Decision tree

ISL / 강한솔

Index

- ✓ Decision Tree
- ✓ ID3
- ✓ C4.5
- ✓ Experiment

01 Decision Tree

❖ Machine Learning

✓ Supervised learning method
ex) Decision Tree, Neural network, etc.

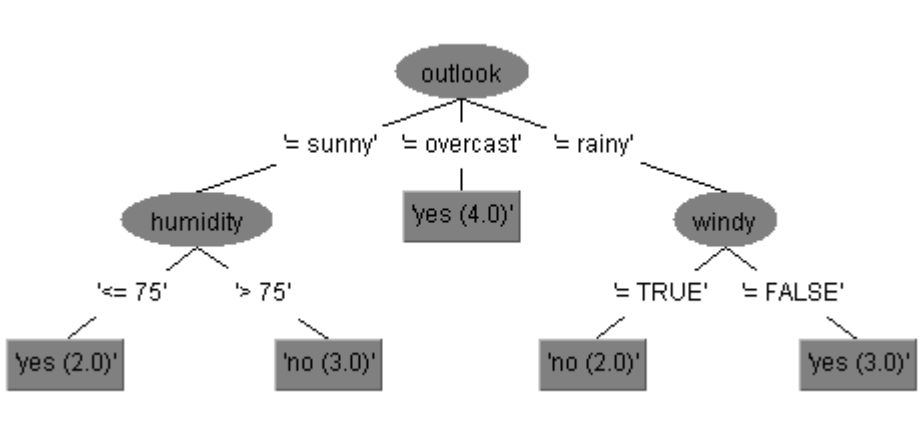
✓ Unsupervised learning method
ex) K-means, Clustering, etc.

알고리즘	평가지수	비고
ID3	Entropy	다지분리(nominal)
C4.5, C5.0	Information Gain	다지분리(nominal) 및 이진분리(numeric)
CHAID	카이제곱(nominal), F검정(numeric)	통계적 접근방식
CART	Gini index(nominal), 분산의 차이 (numeric)	통계적 접근방식, 항상 2진 분리

01 Decision Tree

ex) weather

outlook	temperature	humidity	windy	play
sunny	85	85	false	no
sunny	80	90	true	no
overcast	83	83	false	yes
rain	70	96	false	yes
rain	68	80	false	yes
rain	65	70	true	no
overcast	64	64	true	yes
sunny	72	95	false	no
sunny	69	70	false	yes
rain	75	80	false	yes
sunny	75	70	true	yes
overcast	72	90	true	yes
overcast	81	75	false	yes
rain	71	91	true	no

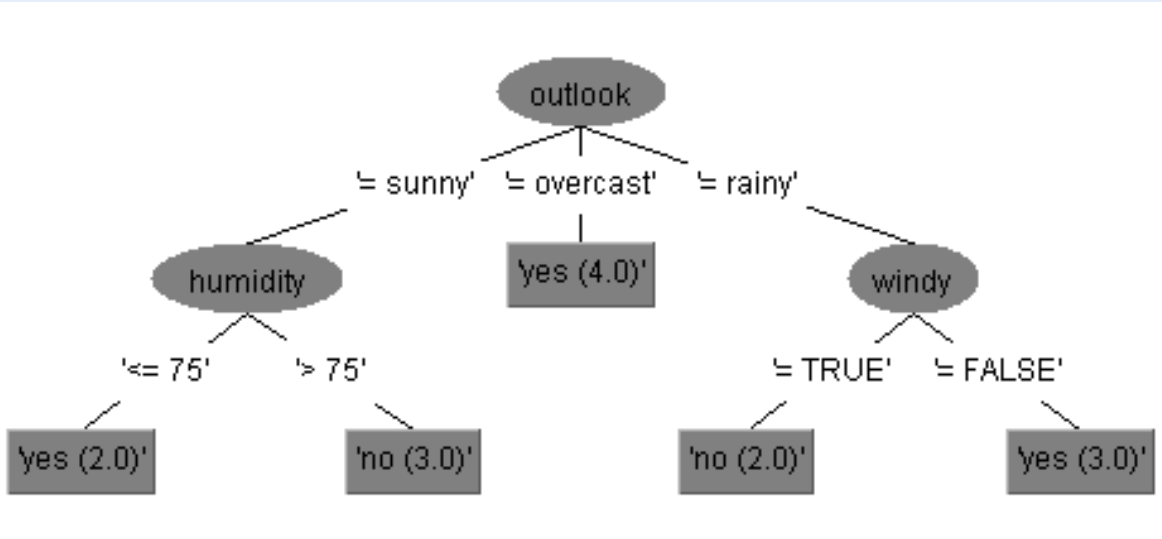


02 ID3

❖ Entropy ➡ 주어진 데이터 집합의 혼잡도

서로 다른 클래스 多 → Entropy ↑

서로 다른 클래스 小 → Entropy ↓



02 ID3

❖ Entropy of the set S.

