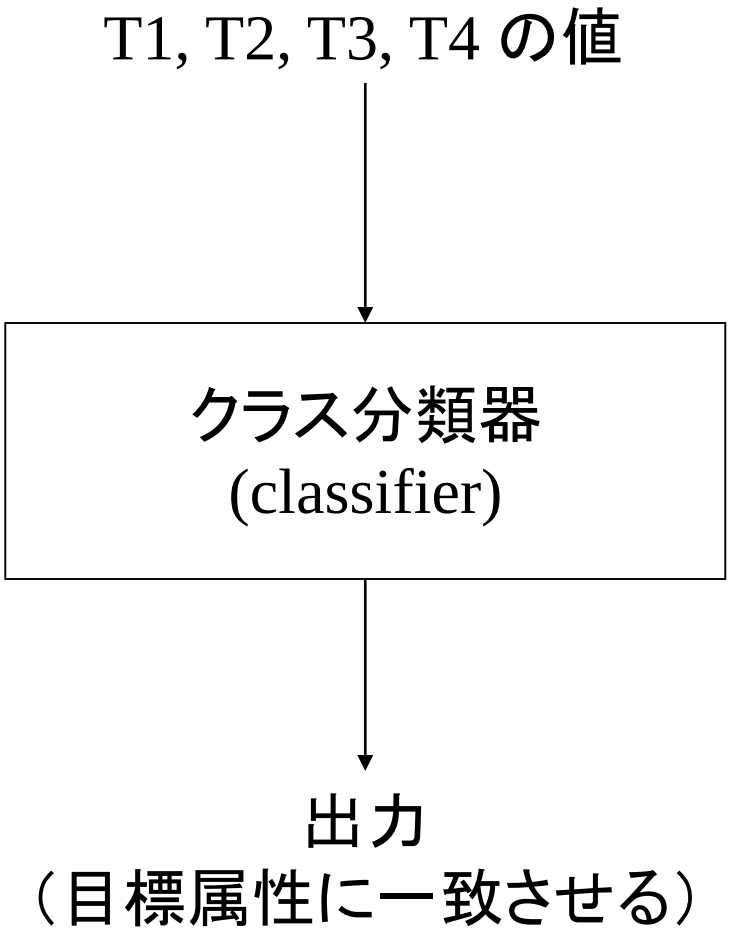
様々な決定木とその構成方法

和田　俊和

T1	T2	T3	T4	目標属性
1	0	1	1	1
1	0	1	1	1
1	1	1	1	1
1	1	1	0	0
1	0	1	0	0
1	1	0	1	0
1	0	0	1	0
1	1	0	1	0
0	1	0	1	1
0	0	1	1	1
0	1	0	1	1
0	1	0	1	1
0	0	0	1	1
0	0	1	0	0
0	1	0	0	0
0	0	1	0	0



分類ルール

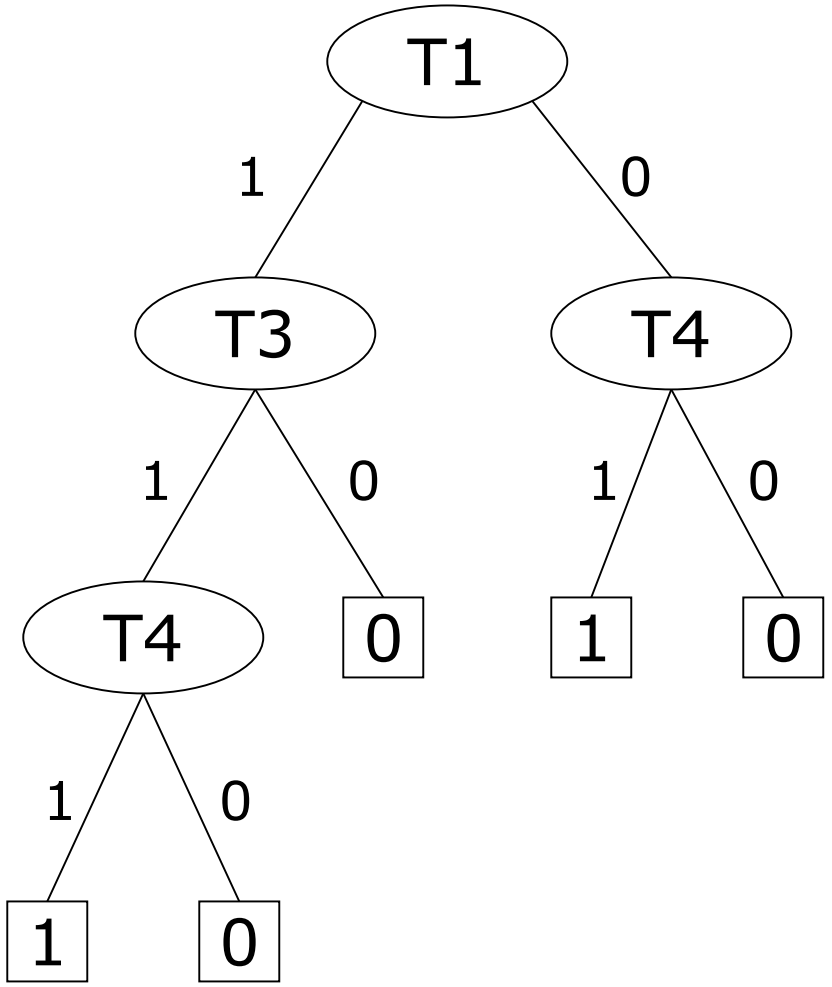
テスト

T1	T2	T3	T4	目標属性 (Objective)
1	0	1	1	1
1	0	1	1	1
1	1	1	1	1
1	1	1	0	0
1	0	1	0	0
1	1	0	1	0
1	0	0	1	0
1	1	0	1	0
0	1	0	1	1
0	0	1	1	1
0	1	0	1	1
0	1	0	1	1
0	0	0	1	1
0	0	1	0	0
0	1	0	0	0
0	0	1	0	0

```
If T1=1
Then
        If T3=1
        Then
                If T4=1
                Then  目標属性＝1
                Else  目標属性＝0
        Else
                目標属性＝0
Else
        If T4=1
        Then  目標属性＝1
        Else  目標属性＝0
```

