知能機械と自然言語処理
知能機械部　第10回

ソフトウェア情報学部

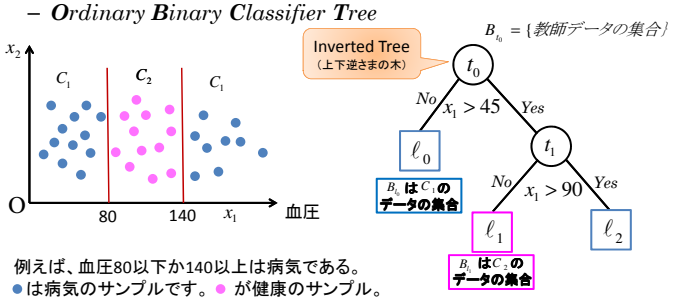
Goutam Chakraborty

決定木
非線形分類方法の一つです

特色空間は順番に分割して、各部分に不純物（Impurity - 異なる分類のサンプル）がないように分類します。同じクラスのデータが特色空間に散らばっている時、特にデータの数が少ない場合有効である方法です。

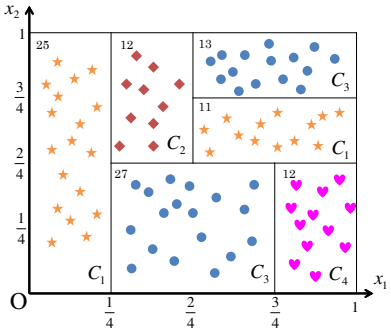
最も簡単な決定木がOBCTと言います

– ***O**rdinary **B**inary **C**lassifier **T**ree*

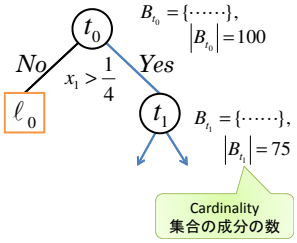


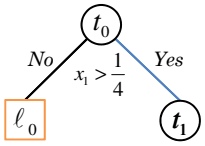
例えば、血圧80以下か140以上は病気である。
● は病気のサンプルです。● が健康のサンプル。

決定木（2）



左の図より、以下の決定木を完成させよ。



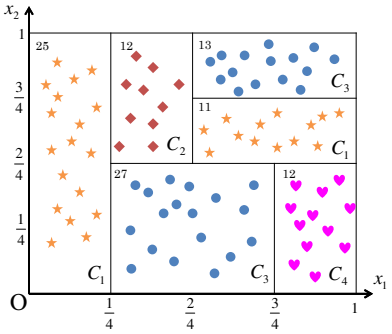


各ノードで、どのような条件に設定すれば簡潔な決定木を作成できるか？

決定木を作成するルール

1. バイナリ決定木 – 全ての決定してないノード（t）から二つのノードに分ける。該当するノード t_i のサンプル集合を B_{t_i} とする。
2. 条件は（各特色の値＞定数）として、左はNo、右はYesとする。
3. ノード二つに分ける際、分けた後のノードに不純物がなるべく入らないようにする。詳細は次のスライドで説明する。
4. 分けた後のノードに不純物が入っていなければ、それを決定したノード（ℓ - leaf 葉っぱ）としてそこで収束する。

下記のデータを用いて不純物の量の定義を説明します



不純物の量の定義
Node Impurity Definition

$$I_t = -\sum_{i=1}^{k} P(C_i)\lg P(C_i)$$

I_t：不純物の量
$P(C_i)$：クラス C_i の確率
$\lg x = \log_2 x = \frac{\log_{10} x}{\log_{10} 2}$

さまざまなクラスのサンプルが混ざっている程 I_t の 値は大きくなる。

$$I_{t_0} = -P(C_1)\lg P(C_1) - P(C_2)\lg P(C_2) - P(C_3)\lg P(C_3) - P(C_4)\lg P(C_4) = -0.36\lg 0.36 - 0.12\lg 0.12 - 0.40\lg 0.40 - 0.12\lg 0.12 = 1.79$$

不純物の量の計算
Impurity Calculation

$$I_{t_1} = -\frac{11}{75}\lg\frac{11}{75} - \frac{12}{75}\lg\frac{12}{75} - \frac{40}{75}\lg\frac{40}{75} - \frac{12}{75}\lg\frac{12}{75} = 1.74$$

$$I_{\ell_1} = -\frac{25}{25}\lg\frac{25}{25} - \frac{0}{25}\lg\frac{0}{25} - \frac{0}{25}\lg\frac{0}{25} - \frac{0}{25}\lg\frac{0}{25} = 0$$

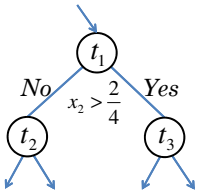
$$(\because \lim_{x\to 0} x\lg x = 0)$$

$$\Delta I = I_{t_0} - \left(\frac{|B_{\ell_0}|}{|B_{t_0}|} I_{\ell_0} + \frac{|B_{t_1}|}{|B_{t_0}|} I_{t_1}\right)$$

$$= 1.79 - \left(\frac{25}{100}\times 0 + \frac{75}{100}\times 1.74\right) = 0.485$$

ノードを分けた際の不純物量の変化の計算

$$|B_{t_2}| = 39,\ |B_{t_3}| = 36$$



$$I_{t_2} = -\frac{27}{39}\lg\frac{27}{39} - \frac{12}{39}\lg\frac{12}{39} = 0.89$$

$$I_{t_3} = -\frac{12}{36}\lg\frac{12}{36} - \frac{11}{36}\lg\frac{11}{36} - \frac{13}{36}\lg\frac{13}{36} = 1.58$$

$$\Delta I = 1.74 - \left(\frac{39}{75}\times 0.89 + \frac{36}{75}\times 1.58\right) = 0.52$$

分ける前のノードの不純物量と分けた後の二つのノードの不純物量の差が一番大きくなるように条件式を設定する。そのために全ての特色に対して検証を行う。

決定木の解答