$$Entropy(S) = -\sum_{i=1}^{m} p_i \log_2(p_i)$$

$$p_i = \frac{freq(C_i, S)}{|S|}$$

S : 주어진 데이터들의 집합

$C = \{C_1, C_2, \ldots, C_k\}$: 클래스 값들의 집합

$freq(C_i, S)$: S에서 class C_i에 속하는 레코드의 수

$|S|$: 주어진 데이터들의 집합 데이터 개수

02 ID3

❖ ex) weather

play
no
no
yes
yes
yes
no
yes
no
yes
yes
yes
yes
yes
no

$$Entropy(S) = -\sum_{i=1}^{m} p_i \log_2(p_i) \quad p_i = \frac{freq(C_i, S)}{|S|}$$

$$freq(C_1, S) = 9, \quad freq(C_2, S) = 5$$

$$|S| = 14$$

$$Entropy(9,5) = -\frac{9}{14}\log_2(\frac{9}{14}) - \frac{5}{14}\log_2(\frac{5}{14}) = 0.9402$$

02 ID3

❖ Information Gain ➡ 어떤 속성이 데이터를 더 잘 구분하는지 나타내는 지표

$$Entropy_X(T) = \sum_{i=1}^{n} \frac{|T_i|}{|T|} \times Entorpy(T_i)$$

$$gain(X) = Entropy(T) - Entropy_X(T)$$

02 ID3

❖ ex) weather (outlook, windy, play)

outlook	windy	play
sunny	false	no
sunny	true	no
overcast	false	yes
rain	false	yes
rain	false	yes
rain	true	no
overcast	true	yes
sunny	false	no
sunny	false	yes
rain	false	yes
sunny	true	yes
overcast	true	yes
overcast	false	yes
rain	true	no

$$Entropy(T) = -\frac{9}{14}\log_2(\frac{9}{14}) - \frac{5}{14}\log_2(\frac{5}{14}) = 0.9402$$

$$Entropy_o(T) = \frac{5}{14}\times(-\frac{2}{5}\log_2(\frac{2}{5}) - \frac{3}{5}\log_2(\frac{3}{5}))$$
$$+\frac{4}{14}\times(-\frac{4}{4}\log_2(\frac{4}{4}) - \frac{0}{4}\log_2(\frac{0}{4}))$$
$$+\frac{5}{14}\times(-\frac{3}{5}\log_2(\frac{3}{5}) - \frac{2}{5}\log_2(\frac{2}{5}))$$
$$= 0.694$$

$$Gain(outlook) = 0.940 - 0.694 = 0.246$$

02 ID3

❖ ex) weather (outlook, windy, play)

outlook	windy	play
sunny	false	no
sunny	true	no
overcast	false	yes
rain	false	yes
rain	false	yes
rain	true	no
overcast	true	yes
sunny	false	no
sunny	false	yes
rain	false	yes
sunny	true	yes
overcast	true	yes
overcast	false	yes
rain	true	no

$$Entropy(T) = -\frac{9}{14}\log_2(\frac{9}{14}) - \frac{5}{14}\log_2(\frac{5}{14}) = 0.9402$$

$$Entropy_w(T) = \frac{6}{14}\times(-\frac{3}{6}\log_2(\frac{3}{6}) - \frac{3}{6}\log_2(\frac{3}{6})) + \frac{8}{14}\times(-\frac{6}{8}\log_2(\frac{6}{8}) - \frac{2}{8}\log_2(\frac{2}{8})) = 0.892$$

$$Gain(windy) = 0.940 - 0.892 = 0.048$$

02 C4.5

❖ Problem of ID3's algorithm

1) 너무 잘게 분할하는 경우가 발생한다. (1개로 분류되는 경우)

2) 수치형 속성(continuous attribute)을 다루지 못함.