決定木(decision tree)

```
If T1＝1
Then
        If T3＝1
        Then
                If T4＝1
                Then  目標属性＝1
                Else   目標属性＝0
        Else
                目標属性＝0
Else
        If T4＝1
        Then  目標属性＝1
        Else   目標属性＝0
```

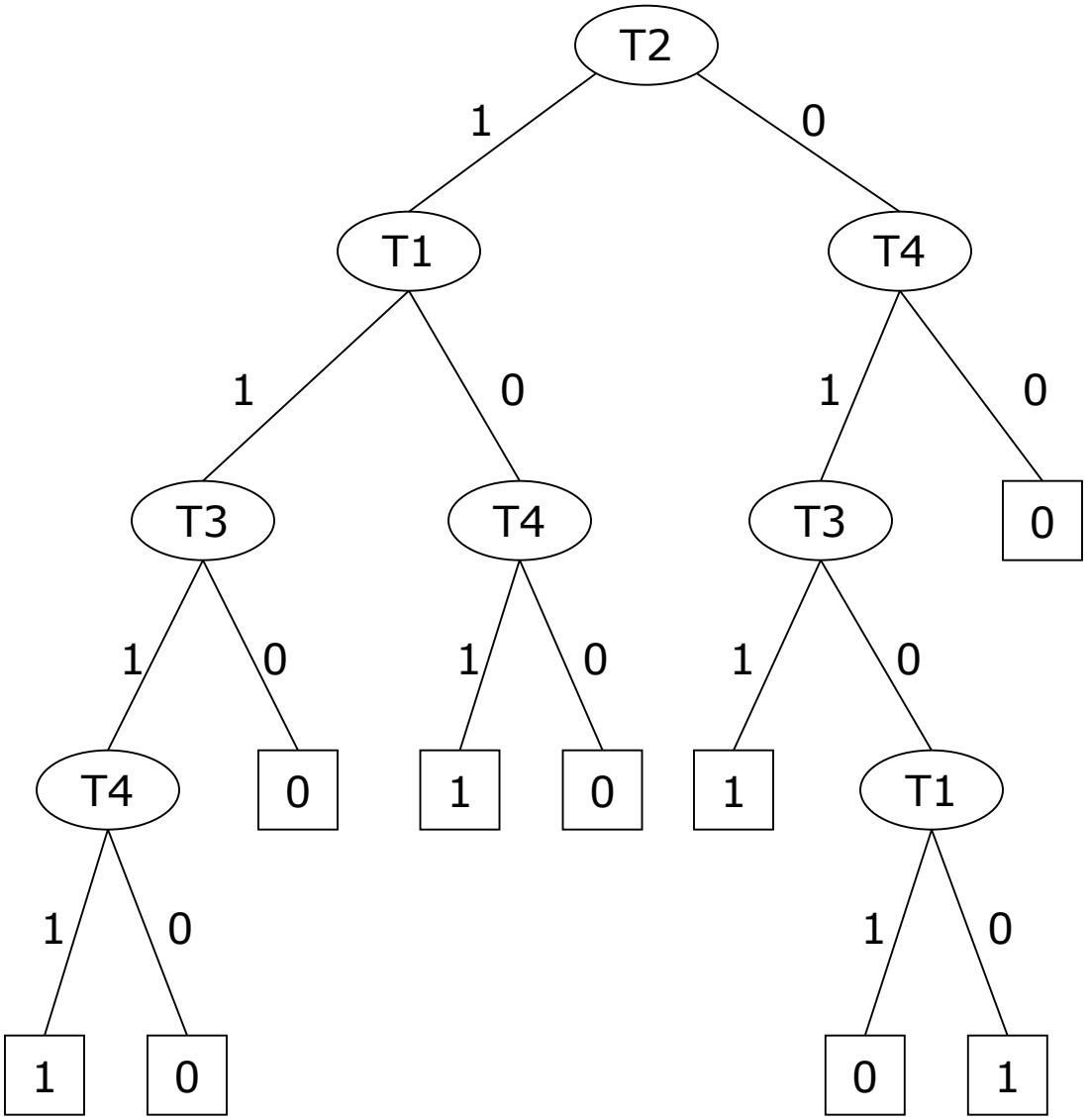


ルールによる表現

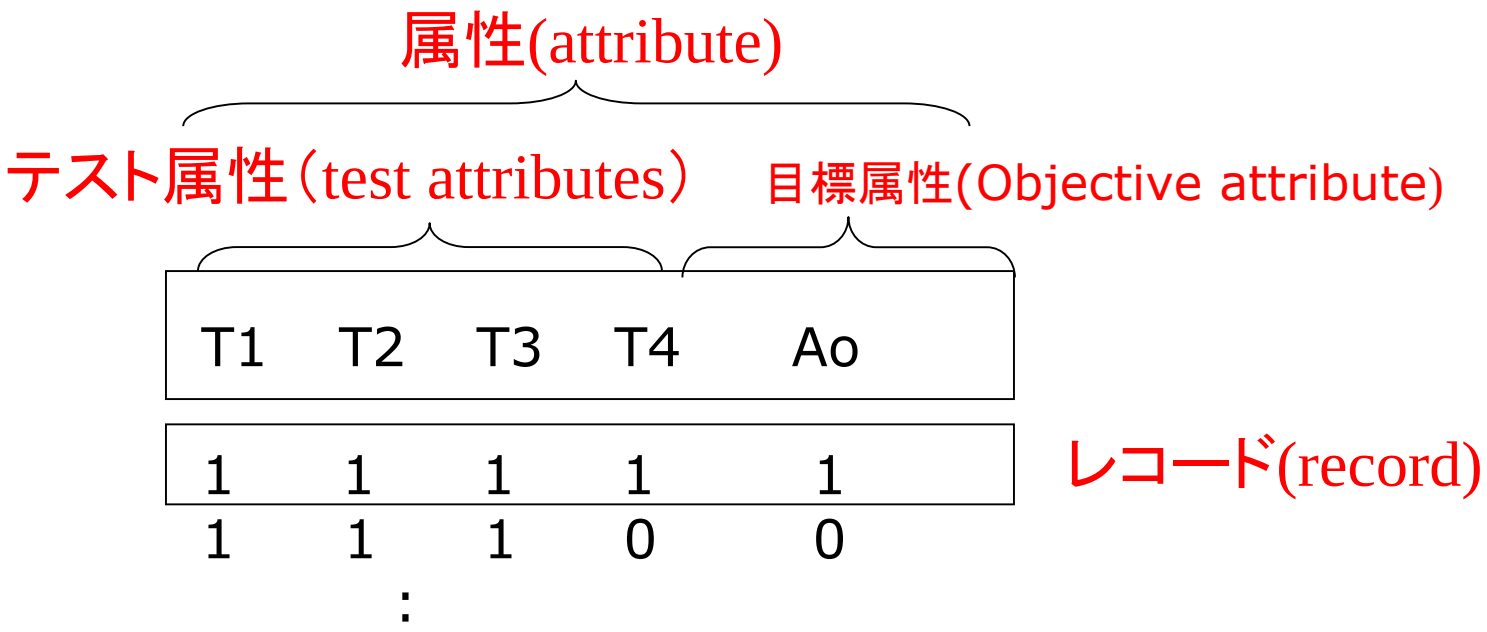
木構造表現

テストの順番を変えると異なる決定木が得られる

T1	T2	T3	T4	目標属性
1	1	1	1	1
1	1	1	0	0
1	1	0	1	0
1	1	0	1	0
0	1	0	1	1
0	1	0	1	1
0	1	0	1	1
0	1	0	0	0
1	0	1	1	1
1	0	1	1	1
0	0	1	1	1
0	0	0	1	1
1	0	0	1	0
0	0	1	0	0
1	0	1	0	0
0	0	1	0	0



用語



レコード t の属性 A の値を t[A] と表記

目標値　目標属性の値の1つ.　例えば 1

目標属性 Ao、目標値が 1 のとき、

- t[Ao]= 1 となる t を　正(positive)
- t[Ao]≠1となる t を　負(negative)

