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, existance 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.

HouseId	Bedrooms	Basement	Floorplan	Location	Visited
1	3	No	traditional	South	**No**
2	3	Yes	traditional	South	**Yes**
3	3	Yes	open	North	**No**
4	3	Yes	traditional	North	**No**
5	3	No	open	North	**No**
6	3	Yes	traditional	South	**Yes**
7	3	Yes	open	South	**No**
8	3	No	traditional	South	**No**
9	4	No	traditional	South	**Yes**
10	4	Yes	open	North	**No**
11	4	Yes	open	South	**Yes**
12	4	No	traditional	North	**No**
13	4	No	open	South	**Yes**
14	4	Yes	open	South	**Yes**
15	4	No	traditional	North	**No**
16	4	Yes	open	North	**No**

Our goal is to build a decision-tree classifier that predicts whether a faimily will visit a specific house.

C4.5. for the Open Houses dataset

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

Note: Termination conditions on step 1 are false.

Info Gain computation.

$Pr(\text{Visited} = \text{Yes}) = \frac{6}{16} = 0.375.$
$Pr(\text{Visited} = \text{No}) = \frac{10}{16} = 0.625.$

$enthropy(D) = -\frac{3}{8} \cdot \log_2 \frac{3}{8} - \frac{5}{8} \cdot log_2 \frac{5}{8} = 0.9544$

Bedrooms. $D_1 = \{1, 2, 3, 4, 5, 6, 7, 8\}$; $D_2 = \{9, 10, 11, 12, 13, 14, 15, 16\}$.
$enthropy(D_1) = -\frac{2}{8} \log_2 \frac{2}{8} - \frac{6}{8} \log_2 \frac{6}{8} = 0.811$
$enthropy(D_2) = -\frac{4}{8} \log_2 \frac{4}{8} - \frac{4}{8} \log_2 \frac{4}{8} = 1$
$enthropy_{\text{Bedrooms}}(D) = \frac{8}{16} \cdot 0.811 + \frac{8}{16} \cdot 1 = 0.9055$

$$Gain_{\text{Bedrooms}}(D) = 0.9544 - 0.9055 = 0.0489$$

Basement. $D_1 = D_{\text{Yes}} = \{2, 3, 4, 6, 7, 10, 11, 14, 16\}$; $D_2 = D_{\text{No}} = \{1, 5, 8, 9, 12, 13, 15\}$.
$Pr_{D_1}(\text{Visited} = \text{Yes}) = \frac{4}{9}$
$Pr_{D_2}(\text{Visited} = \text{Yes}) = \frac{2}{7}$
$enthropy(D_1) = -\frac{4}{9} \log_2 \frac{4}{9} - \frac{5}{9} \log_2 \frac{5}{9} = 0.99107$
$enthropy(D_2) = -\frac{2}{7} \log_2 \frac{2}{7} - \frac{5}{7} \log_2 \frac{5}{7} = 0.8631$
$enthropy_{\text{Basement}}(D) = \frac{9}{16} \cdot 0.99107 + \frac{7}{16} \cdot 0.8631 = 0.9350$

$$Gain_{\text{Basement}}(D) = 0.9544 - 0.9350 = 0.0193$$

Floorplan. $D_1 = D_{\text{traditional}} = \{1, 2, 4, 6, 8, 9, 12, 15\}$; $D_2 = D_{\text{Open}} = \{3, 5, 7, 10, 11, 13, 14, 16\}$.
$Pr_{D_1}(\text{Visited} = \text{Yes}) = \frac{3}{8}$
$Pr_{D_2}(\text{Visited} = \text{Yes}) = \frac{3}{8}$
$enthropy(D_1) = -\frac{3}{8} \log_2 \frac{3}{9} - \frac{5}{8} \log_2 \frac{5}{8} = 0.9544$
$enthropy(D_2) = -\frac{3}{8} \log_2 \frac{3}{9} - \frac{5}{8} \log_2 \frac{5}{8} = 0.9544$
$enthropy_{\text{Floorplan}}(D) = \frac{8}{16} \cdot 0.9544 + \frac{8}{16} \cdot 0.9544 = 0.9544$

