Data Mining:
Classification/Supervised Learning
Examples

Open Houses

A family is looking to buy a house. Their search leads them to a list of open houses in their town that they can visit. We have information about the number of bedrooms, existence of basement, type of floorplan and geographical location of the house. We also have information on whether the family chose to visit the open house.

<table>
<thead>
<tr>
<th>HouseId</th>
<th>Bedrooms</th>
<th>Basement</th>
<th>Floorplan</th>
<th>Location</th>
<th>Visited</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3</td>
<td>No</td>
<td>traditional</td>
<td>South</td>
<td>No</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>Yes</td>
<td>traditional</td>
<td>South</td>
<td>Yes</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>Yes</td>
<td>open</td>
<td>North</td>
<td>No</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>Yes</td>
<td>traditional</td>
<td>North</td>
<td>No</td>
</tr>
<tr>
<td>5</td>
<td>3</td>
<td>No</td>
<td>open</td>
<td>North</td>
<td>No</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
<td>Yes</td>
<td>traditional</td>
<td>South</td>
<td>Yes</td>
</tr>
<tr>
<td>7</td>
<td>3</td>
<td>Yes</td>
<td>open</td>
<td>South</td>
<td>No</td>
</tr>
<tr>
<td>8</td>
<td>3</td>
<td>No</td>
<td>traditional</td>
<td>South</td>
<td>No</td>
</tr>
<tr>
<td>9</td>
<td>4</td>
<td>No</td>
<td>traditional</td>
<td>South</td>
<td>Yes</td>
</tr>
<tr>
<td>10</td>
<td>4</td>
<td>Yes</td>
<td>open</td>
<td>North</td>
<td>No</td>
</tr>
<tr>
<td>11</td>
<td>4</td>
<td>Yes</td>
<td>open</td>
<td>South</td>
<td>Yes</td>
</tr>
<tr>
<td>12</td>
<td>4</td>
<td>No</td>
<td>traditional</td>
<td>North</td>
<td>No</td>
</tr>
<tr>
<td>13</td>
<td>4</td>
<td>No</td>
<td>open</td>
<td>South</td>
<td>Yes</td>
</tr>
<tr>
<td>14</td>
<td>4</td>
<td>Yes</td>
<td>open</td>
<td>South</td>
<td>Yes</td>
</tr>
<tr>
<td>15</td>
<td>4</td>
<td>No</td>
<td>traditional</td>
<td>North</td>
<td>No</td>
</tr>
<tr>
<td>16</td>
<td>4</td>
<td>Yes</td>
<td>open</td>
<td>North</td>
<td>No</td>
</tr>
</tbody>
</table>

Our goal is to build a decision-tree classifier that predicts whether a family will visit a specific house.
C4.5. for the Open Houses dataset

**Step 1.** Determine the root node. Input: full dataset \( D \), Attributes: \{Bedrooms, Basement, Floorplan, Location\}.

**Note:** Termination conditions on step 1 are false.

Info Gain computation.

\[
Pr(Visited = Yes) = \frac{6}{10} = 0.375.
Pr(Visited = No) = \frac{4}{10} = 0.625.
\]

\[
\text{entropy}(D) = -\frac{3}{8} \cdot \log_2 \frac{3}{8} - \frac{5}{8} \cdot \log_2 \frac{5}{8} = 0.9544
\]

**Bedrooms.** \( D_1 = \{1, 2, 3, 4, 5, 6, 7, 8\}; \ D_2 = \{9, 10, 11, 12, 13, 14, 15, 16\}.\)

\[
\text{entropy}(D_1) = -\frac{2}{5} \cdot \log_2 \frac{2}{5} - \frac{3}{5} \cdot \log_2 \frac{3}{5} = 0.811
\]

\[
\text{entropy}(D_2) = -\frac{2}{5} \cdot \log_2 \frac{2}{5} - \frac{3}{5} \cdot \log_2 \frac{3}{5} = 1
\]

\[
\text{Gain}(\text{Bedrooms}) = 0.9544 - 0.9055 = 0.0489
\]