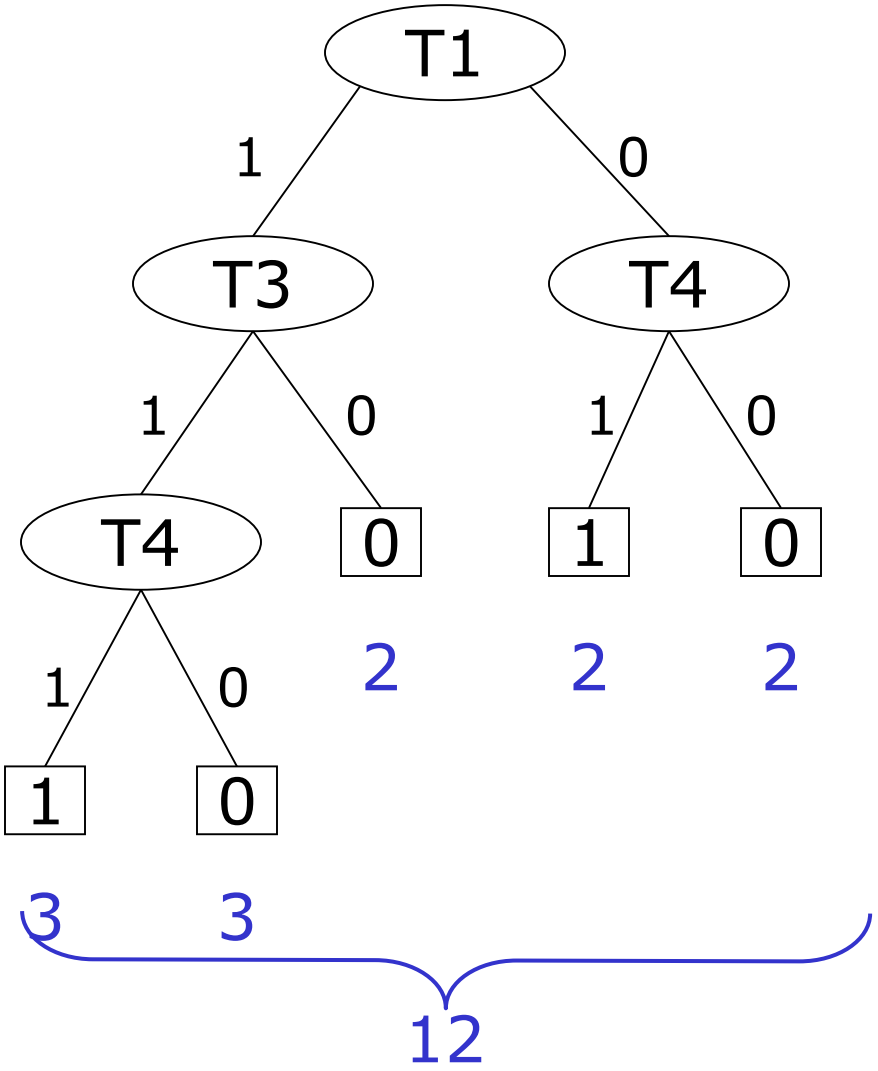
決定木の最小化

決定木のコスト

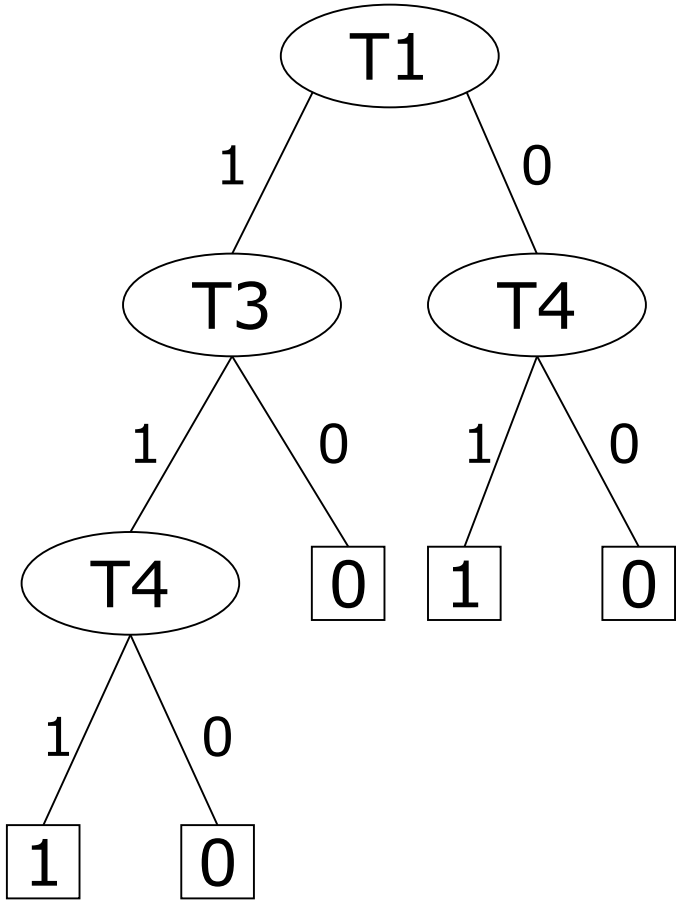
決定木の評価
- 根ノードから葉ノード x に至るテストの数 cost(x)

