Data Mining: Classification: Part 4

Decision Tree Examples

CS 4821 - CS 5831

Some slides adapted from P. Smyth; A. Moore, D. Klein Han, Kamber, Pei; Tan, Steinbach, Kumar; L. Kaebling; R. Tibshirani; T. Taylor; and L. Hannah

Example: Data

Consider a data set on playing tennis:

Outlook	Temp	Humidity	Windy	Play
Sunny	Hot	High	False	No
Sunny	Hot	Low	True	Yes
Sunny	Mild	Low	True	Yes
Rainy	Mild	High	False	Yes
Overcast	Cool	High	False	Yes
Rainy	Cool	High	True	No
Overcast	Cool	Low	True	No
Sunny	Cool	High	True	Yes
Rainy	Mild	High	True	No
Rainy	Cool	Low	False	Yes
Overcast	Hot	Low	True	Yes
Sunny	Cool	Low	False	Yes
Overcast	Mild	High	False	Yes

Use $Outlook$, $Temp$, $Humidity$ and $Windy$ to predict $Play$

What is the Best Root Feature?

Outlook

Sunny	Rainy	Overcast
No	Yes	Yes
Yes	No	No
Yes	No	Yes
Yes	Yes	Yes
Yes		

Temp

Hot	Mild	Cool
No	Yes	Yes
Yes	Yes	No
Yes	No	No
	Yes	Yes
		Yes
		Yes

Humidity

Low	High
Yes	No
Yes	Yes
No	Yes
Yes	No
Yes	Yes
Yes	No
	Yes

Windy

False	True
No	Yes
Yes	Yes
Yes	No
Yes	No
Yes	Yes
Yes	No
	Yes

Example: Humidity

Humidity

Low	High
Yes	No
Yes	Yes
No	Yes
Yes	No
Yes	Yes
Yes	No
	Yes

Parent node: $p_{Hum,N} = \frac{4}{13}$, $p_{Hum,Y} = \frac{9}{13}$

Entropy:

$$H(X_{Hum}) = - \sum_{i=\{N,Y\}} p_{Hum,i} \log p_{Hum,i}$$

$$=$$

Gain in Entropy:

$$Gain(X_{Hum})_H = H(X_{Hum}) - \left(\sum_{l=\{L,H\}} \frac{n_l}{n_{Hum}} H(X_{Hum_l}) \right)$$

$$=$$

Example: Humidity

Humidity

Low	High
Yes	No
Yes	Yes
No	Yes
Yes	No
Yes	Yes
Yes	No
	Yes

Parent node: $p_{Hum,N} = \frac{4}{13}$, $p_{Hum,Y} = \frac{9}{13}$

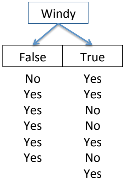
Entropy:

$$
\begin{aligned}
H(X_{Hum}) &= -\sum_{i=\{N,Y\}} p_{Hum,i} \log p_{Hum,i} \\
&= -\frac{4}{13}\log\frac{4}{13} - \frac{9}{13}\log\frac{9}{13} \\
&= 0.8905
\end{aligned}
$$

Gain in Entropy:

$$
\begin{aligned}
Gain(X_{Hum})_H &= H(X_{Hum}) - \left(\sum_{l=\{L,H\}} \frac{n_l}{n_{Hum}} H(X_{Hum_l})\right) \\
&= H(X_{Hum}) - \frac{6}{13}\left(-\frac{1}{6}\log\frac{1}{6} - \frac{5}{6}\log\frac{5}{6}\right) \\
&\quad - \frac{7}{13}\left(-\frac{3}{7}\log\frac{3}{7} - \frac{4}{7}\log\frac{4}{7}\right) \\
&= 0.8905 - 0.8305 = 0.06
\end{aligned}
$$

Example: Windy



Parent node: $p_{W,N} = \frac{4}{13}$, $p_{W,Y} = \frac{9}{13}$

Entropy:

$$H(X_W) = - \sum_{i=\{N,Y\}} p_{W,i} \log p_{W,i}$$

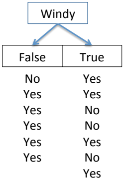
$$=$$

Gain in Entropy:

$$Gain(X_W)_H = H(X_W) - \left(\sum_{l=\{F,T\}} \frac{n_l}{n_W} H(X_{W_l}) \right)$$

$$=$$

Example: Windy



Parent node: $p_{W,N} = \frac{4}{13}$, $p_{W,Y} = \frac{9}{13}$

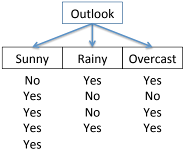
Entropy:

$$
\begin{aligned}
H(X_W) &= -\sum_{i=\{N,Y\}} p_{W,i} \log p_{W,i} \\
&= -\frac{4}{13}\log\frac{4}{13} - \frac{9}{13}\log\frac{9}{13} \\
&= 0.8905
\end{aligned}
$$

