ネットワークコンピューティング(3)
情報集約

関西学院大学理工学部情報科学科
北村泰彦

演習問題

	1:親子丼	2:牛丼	3:海鮮丼	4:カツ丼
1:山田	1	3	*	3
2:田中	*	1	3	*
3:佐藤	2	1	3	1
4:鈴木	1	3	2	*

上の表は，R={1,2,3}とする評価値行列Sである．1:山田の3:海鮮丼に対する推定評価値を求めよ.

演習問題

- 協調フィルタリングの例において，1:山田の3:海鮮丼に対する推定評価値を求めよ.
- 1:山田と2:田中の相関$\rho_{1,2}$は，共通に評価しているアイテムが2:牛丼だけなので，$\rho_{1,2}=0$である.
- 次に，1:山田と3:佐藤の相関を計算する．この二人がともに評価しているアイテムは1:親子丼，2:牛丼，4:カツ丼なので，$Y_{1,3}=\{1,2,4\}$となる．これらのアイテムについての$Y_{1,3}$上の平均評価値はそれぞれ以下の通りである.

$$\bar{s}'_1=\frac{\sum_{k=1,2,4}s_{1,k}}{3}=\frac{1+3+3}{3}=7/3$$
$$\bar{s}'_3=\frac{\sum_{k=1,2,4}s_{3,k}}{3}=\frac{2+1+1}{3}=4/3$$

演習問題

- したがって相関は

$$\rho_{1,3}=\frac{\sum_{k=1,2,4}(s_{1,k}-\bar{s}'_1)(s_{3,k}-\bar{s}'_3)}{\sqrt{\sum_{k=1,2,4}(s_{1,k}-\bar{s}'_1)^2}\sqrt{\sum_{k=1,2,4}(s_{3,k}-\bar{s}'_3)^2}}$$

$$=\frac{(1-7/3)(2-4/3)+(3-7/3)(1-4/3)+(3-7/3)(1-4/3)}{\sqrt{(1-7/3)^2+(3-7/3)^2+(3-7/3)^2}\sqrt{(2-4/3)^2+(1-4/3)^2+(1-4/3)^2}}$$
$$=-1$$

- 同様に計算すると1:山田と4:鈴木の相関は

$$\rho_{1,4}=\frac{\sum_{k=1,2}(s_{1,k}-\bar{s}'_1)(s_{4,k}-\bar{s}'_4)}{\sqrt{\sum_{k=1,2}(s_{1,k}-\bar{s}'_1)^2}\sqrt{\sum_{k=1,2}(s_{4,k}-\bar{s}'_4)^2}}$$
$$=\frac{(1-2)(1-2)+(3-2)(3-2)}{\sqrt{(1-2)^2+(3-2)^2}\sqrt{(1-2)^2+(3-2)^2}}$$
$$=1$$

演習問題

- 次に推定評価値を計算する．まず，1:山田の全評価済みアイテム上の平均評価値を求める．

$$\bar{s}_1 = \frac{\sum_{k=1,2,4} s_{1,k}}{3} = \frac{1+3+3}{3} = 7/3$$

- したがって，

$$\hat{s}_{1,3} = \bar{s}_1 + \frac{\sum_{i=2,3,4} \rho_{1,i}(s_{i,3} - \bar{s'}_i)}{\sum_{i=2,3,4}|\rho_{1,i}|}$$

$$= 7/3 + \frac{0(3-1) + (-1)(3-4/3) + 1(2-2)}{|0| + |-1| + |1|}$$

$$= 1.5$$

よって1:山田は3:海鮮丼がそれほど好きでないと予測される．

情報集約

- インターネット上には膨大な量の情報が存在し，その整理が必要である．
- その整理のための手法として[　　]([　　])と[　　](classification)がある．
- クラスタリングは同じようなアイテムをグループ化する過程を指す．
- 例えば，本のリストを著者毎にグループ化すること．

クラスタリングの必要性

- ニュース記事の集約．インターネット上にある同様のニュース記事をグループ化する．例：Google検索のニュース (http://www.google.co.jp/)
- ユーザの集約．ショッピングサイトに訪問するユーザをグループ化し，それに応じたマーケティング（広告メールの配信など）を行う．