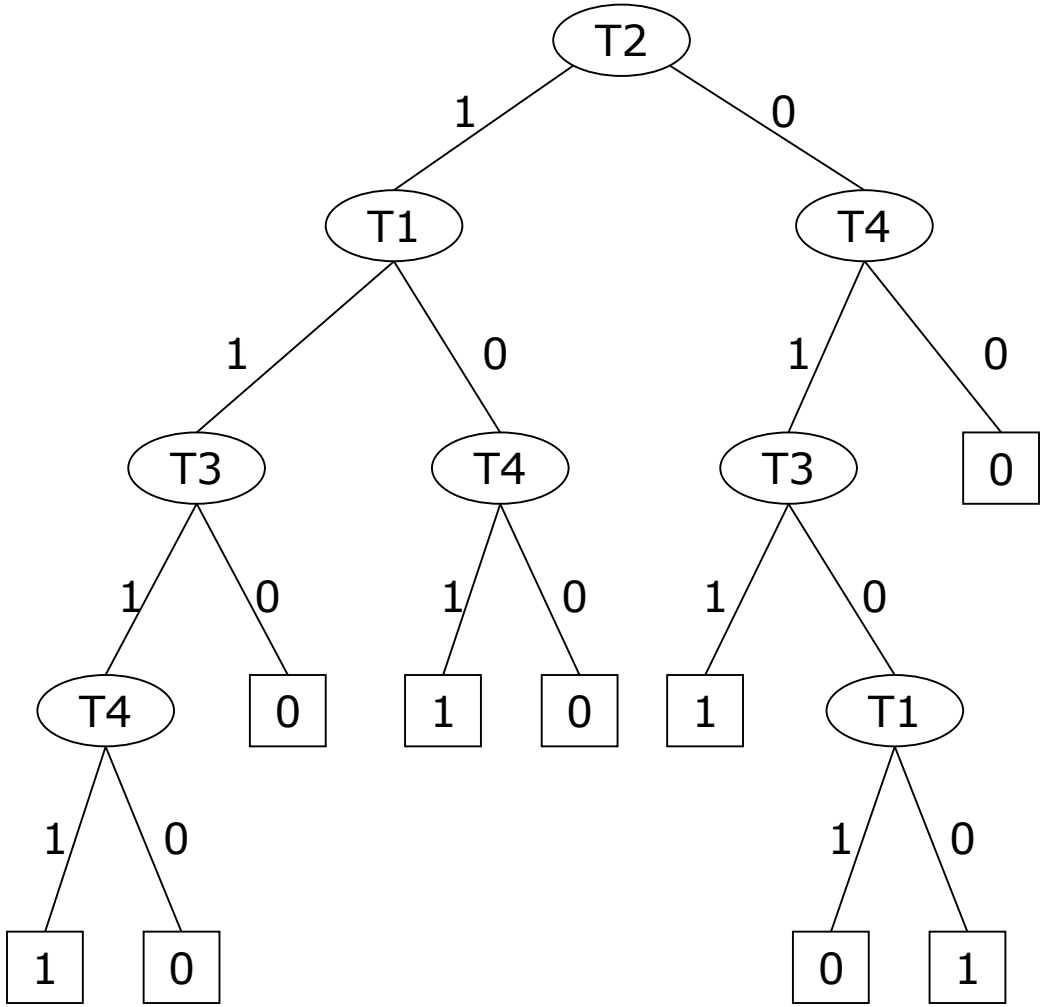
- 決定木のコスト
＝ $\sum$ {cost(x) | x は葉ノード}

- 決定木のコストは小さいほうが望ましい



コストの比較





3+3+2+2+2=12

4+4+3+3+3+3+4+4+2=30

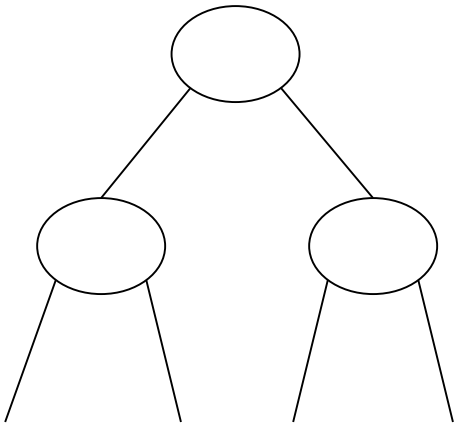
近似的解法

Greedy Algorithm

