ネットワークコンピューティング(4)
情報分類

関西学院大学理工学部情報科学科
北村泰彦

演習問題

	属性a	属性b
x_1	5	1
x_2	4	2
x_3	1	5
x_4	5	4
x_5	5	5

クラスタ数を3，初期データ分割を
C_1={x_1,x_2},C_2={x_3,x_4},C_3={x_5}とする.

C_1，C_2，C_3の重心はそれぞれ

$$\left(\frac{5+4}{2},\frac{1+2}{2}\right)=(4.5,1.5)$$

$$\left(\frac{1+5}{2},\frac{5+4}{2}\right)=(3,4.5)$$

$$(5,5)$$

	C_1	C_2	C_3
x_1	0.70	4.03	4
x_2	0.70	2.69	3.16
x_3	4.94	2.06	4
x_4	2.54	2.06	1
x_5	3.53	2.06	0

x_4 はC_3へ移動させる.

演習問題

	属性a	属性b
x_1	5	1
x_2	4	2
x_3	1	5
x_4	5	4
x_5	5	5

クラスタ数を3，データ分割を
C_1={x_1,x_2},C_2={x_3},C_3={x_4,x_5}とする.

C_1，C_2，C_3の重心はそれぞれ

$$\left(\frac{5+4}{2},\frac{1+2}{2}\right)=(4.5,1.5)$$

$$(1,5)$$

$$\left(\frac{5+5}{2},\frac{4+5}{2}\right)=(5,4.5)$$

	C_1	C_2	C_3
x_1	0.70	5.65	3.5
x_2	0.70	4.24	2.69
x_3	4.94	0	4.03
x_4	2.54	4.12	0.5
x_5	3.53	4	0.5

終了

情報分類

- クラスタリングとは複数の事例をいくつかのグループに分割すること.
- [　　](classification)はある事例を適切なクラス（あるいはカテゴリ）に割り当てること.
- 今回は，すでに分類された事例集合を解析して，新しい事例を自動的に分類できる機械学習アルゴリズムである,[　　],[　　],[　　]を解説する.

分類の必要性

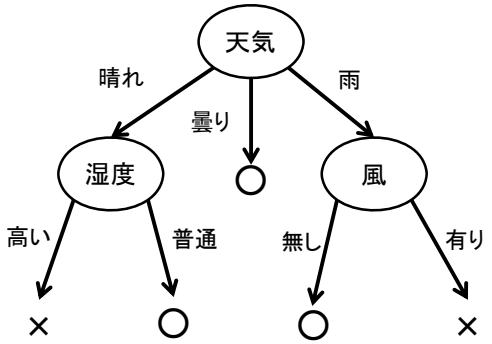
- 電子メールフィルタリング：メールを分類し，スパムメールを排除する．
- 電子ニュースの分類：ニュース記事をカテゴリ毎に分類する．

- [　　]は与えられた事例集合からクラスを生成すること．[　　]は事例をクラスに分類すること．

[　　]

- 決定木は，意思決定や物事の分類を多段階で繰り返し実行する場合，その多段の分岐過程を階層化して樹形図で表現したグラフ表現．目的属性がカテゴリー型である[　　]と，数値型である[　　]に分かれる．
- [　　]は，木の根ノードから葉ノードまでに記述されているテストを繰り返し実行することで，事例を分類する木である．

決定木



気象条件からゴルフプレイの可否を決めるための決定木

決定木学習アルゴリズム

1. 根ノードに置く属性を決定し，その属性値に応じて分岐を作成する．
2. 事例の集合を各分岐に応じて部分集合に分割して子ノードを作成し，その子ノードを根ノードとする．
3. 1と2のプロセスを再帰的に繰り返し，決定木を成長させる．
4. 子ノードの全ての事例が同一クラスに属していれば，終了する．

決定木学習アルゴリズム

- 決定木学習アルゴリズムで問題になるのは，事例集合を分割する属性の選定である．基本的な考え方としては，決定の不確実さがもっとも減少させるような属性を選定すればよい．
- 情報の不確実さを表現する指標に　　ビットとして計算される　　がある．ここでp_iはi番目の事象の生起確率である．
- 　　（＝分割前の平均情報量ー分割後の平均情報量）が最大となる属性を，順次選択して決定木を構成する．