$$Gain_{\text{Floorplan}}(D) = 0.9544 - 0.9544 = 0$$

Location. $D_1 = D_{\text{North}} = \{3, 4, 5, 10, 12, 15, 16\}$; $D_2 = D_{\text{South}} = \{1, 2, 6, 7, 8, 9, 11, 13, 14\}$.
$Pr_{D_1}(\text{Visited} = \text{Yes}) = \frac{0}{7} = 0$
$Pr_{D_2}(\text{Visited} = \text{Yes}) = \frac{6}{9}$
$enthropy(D_1) = -\frac{0}{7} \log_2 \frac{0}{7} - \frac{7}{7} \log_2 \frac{7}{7} = 0$
$enthropy(D_2) = -\frac{6}{9} \log_2 \frac{6}{9} - \frac{3}{9} \log_2 \frac{3}{9} = 0.918$
$enthropy_{\text{Location}}(D) = \frac{7}{16} \cdot 0 + \frac{9}{16} \cdot 0.918 = 0.516$

$$Gain_{\text{Location}}(D) = 0.9544 - 0.516 = 0.438$$

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: $\{\text{Bedrooms}, \text{Basement}, \text{Floorplan}\}$.

Note: This dataset is **homogenous**: no houses from D_1 were visited.

$Class(D_1) = \text{No}$.

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

$enthropy(D_2) = 0.918$

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

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

$Pr_{D_{22}}(\text{Visited} = \text{Yes}) = \frac{4}{4}$

$enthropy(D_{21}) = -\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5} = 0.5743$

$enthropy(D_{22}) = -\frac{4}{4}\log_2\frac{4}{4} - \frac{0}{4}\log_2\frac{0}{4} = 0$

$enthropy_{\text{Bedrooms}}(D_2) = \frac{5}{9} \cdot 0.5743 + \frac{4}{9} \cdot 0 = 0.319$

$$Gain_{\text{Bedrooms}}(D) = 0.918 - 0.319 = 0.599$$

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

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

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

$enthropy(D_{21}) = -\frac{4}{5}\log_2\frac{4}{5} - \frac{1}{5}\log_2\frac{1}{5} = 0.7219$

$enthropy(D_{22}) = -\frac{2}{4}\log_2\frac{2}{4} - \frac{2}{4}\log_2\frac{2}{4} = 1$

$enthropy_{\text{Basement}}(D) = \frac{5}{9} \cdot 0.7219 + \frac{4}{9} \cdot 1 = 0.8455$

$$Gain_{\text{Basement}}(D) = 0.918 - 0.8455 = 0.0725$$

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

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

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

$enthropy(D_{21}) = -\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5} = 0.9709$

$enthropy(D_{22}) = -\frac{3}{4}\log_2\frac{3}{4} - \frac{1}{4}\log_2\frac{1}{4} = 0.8112$

$enthropy_{\text{Basement}}(D) = \frac{5}{9} \cdot 0.9709 + \frac{4}{9} \cdot 0.8112 = 0.8999$

$$Gain_{\text{Basement}}(D) = 0.918 - 0.8999 = 0.0181$$

Step 3 result: Bedrooms yields the best Information Gain.

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

$enthropy(D_{21}) = 0.7219$

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

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

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

$enthropy(D_{211}) = -\frac{2}{3}\log_2\frac{2}{3} - \frac{1}{3}\log_2\frac{1}{3} = 0.918$

$enthropy(D_{212}) = -\frac{0}{2}\log_2\frac{0}{2} - \frac{2}{2}\log_2\frac{2}{2} = 0$