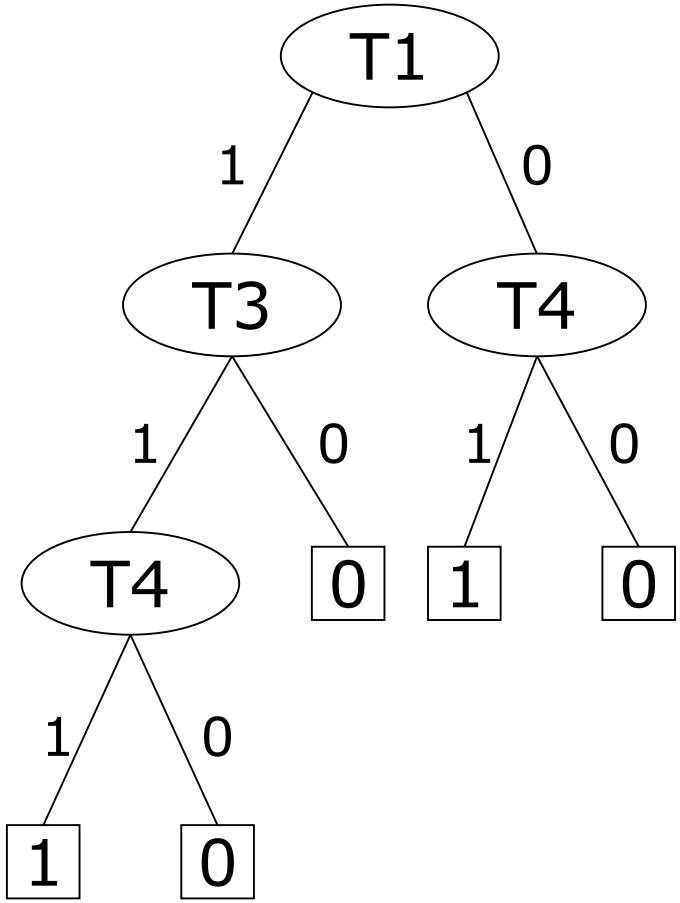
エントロピー

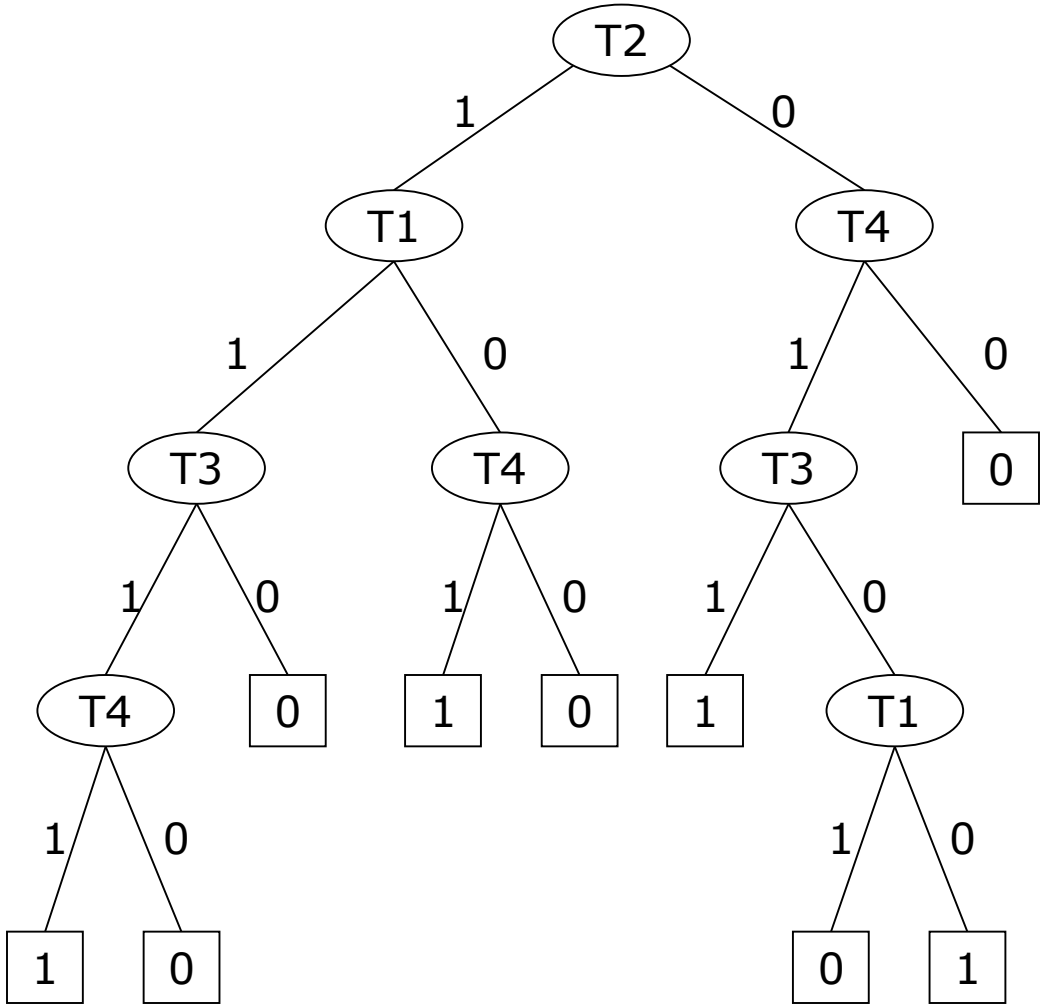
近似的解法 貪欲(greedy)アルゴリズム

- 「最も良さそうな」テストを選択し、
  レコード集合を分割
- 分割したレコード集合に
  同じ戦略を繰り返し適用する
- 一度選択したテストは変更しない
- 「最も良さそうな」 の基準を工夫



