Decision trees

Oliver W. Layton

CS251: Data analysis and visualization

Lecture 32, Spring 2019

Friday April 26
Plan

• Information entropy and feature selection
• ID3 algorithm
Designing a decision tree: Number of branches

• **What we just learned:** How to judge the quality of different split rules for a fixed feature. Other tree design considerations:
  
  • Number of branches/splits allowed per node?
    
    • Often fixed by designer (e.g. always 2). Can vary within tree (e.g. at most $M$).
Designing a decision tree: Height of tree

- Can implement hard limit on maximum height (e.g. 2 nodes deep).
- Enforce height by purity threshold (node is "pure enough" to become a leaf node).
  Example: stop branching if we have 3 or fewer counts in the non-majority classes:
  
  \[10, 3, 1]\n
- We need to update our definition of a **leaf node**.
  - Whether pure or impure, classify data point according to **majority class**: 
    
    \[3, 0]; [2, 1]\n
- "**Pure enough**" leaf nodes allow decision trees to handle noisy data (Conflicting class labels for same combo of features, Allen's point)
ID3 Algorithm: third iterative dichotomizer

How do we know which feature to put in which node?

1. Compute information gain for each unused feature.

2. Also use information gain to determine "best" split point WITHIN each unused feature.

3. Assign to the next node the feature with maximal information gain.

4. Stop when every branch ends with a leaf node (pure or "pure enough" node) OR when all the features are used up.

• Let's use ID3 to create a decision tree based on the tennis data.
<table>
<thead>
<tr>
<th>Day</th>
<th>Outlook</th>
<th>Temp.</th>
<th>Humidity</th>
<th>Wind</th>
<th>Play Tennis</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>Weak</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>Cold</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>Strong</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>