Gain in Entropy:

$$
\begin{aligned}
Gain(X_W)_H &= H(X_W) - \left(\sum_{l=\{F,T\}} \frac{n_l}{n_W} H(X_{W_l})\right) \\
&= H(X_W) - \frac{6}{13}\left(-\frac{1}{6}\log\frac{1}{6} - \frac{5}{6}\log\frac{5}{6}\right) \\
&\quad - \frac{7}{13}\left(-\frac{3}{7}\log\frac{3}{7} - \frac{4}{7}\log\frac{4}{7}\right) \\
&= 0.8905 - 0.8305 = 0.06
\end{aligned}
$$

Example: Outlook



Parent node: $p_{Out,N} = \frac{4}{13}$, $p_{Out,Y} = \frac{9}{13}$

Entropy:

$$H(X_{Out}) = - \sum_{i=\{N,Y\}} p_{Out,i} \log p_{Out,i}$$

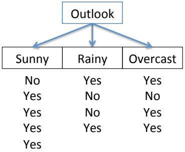
$$=$$

Gain in Entropy:

$$Gain(X_{Out})_H = H(X_{Out}) - \left(\sum_{l=\{S,R,O\}} \frac{n_l}{n_{Out}} H(X_{Out_l}) \right)$$

$$=$$

Example: Outlook



Parent node: $p_{Out,N} = \frac{4}{13}$, $p_{Out,Y} = \frac{9}{13}$

Entropy:

$$
\begin{aligned}
H(X_{Out}) &= - \sum_{i=\{N,Y\}} p_{Out,i} \log p_{Out,i} \\
&= -\frac{4}{13} \log \frac{4}{13} - \frac{9}{13} \log \frac{9}{13} \\
&= 0.8905
\end{aligned}
$$

Gain in Entropy:

$$
Gain(X_{Out})_H = H(X_{Out}) - \left(\sum_{l=\{S,R,O\}} \frac{n_l}{n_{Out}} H(X_{Out_l}) \right)
$$

$$
\begin{aligned}
&= H(X_{Out}) - \frac{4}{13}\left(-\frac{1}{4}\log\frac{1}{4} - \frac{3}{4}\log\frac{3}{4}\right) \\
&\qquad - \frac{4}{13}\left(-\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2}\right) - \frac{5}{13}\left(-\frac{1}{5}\log\frac{1}{5} - \frac{4}{5}\log\frac{4}{5}\right) \\
&= 0.8905 - 0.8345 = 0.056
\end{aligned}
$$

Example: Temperature

Temp

Hot	Mild	Cool
No	Yes	Yes
Yes	Yes	No
Yes	No	No
	Yes	Yes
		Yes
		Yes

Parent node: $p_{T,N} = \frac{4}{13}$, $p_{T,Y} = \frac{9}{13}$

Entropy:

$$H(X_T) = - \sum_{i=\{N,Y\}} p_{T,i} \log p_{T,i}$$

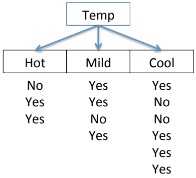
$$=$$

Gain in Entropy:

$$Gain(X_T)_H = H(X_T) - \left(\sum_{l=\{H,M,C\}} \frac{n_l}{n_T} H(X_{T_l}) \right)$$

$$=$$

Example: Temperature



Parent node: $p_{T,N} = \frac{4}{13}$, $p_{T,Y} = \frac{9}{13}$

Entropy:

$$
\begin{aligned}
H(X_T) &= -\sum_{i=\{N,Y\}} p_{T,i} \log p_{T,i} \\
&= -\frac{4}{13} \log \frac{4}{13} - \frac{9}{13} \log \frac{9}{13} \\
&= 0.8905
\end{aligned}
$$

Gain in Entropy:

$$
Gain(X_T)_H = H(X_T) - \left(\sum_{l=\{H,M,C\}} \frac{n_l}{n_T} H(X_{T_l}) \right)
$$

$$
\begin{aligned}
&= H(X_T) - \frac{3}{13}\left(-\frac{1}{3}\log\frac{1}{3} - \frac{2}{3}\log\frac{2}{3}\right) \\
&\qquad - \frac{4}{13}\left(-\frac{1}{4}\log\frac{1}{4} - \frac{3}{4}\log\frac{3}{4}\right) - \frac{6}{13}\left(-\frac{1}{3}\log\frac{1}{3} - \frac{2}{3}\log\frac{2}{3}\right) \\
&= 0.8905 - 0.8854 = 0.051
\end{aligned}
$$

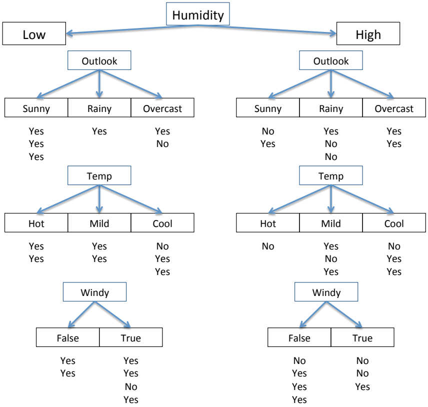
ID3 Example: Best Feature

Compare the Gain in Entropy:

Variable	GainEntropy
Outlook	0.056
Temp	0.051
Humidity	0.060
Windy	0.060

Select the **maximum** - Tie between Humidity and Windy

Example - Next Feature?