❖ Gain ratio

$$split\ info(X) = -\sum_{i=1}^{n} \frac{|T_i|}{|T|} \times \log_2\left(\frac{|T_i|}{|T|}\right) \qquad gain\ ratio(X) = \frac{gain(X)}{split\ info(X)}$$

❖ Continuous attributes

$$A = \{v_1, v_2, \ldots, v_m\} \quad between\ v_i\ and\ v_{i+1}$$

$$\{v_1, v_2, \ldots, v_i\} \quad \{v_{i+1}, v_{i+2}, \ldots, v_m\}$$

$$midpoint = \frac{v_i + v_{i+1}}{2}$$

02 C4.5

❖ ex) weather (temperature, play)

temperature	85	80	83	70	68	65	64	72	69	75	75	72	81	71
play	no	no	yes	yes	yes	no	yes	no	yes	yes	yes	yes	yes	no

temperature	64	65	68	69	70	71	72	75	80	81	83	85
play	yes	no	yes	yes	yes	no	no yes	yes yes	no	yes	yes	no

break points

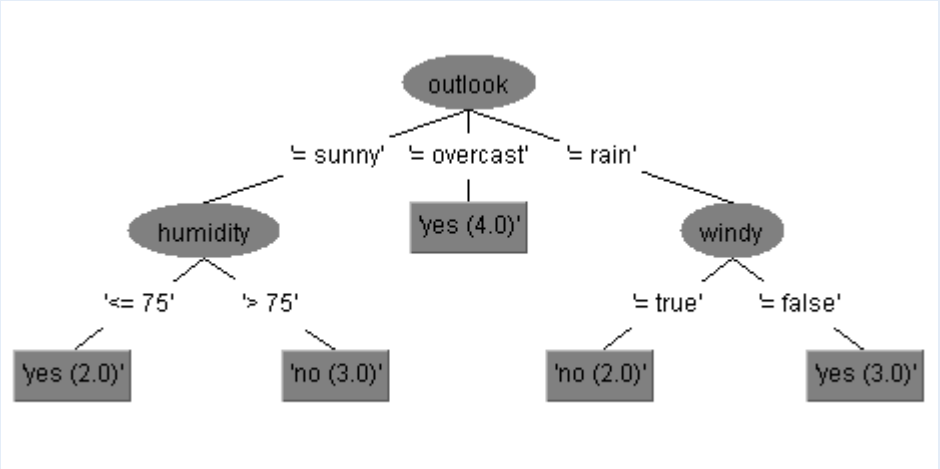
ex) *break point*: 71.5

$$Entropy(T) = -\frac{9}{14}\log_2(\frac{9}{14}) - \frac{5}{14}\log_2(\frac{5}{14}) = 0.940$$

$$Entropy_{71.5}(T) = \frac{6}{14}(-\frac{4}{6}\log_2\frac{4}{6} - \frac{2}{6}\log_2\frac{2}{6}) + \frac{8}{14}(-\frac{5}{8}\log_2\frac{5}{8} - \frac{3}{8}\log_2\frac{3}{8}) = 0.939$$

03 Experiment

@relation weather
@attribute outlook {sunny, overcast, rain}
@attribute temperature numeric
@attribute humidity numeric
@attribute windy {true, false}
@attribute play {yes, no}
@data
sunny, 85, 85, false, no
sunny, 80, 90, true, no
overcast, 83, 78, false, yes
rain, 70, 96, false, yes
rain, 68, 80, false, yes
rain, 65, 70, true, no
overcast, 64, 65, true, yes
sunny, 72, 95, false, no
sunny, 69, 70, false, yes
rain, 75, 80, false, yes
sunny, 75, 70, true, yes
overcast, 72, 90, true, yes
overcast, 81, 75, false, yes
rain, 71, 80, true, no



03 Experiment

```
C4.5 [release 8] decision tree generator        Sun Oct 19 15:08:43 2014
----------------------------------------

    Options:
        File stem <golf>

Read 14 cases (4 attributes) from golf.data

Decision Tree:

outlook = overcast: Play (4.0)
outlook = sunny:
|   humidity <= 75 : Play (2.0)
|   humidity > 75 : Don't Play (3.0)
outlook = rain:
|   windy = true: Don't Play (2.0)
|   windy = false: Play (3.0)


Tree saved


Evaluation on training data (14 items):

         Before Pruning           After Pruning
         ----------------   ---------------------------
         Size      Errors   Size      Errors    Estimate

            8    0(  0.0%)      8    0(  0.0%)    (38.5%)   <<
```

```
=== Run information ===

Scheme:weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     weather
Instances:    14
Attributes:   5
              outlook
              temperature
              humidity
              windy
              play
Test mode:10-fold cross-validation

=== Classifier model (full training set) ===

J48 pruned tree
------------------

outlook = sunny
|   humidity <= 75: yes (2.0)
|   humidity > 75: no (3.0)
outlook = overcast: yes (4.0)
outlook = rain
|   windy = true: no (2.0)
|   windy = false: yes (3.0)

Number of Leaves  :     5

Size of the tree :      8
```

Q & A