気象条件とゴルフプレイに関するデータ

天気	温度	湿度	風	ゴルフプレイ
晴れ	暑い	高い	無し	×
晴れ	暑い	高い	有り	×
曇り	暑い	高い	無し	○
雨	暖かい	高い	無し	○
雨	涼しい	普通	無し	○
雨	涼しい	普通	有り	×
曇り	涼しい	普通	有り	○
晴れ	暖かい	高い	無し	×
晴れ	涼しい	普通	無し	○
雨	暖かい	普通	無し	○
晴れ	暖かい	普通	有り	○
曇り	暖かい	高い	有り	○
曇り	暑い	普通	無し	○
雨	暖かい	高い	有り	×

情報利得

- 天気が「晴れ」の場合，○と×の数はそれぞれ2と3である．従って，その平均情報量は $-\left(\frac{2}{5}\log_2\frac{2}{5}+\frac{3}{5}\log_2\frac{3}{5}\right)=0.971$
- 同様に「曇り」と「雨」の場合の情報量は0と0.971である．
- したがって，天気の平均情報量（エントロピ）は $\frac{5}{14}\times 0.971+\frac{4}{14}\times 0+\frac{5}{14}\times 0.971=0.693$

情報利得

- 根ノードの平均情報量は○と×の数が9と5に分かれるので，$-\left(\frac{9}{14}\log_2\frac{9}{14}+\frac{5}{14}\log_2\frac{5}{14}\right)=0.940$.
- したがって，分割属性を天気とした場合の情報利得は$0.940-0.693=0.247$である．
- 同様に分割属性を温度，湿度，風とした場合の情報利得はそれぞれ0.029,0.152,0.048となり，天気を選択することが情報利得を最大化する．

情報利得

属性	属性値	○	×	情報量	平均情報量	情報利得
天気	晴れ	2	3	0.971	0.693	0.247
	曇り	4	0	0.000		
	雨	3	2	0.971		
温度	暑い	2	2	1.000	0.911	0.029
	暖かい	4	2	0.918		
	涼しい	3	1	0.811		
湿度	高い	3	4	0.985	0.788	0.152
	普通	6	1	0.592		
風	無し	6	2	0.811	0.892	0.048
	有り	3	3	1.000		

情報利得

- 天気が曇りの場合は○と確定するので，子ノードの分割は必要ないが，晴れと雨の場合は○と×が混在するので，同様のプロセスで決定木を成長させる必要がある．

「晴れ」とゴルフプレイに関するデータ

天気	温度	湿度	風	ゴルフプレイ
晴れ	暑い	高い	無し	×
晴れ	暑い	高い	有り	×
晴れ	暖かい	高い	無し	×
晴れ	涼しい	普通	無し	○
晴れ	暖かい	普通	有り	○

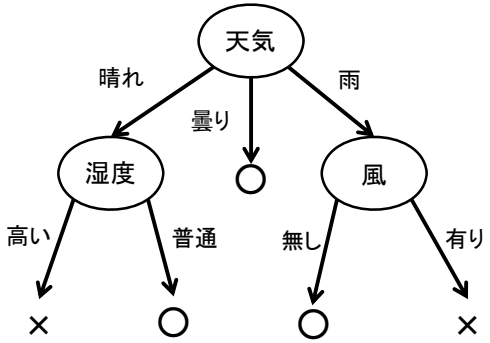
属性	属性値	○	×	情報量	平均情報量	情報利得
温度	暑い	0	2	0.000	0.400	0.571
	暖かい	1	1	1.000		
	涼しい	1	0	0.000		
湿度	高い	0	3	0.000	0.000	0.971
	普通	2	0	0.000		
風	無し	1	2	0.918	0.951	0.020
	有り	1	1	1.000		

「雨」とゴルフプレイに関するデータ

天気	温度	湿度	風	ゴルフプレイ
雨	暖かい	高い	無し	○
雨	涼しい	普通	無し	○
雨	涼しい	普通	有り	×
雨	暖かい	普通	無し	○
雨	暖かい	高い	有り	×

属性	属性値	○	×	情報量	平均情報量	情報利得
温度	暑い	0	0	0.000	0.951	0.020
	暖かい	2	1	0.918		
	涼しい	1	1	1.000		
湿度	高い	1	1	1.000	0.951	0.020
	普通	2	1	0.918		
風	無し	3	0	0.000	0.000	0.971
	有り	0	2	0.000		