**Basement.** \( D_1 = D_{\text{Yes}} = \{2, 3, 4, 6, 7, 10, 11, 14, 16\}; \ D_2 = D_{\text{No}} = \{1, 5, 8, 9, 12, 13, 15\}.\)

\[
Pr_{D_1}(Visited = Yes) = \frac{1}{4}
Pr_{D_2}(Visited = Yes) = \frac{3}{4}
\]

\[
\text{entropy}(D_1) = -\frac{2}{7} \cdot \log_2 \frac{2}{7} - \frac{5}{7} \cdot \log_2 \frac{5}{7} = 0.9107
\]

\[
\text{entropy}(D_2) = -\frac{2}{7} \cdot \log_2 \frac{2}{7} - \frac{5}{7} \cdot \log_2 \frac{5}{7} = 0.8631
\]

\[
\text{Gain}(\text{Basement}) = 0.9544 - 0.9350 = 0.0193
\]

**Floorplan.** \( D_1 = D_{\text{traditional}} = \{1, 2, 4, 6, 8, 9, 12, 15\}; \ D_2 = D_{\text{Open}} = \{3, 5, 7, 10, 11, 13, 14, 16\}.\)

\[
Pr_{D_1}(Visited = Yes) = \frac{3}{8}
Pr_{D_2}(Visited = Yes) = \frac{5}{8}
\]

\[
\text{entropy}(D_1) = -\frac{3}{8} \cdot \log_2 \frac{3}{8} - \frac{5}{8} \cdot \log_2 \frac{5}{8} = 0.9544
\]

\[
\text{entropy}(D_2) = -\frac{3}{8} \cdot \log_2 \frac{3}{8} - \frac{5}{8} \cdot \log_2 \frac{5}{8} = 0.9544
\]

\[
\text{Gain}(\text{Floorplan}) = 0.9544 - 0.9544 = 0
\]

**Location.** \( D_1 = D_{\text{North}} = \{3, 4, 5, 10, 12, 15, 16\}; \ D_2 = D_{\text{South}} = \{1, 2, 6, 7, 8, 9, 11, 13, 14\}.\)

\[
Pr_{D_1}(Visited = Yes) = \frac{0}{9} = 0
Pr_{D_2}(Visited = Yes) = \frac{9}{9} = 1
\]

\[
\text{entropy}(D_1) = -\frac{2}{7} \cdot \log_2 \frac{2}{7} - \frac{5}{7} \cdot \log_2 \frac{5}{7} = 0
\]

\[
\text{entropy}(D_2) = -\frac{2}{7} \cdot \log_2 \frac{2}{7} - \frac{5}{7} \cdot \log_2 \frac{5}{7} = 0.918
\]

\[
\text{Gain}(\text{Location}) = \frac{7}{16} \cdot 0 + \frac{9}{16} \cdot 0.918 = 0.516
\]

\[
\text{Gain}(\text{Location}) = 0.9544 - 0.516 = 0.438
\]

**Step 1 result:** Location yields the best Information Gain.

Splitting the dataset on Location attribute: \( D_1 = D_{\text{North}} = \{3, 4, 5, 10, 12, 15, 16\}; \)

\( D_2 = D_{\text{South}} = \{1, 2, 6, 7, 8, 9, 11, 13, 14\}.\)
Step 2: Input: \(D_1 = \{3, 4, 5, 10, 12, 15, 16\}\). Attributes: \{Bedrooms, Basement, Floorplan\}.