最初にT2ではなくT1を選ぶべき





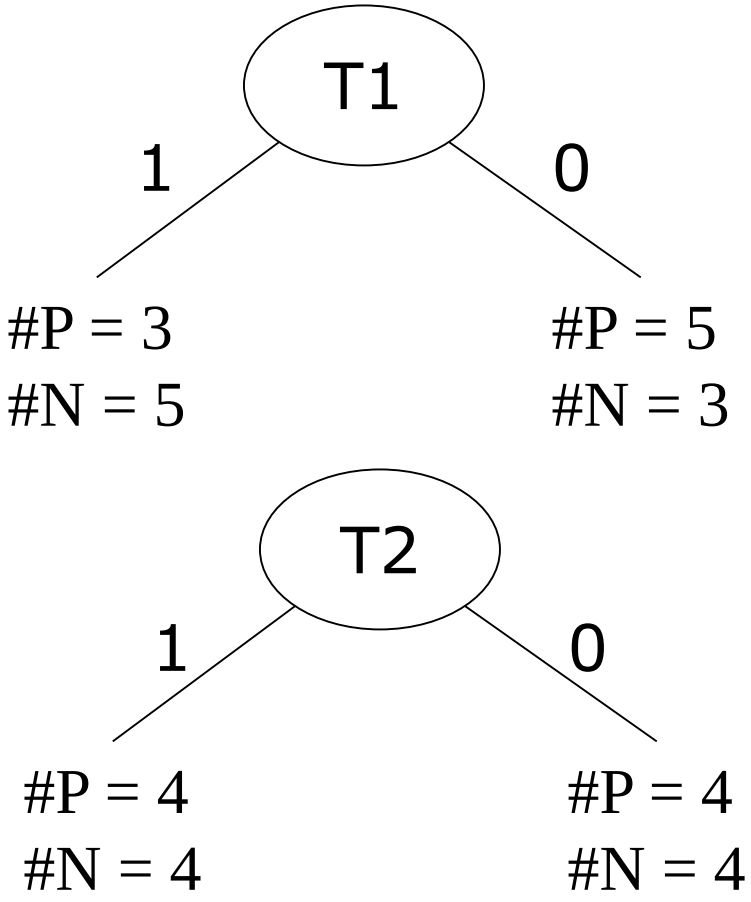
3+3+2+2+2=12

4+4+3+3+3+3+4+4+2=30

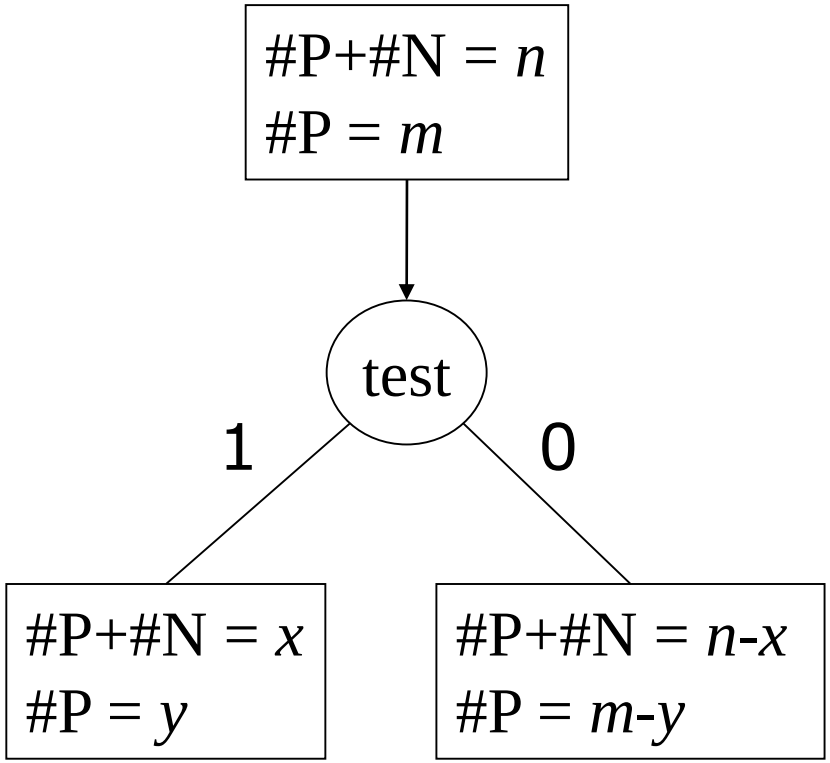
T2 ではなく T1 を選択する評価基準は？

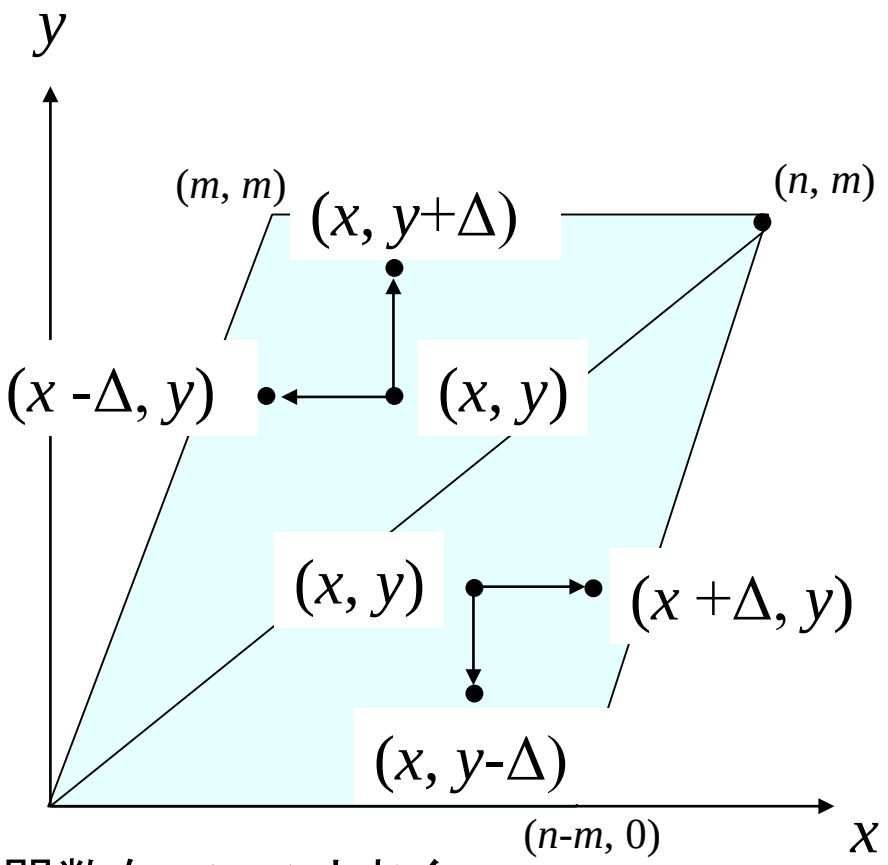
T1	T2	T3	T4	目標属性
1	0	1	1	1
1	0	1	1	1
1	1	1	1	1
1	1	1	0	0
1	0	1	0	0
1	1	0	1	0
1	0	0	1	0
1	1	0	1	0
0	1	0	1	1
0	0	1	1	1
0	1	0	1	1
0	1	0	1	1
0	0	0	1	1
0	0	1	0	0
0	1	0	0	0
0	0	1	0	0

#P : 正レコードの数
#N: 負レコードの数



テスト選択のための評価値を定める





- test を選択すると (x, y) が定まる．評価関数を $\varphi(x, y)$ とおく
- 評価基準 $\varphi(x, y)$ が満たしてほしい条件

  $\varphi(x, y)$ は $m/n = y/x$ のとき最小