$enthropy_{\text{Basement}}(D) = \frac{3}{5} \cdot 0.918 + \frac{2}{5} \cdot 0 = 0.5509$

$$Gain_{\mathsf{Basement}}(D) = 0.7219 - 0.5509 = 0.1709$$

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

$Pr_{D_{211}}(\mathsf{Visited} = \mathsf{Yes}) = \frac{2}{4}$
$Pr_{D_{221}}(\mathsf{Visited} = \mathsf{Yes}) = \frac{0}{1} = 0$
$enthropy(D_{211}) = -\frac{2}{4}\log_2\frac{2}{4} - \frac{2}{4}\log_2\frac{2}{4} = 1$
$enthropy(D_{212}) = -\frac{0}{1}\log_2\frac{0}{1} - \frac{0}{1}\log_2\frac{0}{1} = 0$
$enthropy_{\mathsf{Basement}}(D) = \frac{4}{5}\cdot 1 + \frac{1}{5}\cdot 0 = 0.8$

$$Gain_{\mathsf{Basement}}(D) = 0.7219 - 0.8 = -0.781$$

Step 4 result: Basement yields the best Information Gain

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

$enthropy(D_{211}) = 0.918$

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

$Pr_{D_{2111}}(\mathsf{Visited} = \mathsf{Yes}) = \frac{2}{2} = 1$
$Pr_{D_{2112}}(\mathsf{Visited} = \mathsf{Yes}) = \frac{0}{1} = 0$
$enthropy(D_{2111}) = -\frac{2}{2}\log_2\frac{2}{2} - \frac{0}{2}\log_2\frac{0}{2} = 0$
$enthropy(D_{221}) = -\frac{0}{1}\log_2\frac{0}{1} - \frac{0}{1}\log_2\frac{0}{1} = 0$
$enthropy_{\mathsf{Basement}}(D) = \frac{2}{3}\cdot 0 + \frac{2}{3}\cdot 0 = 0$

$$Gain_{\mathsf{Basement}}(D) = 0.918 - 0 = -0.918$$

Step 5: result Floorplan achieves significant Information gain.

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

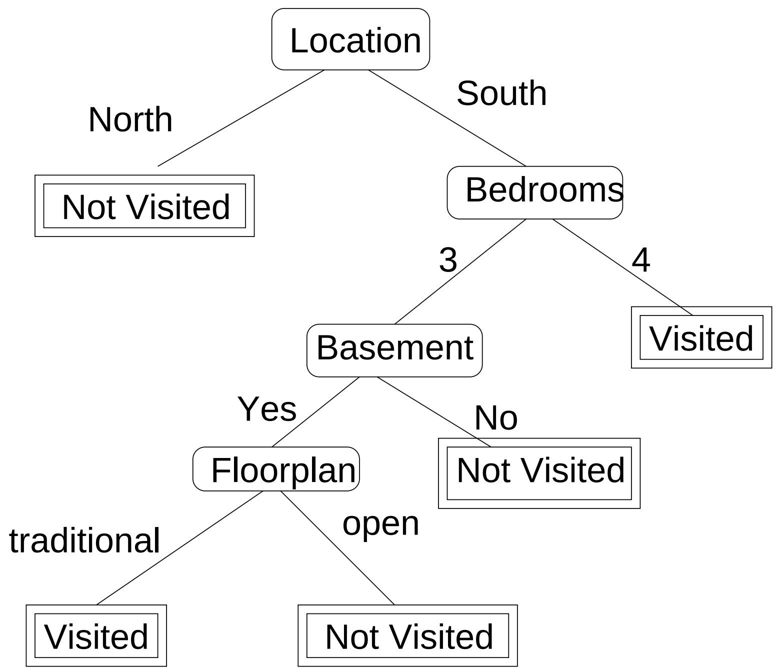
$Class(D_{2111}) = \mathsf{Yes}$.
$Class(D_{2112}) = \mathsf{No}$.

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

$Pr_{D_{22}}(\mathsf{Visited} = \mathsf{Yes}) = 1$, therefore:

$Class(D_{22}) = \mathsf{Yes}$.

Resulting Tree: Resulting tree is depicted below:


Location
North
South
Not Visited
Bedrooms
3
4
Basement
Visited
Yes
No
Floorplan
Not Visited
traditional
open
Visited
Not Visited