Note: This dataset is homogenous: no houses from \(D_1\) were visited.

\[\text{Class}(D_1) = \text{No}.\]

Step 3: Input \(D_2 = D_{\text{South}} = \{1, 2, 6, 7, 8, 9, 11, 13, 14\}\). Attributes: \{Bedrooms, Basement, Floorplan\}.

\[\text{entropy}(D_2) = 0.918\]

Bedrooms. \(D_{21} = D_{\text{South,3br}} = \{1, 2, 6, 7, 8\}\); \(D_{22} = D_{\text{South,4br}} \{9, 11, 13, 14\}\).

\[\Pr_{D_{21}}(\text{Visited} = \text{Yes}) = \frac{2}{5}\]

\[\Pr_{D_{22}}(\text{Visited} = \text{Yes}) = \frac{2}{4}\]

\[\text{entropy}(D_{21}) = -\frac{2}{5} \log_2 \frac{2}{5} - \frac{3}{5} \log_2 \frac{3}{5} = 0.9709\]

\[\text{entropy}(D_{22}) = -\frac{2}{4} \log_2 \frac{2}{4} - \frac{2}{4} \log_2 \frac{2}{4} = 0\]

\[\text{entropy}_{\text{Bedrooms}}(D_2) = \frac{2}{5} \cdot 0.9709 + \frac{3}{5} \cdot 0 = 0.5391\]

\[\text{Gain}_{\text{Bedrooms}}(D) = 0.918 - 0.5391 = 0.3785\]

Basement. \(D_{21} = D_{\text{South,Yes}} = \{2, 6, 7, 11, 14\}; D_{22} = D_{\text{No}} = \{1, 8, 9, 13\}\).

\[\Pr_{D_{21}}(\text{Visited} = \text{Yes}) = \frac{4}{5}\]

\[\Pr_{D_{22}}(\text{Visited} = \text{Yes}) = \frac{3}{4}\]

\[\text{entropy}(D_{21}) = -\frac{4}{5} \log_2 \frac{4}{5} - \frac{1}{5} \log_2 \frac{1}{5} = 0.7219\]

\[\text{entropy}(D_{22}) = -\frac{3}{4} \log_2 \frac{3}{4} - \frac{1}{4} \log_2 \frac{1}{4} = 1\]

\[\text{entropy}_{\text{Basement}}(D) = \frac{3}{5} \cdot 0.7219 + \frac{2}{5} \cdot 1 = 0.8455\]

\[\text{Gain}_{\text{Basement}}(D) = 0.918 - 0.8455 = 0.0725\]

Floorplan. \(D_{21} = D_{\text{South,traditional}} = \{1, 2, 6, 8, 9\}; D_{22} = D_{\text{No}} = \{7, 11, 13, 14\}\).

\[\Pr_{D_{21}}(\text{Visited} = \text{Yes}) = \frac{3}{5}\]

\[\Pr_{D_{22}}(\text{Visited} = \text{Yes}) = \frac{3}{4}\]

\[\text{entropy}(D_{21}) = -\frac{3}{5} \log_2 \frac{3}{5} - \frac{2}{5} \log_2 \frac{2}{5} = 0.9709\]

\[\text{entropy}(D_{22}) = -\frac{3}{4} \log_2 \frac{3}{4} - \frac{1}{4} \log_2 \frac{1}{4} = 0.8112\]

\[\text{entropy}_{\text{Floorplan}}(D) = \frac{3}{5} \cdot 0.9709 + \frac{2}{5} \cdot 0.8112 = 0.8999\]

\[\text{Gain}_{\text{Floorplan}}(D) = 0.918 - 0.8999 = 0.0181\]

Step 3 result: Bedroms yields the best Information Gain.