決定木



気象条件からゴルフプレイの可否を決めるための決定木

- は，ある基準により選択された分割属性により事例を分割することを再帰的に繰り返す，である．
- は，全事例を分割するのではなく，全事例からルールで説明できる事例を取り除くことを再帰的に繰り返す，である．

- カバーリングアルゴリズムは，クラス分類の確率を最大化するような「属性とその値」の対を選択しながら，ルールの条件部を操作（追加，削除）し，最大の精度をもつルールを生成する．

眼年齢	処方箋	乱視	涙量率	推奨レンズ
若い	近視	なし	減少	なし
若い	近視	なし	正常	ソフト
若い	近視	あり	減少	なし
若い	近視	あり	正常	ハード
若い	遠視	なし	減少	なし
若い	遠視	なし	正常	ソフト
若い	遠視	あり	減少	なし
若い	遠視	あり	正常	ハード
老眼前	近視	なし	減少	なし
老眼前	近視	なし	正常	ソフト
老眼前	近視	あり	減少	なし
老眼前	近視	あり	正常	ハード
老眼前	遠視	なし	減少	なし
老眼前	遠視	なし	正常	ソフト
老眼前	遠視	あり	減少	なし
老眼前	遠視	あり	正常	なし
老眼	近視	なし	減少	なし
老眼	近視	なし	正常	なし
老眼	近視	あり	減少	なし
老眼	近視	あり	正常	ハード
老眼	遠視	なし	減少	なし
老眼	遠視	なし	正常	ソフト
老眼	遠視	あり	減少	なし
老眼	遠視	あり	正常	なし

カバーリングアルゴリズム

- 以下のルールを学習することを目標とする.
  If [?] Then 推奨レンズ=ハード
- 未知の項?には，9つの選択肢が存在する.
  - 眼年齢=若い:2/8
  - 眼年齢=老眼前:1/8
  - 眼年齢=老眼:1/8
  - 処方箋=近視:3/12
  - 処方箋=遠視:1/12
  - 乱視=なし:0/12
  - 乱視=あり:4/12
  - 涙量率=減少:0/12
  - 涙量率=正常:4/12

カバーリングアルゴリズム

- 最大正答率の選択肢は「乱視=あり」なので，以下のルールを選択する.
  If [乱視=あり] Then 推奨レンズ=ハード
- このルールがカバーする範囲は4/12であるので，新しい条件を付加して，ルールを洗練する.
  If [乱視=あり] and [?] Then 推奨レンズ=ハード

眼年齢	処方箋	乱視	涙量率	推奨レンズ
若い	近視	あり	減少	なし
若い	近視	あり	正常	ハード
若い	遠視	あり	減少	なし
若い	遠視	あり	正常	ハード
老眼前	近視	あり	減少	なし
老眼前	近視	あり	正常	ハード
老眼前	遠視	あり	減少	なし
老眼前	遠視	あり	正常	なし
老眼	近視	あり	減少	なし
老眼	近視	あり	正常	ハード
老眼	遠視	あり	減少	なし
老眼	遠視	あり	正常	なし

カバーリングアルゴリズム

- ここで未知の項?で考えられる条件項は，以下の7つ.
  - 眼年齢=若い:2/4
  - 眼年齢=老眼前:1/4
  - 眼年齢=老眼:1/4
  - 処方箋=近視:3/6
  - 処方箋=遠視:1/6
  - 涙量率=減少:0/6
  - 涙量率=正常:4/6
- 「涙量率=正常」が最大正答率なので，以下のルールを得る.
  If [乱視=あり] and [涙量率=正常] Then 推奨レンズ=ハード

眼年齢	処方箋	乱視	涙量率	推奨レンズ
若い	近視	あり	正常	ハード
若い	遠視	あり	正常	ハード
老眼前	近視	あり	正常	ハード
老眼前	遠視	あり	正常	なし
老眼	近視	あり	正常	ハード
老眼	遠視	あり	正常	なし

カバーリングアルゴリズム

- さらに正答率の向上を目指して，以下の条件項を検討する.
  - 眼年齢=若い：2/2
  - 眼年齢=老眼前：1/2
  - 眼年齢=老眼：1/2
  - 処方箋=近視：3/3
  - 処方箋=遠視：1/3
- 「眼年齢=若い」と「処方箋=近視」が最大正答率100%だが，カバー数の多い「処方箋=近視」を採用し，以下のルールを得る.
  If [乱視=あり] and [涙量率=正常] and [処方箋=近視] Then 推奨レンズ=ハード
- このルールは24事例のうちの3事例しかカバーしていないので，3事例をデータセットから削除し，ルール学習を繰り返す.

