td>
<td>Green</td>
<td>Female</td>
</tr>
<tr>
<td>64</td>
<td>120</td>
<td>Brown</td>
<td>Female</td>
</tr>
<tr>
<td>65</td>
<td>150</td>
<td>Blue</td>
<td>Female</td>
</tr>
<tr>
<td>66</td>
<td>170</td>
<td>Blue</td>
<td>Male</td>
</tr>
<tr>
<td>66</td>
<td>130</td>
<td>Green</td>
<td>Female</td>
</tr>
<tr>
<td>67</td>
<td>140</td>
<td>Blue</td>
<td>Female</td>
</tr>
<tr>
<td>68</td>
<td>165</td>
<td>Brown</td>
<td>Female</td>
</tr>
<tr>
<td>68</td>
<td>155</td>
<td>Green</td>
<td>Male</td>
</tr>
<tr>
<td>70</td>
<td>180</td>
<td>Blue</td>
<td>Male</td>
</tr>
<tr>
<td>72</td>
<td>165</td>
<td>Green</td>
<td>Male</td>
</tr>
<tr>
<td>73</td>
<td>210</td>
<td>Brown</td>
<td>Male</td>
</tr>
<tr>
<td>74</td>
<td>185</td>
<td>Brown</td>
<td>Male</td>
</tr>
</tbody>
</table>
Training Example

- Possible branches

\[
\begin{align*}
\text{height} \\
<68 &\quad 1M / 5F \\
&\quad 5M / 1F
\end{align*}
\]

50% Accuracy before Branch

83.3% Accuracy after Branch

33.3% Accuracy Improvement
## Weight

<table>
<thead>
<tr>
<th>Height</th>
<th>Weight</th>
<th>Eye Color</th>
<th>Gender</th>
</tr>
</thead>
<tbody>
<tr>
<td>64</td>
<td>120</td>
<td>Brown</td>
<td>Female</td>
</tr>
<tr>
<td>63</td>
<td>125</td>
<td>Green</td>
<td>Female</td>
</tr>
<tr>
<td>66</td>
<td>130</td>
<td>Green</td>
<td>Female</td>
</tr>
<tr>
<td>67</td>
<td>140</td>
<td>Blue</td>
<td>Female</td>
</tr>
<tr>
<td>65</td>
<td>150</td>
<td>Blue</td>
<td>Female</td>
</tr>
<tr>
<td>68</td>
<td>155</td>
<td>Green</td>
<td>Male</td>
</tr>
<tr>
<td>68</td>
<td>165</td>
<td>Brown</td>
<td>Female</td>
</tr>
<tr>
<td>72</td>
<td>165</td>
<td>Green</td>
<td>Male</td>
</tr>
<tr>
<td>66</td>
<td>170</td>
<td>Blue</td>
<td>Male</td>
</tr>
<tr>
<td>70</td>
<td>180</td>
<td>Blue</td>
<td>Male</td>
</tr>
<tr>
<td>74</td>
<td>185</td>
<td>Brown</td>
<td>Male</td>
</tr>
<tr>
<td>73</td>
<td>210</td>
<td>Brown</td>
<td>Male</td>
</tr>
</tbody>
</table>
Training Example

• Possible branches

- weight
  - <155
    - 0M / 5F
    - 50% Accuracy before Branch
    - 91.7% Accuracy after Branch
    - 41.7% Accuracy Improvement
  - 6M / 1F
Training Example

• Recursively train child nodes.
Training Example

• Finished Tree

- **color**
  - blue
  - brown
  - green

- **height**
  - <70
    - 2M / 1F
  - >70
    - 6M / 1F

- **weight**
  - <155
    - 5F
  - >155
    - 4M
Generalization

• What is the performance of the tree on the training data?
  – Is there any way we could get less than 100% accuracy?

• What performance can we expect on unseen data?
Evaluation

• Evaluate performance on data that was not used in training.
• Isolate a subset of data points to be used for evaluation.
• Evaluate generalization performance.
Evaluation of our Decision Tree

• What is the Training performance?
• What is the Evaluation performance?
  – Never classify female over 165
  – Never classify male under 165, and under 68.
  – The middle section is trickier.
• What are some ways to make these similar?
Pruning

- There are many pruning techniques.
- A simple approach is to have a minimum membership size in each node.
Decision Trees

- Training via **Recursive Partitioning**.
- Simple, interpretable models.
- Different node selection criteria can be used.
  - Information theory is a common choice.
- Pruning techniques can be used to make the model more robust to unseen data.
Entropy and Information Theory

• Entropy is a measure of how homogenous a data set is.
  – Also how even a probability distribution or a random variable is.
• The unit of Entropy is the bit.
• Under an Information Theory perspective entropy represents the fewest bits it would take on average to transmit information in a signal (i.e. a random variable)
Entropy

• Say I have a vocabulary of 4 items.
  – A, B, C, D.

• A standard encoding of these might be
  – 00, 01, 10, 11.

• 2 bits per vocabulary item.

• However, if A is much more common, it might be more efficient to use this coding
  – 0, 10, 111, 110

• Exercise: What is the average bit length if there are 150 As, 40 Bs, 5 Cs, and 5Ds?

\[
\frac{150 \times 1 + 40 \times 2 + 5 \times 3 + 5 \times 3}{250} = \frac{250}{250} = 1.3
\]
Calculating Entropy

\[ H(X) = - \sum_{i \in X} p_i \log p_i \]

- Where \( p_i \) is the probability of selecting the ith value.
- For example, say \( X = \{A A A B B B B B B\} \)
- In the calculation of entropy \( 0 \log 0 = 0 \)

\[ H(X) = - \left( \frac{3}{8} \log \frac{3}{8} + \frac{5}{8} \log \frac{5}{8} \right) \]
Information Gain

• In our previous example we examined the improvement to classification performance.
  – Error reduction or change to overall accuracy.

• Using entropy the measure that is optimized is Information Gain.
  – The difference in the entropy of the label or class distribution before or after a particular decision tree split.
Calculating Information Gain

$$Gain(S, F) = H(S) - \sum_{f \in \text{values}(F)} \frac{|S_f|}{|S|} H(S_f)$$

$$H(S) = -\frac{3}{4} \log_2 \frac{3}{4} - \frac{1}{4} \log_2 \frac{1}{4}$$

$$H(S_{VW}) = -\frac{1}{4} \log_2 \frac{1}{4} - \frac{1}{2} \log_2 \frac{1}{2} = 0$$

$$H(S_{Ford}) = -\frac{1}{4} \log_2 \frac{1}{4} - \frac{1}{2} \log_2 \frac{1}{2} = 1$$

$$H(S_{BMW}) = -\frac{1}{4} \log_2 \frac{1}{4} - \frac{1}{2} \log_2 \frac{1}{2} = 0$$

$$Gain(S, F) = H(S) - \frac{1}{4} \times 0 - \frac{3}{4} \times 1 - \frac{1}{4} \times 0 = 0$$
Calculating Information Gain

\[ \text{Gain}(S, F) = H(S) - \sum_{f \in \text{values}(F)} \frac{|S_f|}{|S|} H(S_f) \]

Identify the feature with the greatest Information Gain and repeat this process recursively!
Visualization of Decision Tree Training
Visualization of Decision Tree Training
Visualization of Decision Tree Training
Visualization of Decision Tree Training
Visualization of Decision Tree Training
Visualization of Decision Tree Training
Visualization of Decision Tree Training
Next Time: Math Primer

- Probability
  - Bayes Rule
  - Naïve Bayes Classification

- Statistics
  - Normal Distribution
  - Multinomial Distribution