Step 4: \(D_{21} = D_{\text{South,3br}} = \{1, 2, 6, 7, 8\}\). Attributes: \{Basement, Floorplan\}.

\[\text{entropy}(D_{21}) = 0.97095\]

Basement. \(D_{211} = D_{\text{South,3br,Yes}} = \{2, 6, 7\}; D_{212} = D_{\text{South,3br,No}} = \{1, 8\}\).

\[\Pr_{D_{211}}(\text{Visited} = \text{Yes}) = \frac{2}{3}\]

\[\Pr_{D_{212}}(\text{Visited} = \text{Yes}) = \frac{2}{2} = 0\]

\[\text{entropy}(D_{211}) = -\frac{2}{3} \log_2 \frac{2}{3} - \frac{1}{3} \log_2 \frac{1}{3} = 0.918\]

\[\text{entropy}(D_{212}) = -\frac{2}{2} \log_2 \frac{2}{2} = 0\]

\[\text{entropy}_{\text{Basement}}(D) = \frac{2}{3} \cdot 0.918 + \frac{1}{3} \cdot 0 = 0.5509\]
Gain_{Basement}(D) = 0.97095 - 0.5509 = 0.42005

Floorplan. \( D_{211} = D_{South,3br,traditional} = \{1, 2, 6, 8\}; D_{212} = D_{South,3br,open} = \{7\} \).

\begin{align*}
Pr_{D_{211}} (Visited = Yes) &= \frac{2}{3} \\
Pr_{D_{212}} (Visited = Yes) &= \frac{0}{3} = 0 \\
entropy(D_{211}) &= -\frac{2}{3} \log_2 \frac{2}{3} - \frac{1}{3} \log_2 \frac{1}{3} = 1 \\
entropy(D_{212}) &= -\frac{1}{3} \log_2 \frac{1}{3} - \frac{2}{3} \log_2 \frac{2}{3} = 0 \\
entropy_{Floorplan}(D) &= \frac{4}{9} \cdot 1 + \frac{1}{9} \cdot 0 = 0.8
\end{align*}

Gain_{Floorplan}(D) = 0.97095 - 0.8 = 0.1709

Step 4 result: Basement yields the best Information Gain

Step 5: \( D_{211} = D_{South,3br,Yes} = \{2, 6, 7\}; \) Attributes: \{Floorplan\}

\begin{align*}
entropy(D_{211}) &= 0.918
\end{align*}

Floorplan. \( D_{2111} = D_{South,3br,Yes,traditional} = \{2, 6\}; D_{2112} = D_{South,3br,Yes,open} = \{7\} \).

\begin{align*}
Pr_{D_{2111}} (Visited = Yes) &= \frac{2}{3} = 1 \\
Pr_{D_{2112}} (Visited = Yes) &= \frac{0}{3} = 0 \\
entropy(D_{2111}) &= -\frac{2}{3} \log_2 \frac{2}{3} - \frac{1}{3} \log_2 \frac{1}{3} = 0 \\
entropy(D_{2112}) &= -\frac{1}{3} \log_2 \frac{1}{3} - \frac{2}{3} \log_2 \frac{2}{3} = 0 \\
entropy_{Basement}(D) &= \frac{4}{9} \cdot 0 + \frac{1}{9} \cdot 0 = 0
\end{align*}

Gain_{Basement}(D) = 0.918 - 0.8 = 0.918

Step 5: result Floorplan achieves significant Information gain.

Steps 6 and 7: \( D_{2111} = D_{South,3br,Yes,traditional} = \{2, 6\} \) and \( D_{2112} = D_{South,3br,Yes,open} = \{7\} \): the list of attributes is exhausted.

\begin{align*}
Class(D_{2111}) &= Yes \\
Class(D_{2112}) &= No
\end{align*}

Step 8: Input: \( D_{22} = D_{South,4br} \{9, 11, 13, 14\} \). Attributes: \{Basement, Floorplan\}.

\begin{align*}
Pr_{D_{22}} (Visited = Yes) &= 1, \text{ therefore:} \\
Class(D_{22}) &= Yes
\end{align*}

Resulting Tree: Resulting tree is depicted below:
Tree structure for property details:

- **Location**
  - North
    - Not Visited
  - South

- **Bedrooms**
  - 3
  - 4

- **Basement**
  - Yes
    - Floorplan
      - traditional
        - Visited
      - open
        - Not Visited
  - No
    - Not Visited

- **Visited**
  - 5