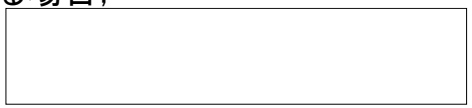
眼年齢	処方箋	乱視	涙量率	推奨レンズ
若い	近視	なし	減少	なし
若い	近視	なし	正常	ソフト
若い	近視	あり	減少	なし
若い	近視	あり	正常	ハード
若い	遠視	なし	減少	なし
若い	遠視	なし	正常	ソフト
若い	遠視	あり	減少	なし
若い	遠視	あり	正常	ハード
老眼前	近視	なし	減少	なし
老眼前	近視	なし	正常	ソフト
老眼前	近視	あり	減少	なし
老眼前	近視	あり	正常	ハード
老眼前	遠視	なし	減少	なし
老眼前	遠視	なし	正常	ソフト
老眼前	遠視	あり	減少	なし
老眼前	遠視	あり	正常	なし
老眼	近視	なし	減少	なし
老眼	近視	なし	正常	なし
老眼	近視	あり	減少	なし
老眼	近視	あり	正常	ハード
老眼	遠視	なし	減少	なし
老眼	遠視	なし	正常	ソフト
老眼	遠視	あり	減少	なし
老眼	遠視	あり	正常	なし

カバーリングアルゴリズム

- 同様にして正答率1/1の以下のルールが得られる.
  If [眼年齢=若い] and [乱視=あり] and [涙量率=正常] Then 推奨レンズ=ハード
- 次に，推奨レンズが「ソフト」と「なし」のルールを生成する.

- [redacted]と[redacted]のいずれにおいても属性は互いに[redacted]であることを仮定していた．現実世界では属性が互いに独立であるという前提は成立しないこともある．
- [redacted]は属性が互いに独立であることを前提としない．

- ナイーブベイズ学習はベイズ則に基づく分類法である．
- ベイズ則は，もし仮説Hとその仮説を受け入れる証拠Eがある場合，



で表される．
- Hを「ゴルフをする」，Eを「雨が降る」としたとき，「雨が降るときにゴルフをする」確率は「ゴルフをしているときに雨が降っている」確率に「ゴルフをする」確率をかけ，「雨が降る」確率で割った値に等しい．
- P(E|H)とP(H)はこれまでの経験から算出可能である．P(E)は分からなくても問題ない．

気象データのクラス分布と生起確率

天気			温度			湿度			風			ゴルフ	
	○	×		○	×		○	×		○	×	○	×
晴れ	2	3	暑い	2	2	高い	3	4	無し	6	2	9	5
曇り	4	0	暖かい	4	2	普通	6	1	有り	3	3		
雨	3	2	涼しい	3	1								
晴れ	2/9	3/5	暑い	2/9	2/5	高い	3/9	4/5	無し	6/9	2/5	9/14	5/14
曇り	4/9	0/5	暖かい	4/9	2/5	普通	6/9	1/5	有り	3/9	3/5		
雨	3/9	2/5	涼しい	3/9	1/5								

例

- 「天気が晴れ，温度が涼しく，湿度が高く，風があるとき，ゴルフをすべきか？」
- ベイズ則より「ゴルフをする」条件付き確率を求める．

$$\frac{2/9 \times 3/9 \times 3/9 \times 3/9 \times 9/14}{P(E)} = \frac{0.0053}{P(E)}$$

- 同様に「ゴルフをしない」条件付き確率を求める．

$$\frac{3/5 \times 1/5 \times 4/5 \times 3/5 \times 5/14}{P(E)} = \frac{0.0206}{P(E)}$$

- したがって，「ゴルフをしない」

参考文献

- Marmanis and Babenko: Algorithms of the Intelligent Web, Manning, 2009.
- 元田，津本，山口，沼尾：データマイニングの基礎，オーム社，2006.