

知識工学

岡山大学大学院
講師 竹内孔一

本日の内容

- 概念の学習
 - 決定木

決定木による学習

- 決定木
 - ある判断を下したい時にデータのどの属性を見てどう判断するかを記述したもの
- 特徴
 - 属性の組のすべてに対して与えられたクラスから分類を行う
- ヴァージョン空間法との違い
 - 詳細な組み合わせをすべて考える(利点)
 - 詳細すぎる組み合わせを防ぐ機構が必要(欠点)

問題例

- お客がお弁当を買う条件は何か?

| 料理 | おかず | 価格 | 正負 |
|----|-----|----|----|
| 和 | 多 | 中 | 正 |
| 和 | 小 | 高 | 負 |
| 中華 | 小 | 中 | 正 |
| 中華 | 多 | 高 | 負 |
| 洋 | 多 | 中 | 正 |
| 洋 | 小 | 中 | 負 |
| 和 | 小 | 中 | 正 |
| 洋 | 小 | 高 | 負 |

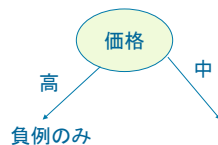
データマイニング:
データから有益な
情報(規則)を
取り出すこと

お弁当が良く売れる
条件は?

左は弁当がうれたときの
条件を記録したものとする

決定木による学習

- 目標
 - 決定木(decision tree)を作成する
各接点が属性、葉が分類となっている木
 - > 属性の順番を決める 例:客がお弁当を買う条件
- 用意
 - 属性と各事例のクラス分類
 - 分類の基準
 - なんの属性から分類すると
情報量を下げられるか(ID3) 負例のみ



決定木の学習

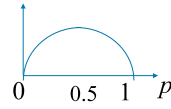
- 分類基準
 - 情報量を計算してもっとも低いものから分類する
- 情報量
 - 小さい: 偏っている (特定の情報がある)
 - 大きい: 偏っていない, 均一, (情報が無い)

方法
接点となる属性について
情報量を測定してもっとも
低いものをその接点とする



情報量について

N: ある属性でのデータ数
 E: ある属性で分割した結果の情報量
 p: 正のクラスに属する確率
 1-p: 負のクラスに属する確率



$$\text{平均情報量} = -p \log p - (1-p) \log(1-p)$$

偏りがあると値が小さい

z_{j+} : ある属性での各属性値jに分類された正事例の数
 z_{j-} : 同様に負事例の数

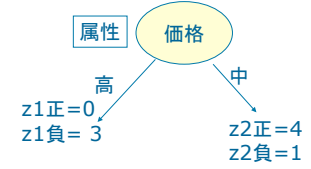
$$E = - \left(\sum_j z_j (p \log p + (1-p) \log(1-p)) / N \right)$$

各属性で計算してEが小さい属性を判別に使う

計算例

- お弁当に関する属性と属性値

まず、価格で分類
 全体の数8



$$E_{\text{価格}} = - \left\{ 3 \left(0 + \frac{3}{3} \log \frac{3}{3} \right) + 5 \left(\frac{4}{5} \log \frac{4}{5} + \frac{1}{5} \log \frac{1}{5} \right) \right\} / 8$$

$$= -(4 \log 4 - 5 \log 5) / 8$$

練習18

- お弁当の表で「おかず」で分類した場合の情報量を求めよ

| 料理 | おかず | 価格 | 正負 |
|----|-----|----|----|
| 和 | 多 | 中 | 正 |
| 和 | 小 | 高 | 負 |
| 中華 | 小 | 中 | 正 |
| 中華 | 多 | 高 | 負 |
| 洋 | 多 | 中 | 正 |
| 洋 | 小 | 中 | 負 |
| 和 | 小 | 中 | 正 |
| 洋 | 小 | 高 | 負 |

• E(おかず)と
 E(価格)では
 どちらが良いか?

学習における難しさ

- 分類の難しさ
 - 決定木が細かすぎる
 - 学習事例に過度に学習
 未知のデータに対してよくない
 - 丁度いいところでとめる必要がある
 -> 枝狩り pruning
- 回避の方法
 - 未知のデータに対する予測
 - 分布を仮定: 二項分布