  $\varphi(x, y) \leqq \varphi(x, y+\Delta), \varphi(x, y) \leqq \varphi(x -\Delta, y)$ if $m/n < y/x$

  $\varphi(x, y) \leqq \varphi(x, y-\Delta), \varphi(x, y) \leqq \varphi(x +\Delta, y)$ if $m/n > y/x$

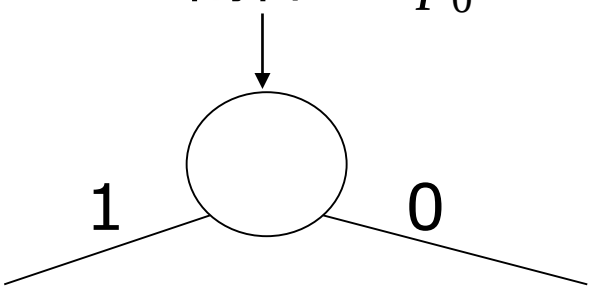
  $\Delta > 0$

評価値:エントロピーゲイン

正レコードの割合	p = #P / (#P + #N)
負レコードの割合	1-p = #N / (#P+#N)
エントロピー	ent(p) = - $p \log_2 p$ - (1-p) $\log_2$ (1-p)

レコード数 n

正レコードの割合 $p_0 = m / n$



$n_1 = x$ $n_2 = n - x$

$p_1 = y / x$ $p_2 = (m - y) / (n - x)$

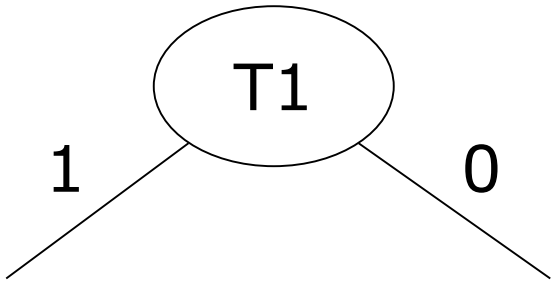
Entropy Gain

$$\mathrm{Ent}(x, y) = \mathrm{ent}(p_0) - (n_1/n)\,\mathrm{ent}(p_1) - (n_2/n)\,\mathrm{ent}(p_2)$$