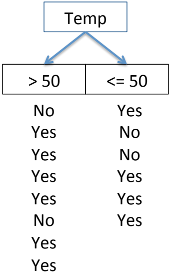


- For the `low humidity` side, compute conditional entropy
- For the `high humidity` side, compute conditional entropy
- Treat each side separately

Example 2: Numeric Data

Outlook	Temp2	Humidity	Windy	Play
Sunny	97	High	False	No
Sunny	85	Low	True	Yes
Sunny	71	Low	True	Yes
Rainy	75	High	False	Yes
Overcast	56	High	False	Yes
Rainy	42	High	True	No
Overcast	34	Low	True	No
Sunny	44	High	True	Yes
Rainy	64	High	True	No
Rainy	49	Low	False	Yes
Overcast	88	Low	True	Yes
Sunny	47	Low	False	Yes
Overcast	69	High	False	Yes

Example 2: Numeric Data



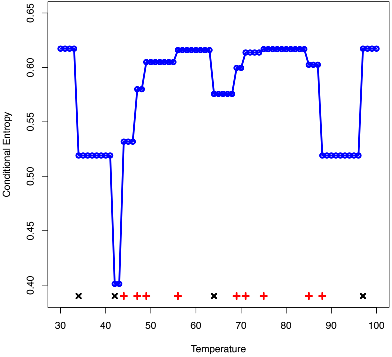
Fix a value for x', say 50:

- find which records have values > 50 and ≤ 50
- compute gain in entropy

Example 2 - Choose best split value

How do we find the best x'?

- calculate conditional entropy for all values in range of X_j
- only need to search over "seen" values (in data)



- Conditional entropy is minimized when $42 \leq x' \leq 43$

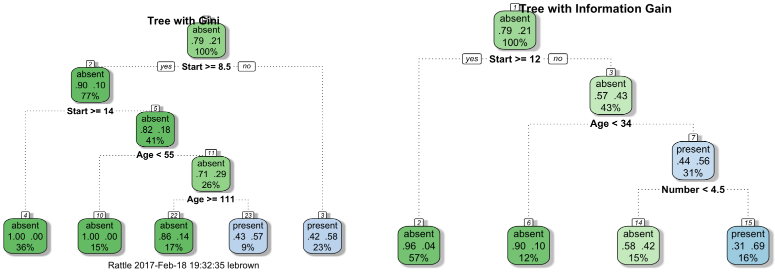
Example 3: Kyphosis

Data set to predict kyphosis (type of deformation) using:

- Age, number, start

```r
library(rpart)
# using Gini to split
m1 <- rpart(Kyphosis ~ Age + Number + Start,
    data=kyphosis)
# using information gain to split
m2 <- rpart(Kyphosis ~ Age + Number + Start,
    data=kyphosis,
             parms=list(split='information'))
```

Example 3: Kyphosis



Example 4: Decision Tree

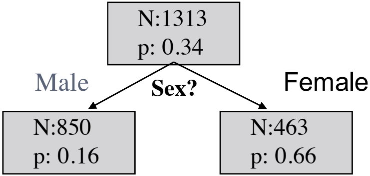
Titanic Data Set

- 1313 passengers
- 34% survived
- was survival random? or did it depend on feature of the individual?
  - gender
  - age
  - class of ticket

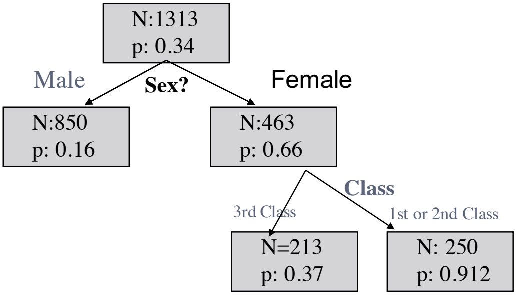
Example 4: Decision Tree

N:1313
p: 0.34

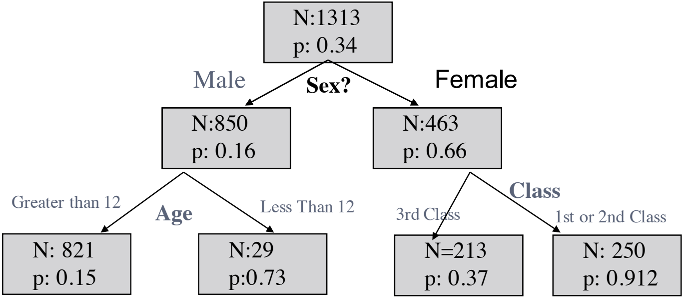
Example 4: Decision Tree



Example 4: Decision Tree



Example 4: Decision Tree



Example 4: Decision Tree