T1を選んだときのエントロピーゲイン

#P = 8
#N = 8

ent(8/16)
= 2 (- (1/2) $\log_2(1/2)$)
= 1



#P = 3
#N = 5

#P = 5
#N = 3

ent(3/8)
= - (3/8)$\log_2(3/8)$ - (5/8)$\log_2(5/8)$
= 0.95444

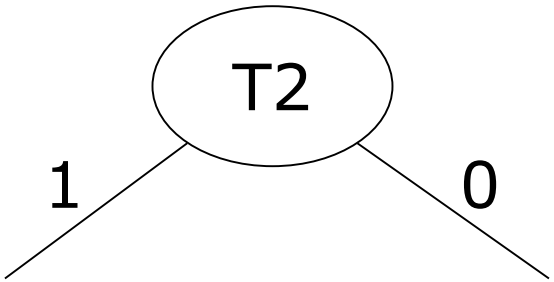
ent(5/8) = ent(3/8)

Entropy Gain = ent(8/16) - (8/16)ent(3/8) - (8/16)ent(5/8)
= 1 - 0.95444
= 0.04556

T2を選んだときのエントロピーゲイン

#P = 8
#N = 8

ent(8/16) = 1



#P = 4
#N = 4

#P = 4
#N = 4

ent(4/8) = 1

ent(4/8) = 1

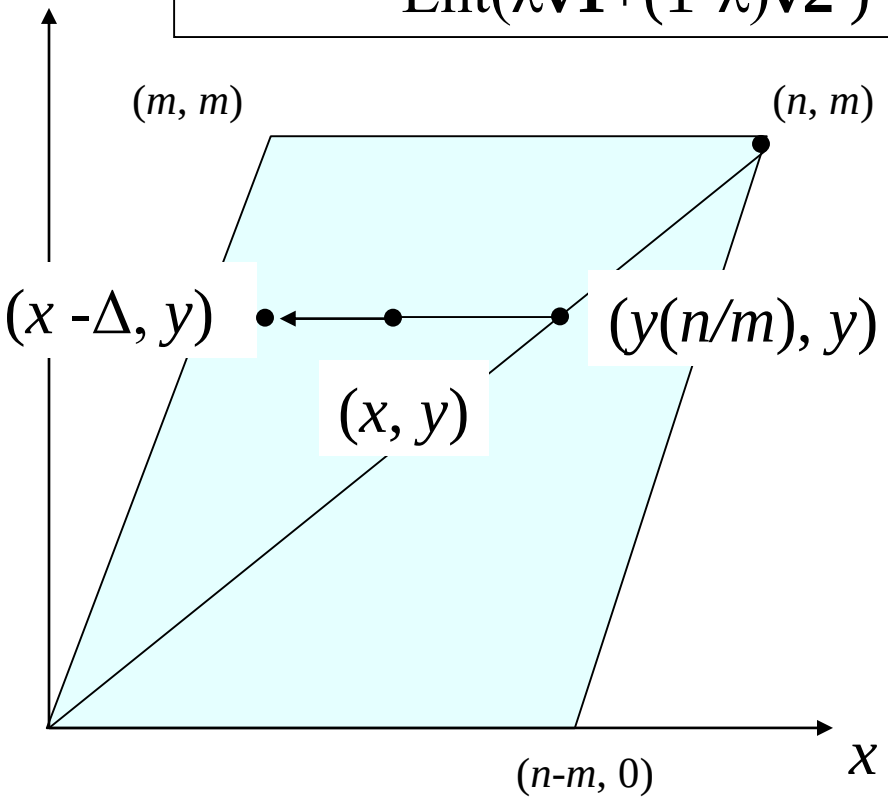
Entropy Gain = ent(8/16) - (8/16)ent(4/8) - (8/16)ent(4/8)
= 0

- Ent(x, y) は $m/n = y/x$ のとき最小
- Ent(x, y) は凸関数　任意の点 **v1**, **v2** と $0 \leqq \lambda \leqq 1$について

$$\mathrm{Ent}(\lambda\mathbf{v1}+(1-\lambda)\mathbf{v2}) \leqq \lambda\mathrm{Ent}(\mathbf{v1}) + (1-\lambda)\,\mathrm{Ent}(\mathbf{v2})$$

すると

$$\mathrm{Ent}(\lambda\mathbf{v1}+(1-\lambda)\mathbf{v2}) \leqq \max\{\mathrm{Ent}(\mathbf{v1}), \mathrm{Ent}(\mathbf{v2})\}$$



Ent(x,y)

$\leqq$ $\max\{\mathrm{Ent}(x-\Delta, y), \mathrm{Ent}(y(n/m), y)\}$

$\leqq$ $\mathrm{Ent}(x-\Delta, y)$

演習問題

T1	T2	T3	T4	目標属性
1	0	1	1	1
1	1	1	1	1
1	1	1	0	0
1	0	1	0	0
1	1	0	1	0
1	0	0	1	0
1	1	0	1	0
0	1	0	1	1
0	0	1	1	1
0	1	0	1	1
0	0	0	1	1
0	0	1	0	0
0	1	0	0	0
0	0	1	0	0

- 左の表のようなテスト-目標属性間の関係があるとき，エントロピーゲインに基づいて決定木を求めなさい