具体例

名前	年齢	収入	学歴	スキル	社交性	常勤
Albert	23	0	0	3	3	0
Alexandra	25	1	2	4	2	0
Athena	24	0	1	3	4	0
Aurora	23	1	2	5	2	0
Babis	21	0	0	3	4	0
Bill	31	1	2	4	2	0
Bob	32	1	1	3	1	1
Carl	30	0	2	4	2	0
Catherine	31	1	1	3	3	0
Charlie	30	1	2	3	2	0
Constantine	37	1	1	3	2	0
Dmitry	35	2	2	1	1	1
Elena	38	1	1	3	2	0
Eric	37	2	2	2	2	0
Frank	39	3	1	3	1	1
George	42	2	2	2	1	1
Jack	43	3	1	1	1	1
John	45	4	2	1	1	1
Lukas	45	3	2	1	1	1
Maria	43	2	1	3	1	0

上記の利用者をどのようにグループ化するか？

(1/2)

名前	年齢	収入	学歴	スキル	社交性	常勤
Babis	21	0	0	3	4	0
Albert	23	0	0	3	3	0
Aurora	23	1	2	5	2	0
Athena	24	0	1	3	4	0
Alexandra	25	1	2	4	2	0
Carl	30	0	2	4	2	0
Charlie	30	1	2	3	2	0
Bill	31	1	2	4	2	0
Catherine	31	1	1	3	3	0
Bob	32	1	1	3	1	1
Dmitry	35	2	2	1	1	1
Constantine	37	1	1	3	2	0
Eric	37	2	2	2	2	0
Elena	38	1	1	3	2	0
Frank	39	3	1	3	1	1
George	42	2	2	2	1	1
Maria	43	2	1	3	1	0
Jack	43	3	1	1	1	1
Lukas	45	3	2	1	1	1
John	45	4	2	1	1	1

年齢でソーティングを行い，同年齢は収入でソーティングする．
ただし，二つの属性しか考慮していない．

ソーティング(2/2)

名前	年齢	収入	学歴	スキル	社交性	常勤	距離
Babis	21	0	0	3	4	0	21.58703
Albert	23	0	0	3	3	0	23.38803
Aurora	23	1	2	5	2	0	23.72762
Athena	24	0	1	3	4	0	24.53569
Alexandra	25	1	2	4	2	0	25.4951
Charlie	30	1	2	3	2	0	30.29851
Carl	30	0	2	4	2	0	30.39737
Catherine	31	1	1	3	3	0	31.32092
Bill	31	1	2	4	2	0	31.40064
Bob	32	1	1	3	1	1	32.20248
Dmitry	35	2	2	1	1	1	35.15679
Constantine	37	1	1	3	2	0	37.20215
Eric	37	2	2	2	2	0	37.21559
Elena	38	1	1	3	2	0	38.19686
Frank	39	3	1	3	1	1	39.26831
George	42	2	2	2	1	1	42.16634
Jack	43	3	1	1	1	1	43.1509
Maria	43	2	1	3	1	0	43.17407
Lukas	45	3	2	1	1	1	45.17743
John	45	4	2	1	1	1	45.25483

原点からの距離$\sqrt{\sum_{i=1}^{n} {a_i}^2}$でソーティングを行う．
ただし，a_iはi番目の属性値で，属性の数はnとする．

(1/2)

- Agglomerative Hierarchical Clustering (AHC)
- 入力：事例の集合$\{x_1, x_2, \ldots, x_n\}$, $G = \{G_1, G_2, \ldots, G_n\}$

1. 全ての事例x_iに対して，初期クラスタを$G_i := \{x_i\}$，各事例間の類似度を$s(G_i, G_j) = s(x_i, x_j)(1 \leq i, j \leq n)$（あるいは非類似度を$d(G_i, G_j) = d(x_i, x_j)$）．クラスタの数：$C = n$，クラスタの番号：$index = n + 1$.

階層併合的クラスタリング(2/2)

2. 類似度最大（非類似度最小）のクラスタ対を結合する．つまり，$s(G_q, G_r) = max_{i,j} s(G_i, G_j)$（または$d(G_q, G_r) = min_{i,j} d(G_i, G_j)$）となる$G_q, G_r$について，$G_{index} = G_q \cup G_r$とし，$G$から$G_q$と$G_r$を除き，$G_{index}$を加える．$C := C$-1, $index := index$+1.
3. もしC=1ならば，終了．C>1ならば，$i < index$なるすべてのG_iについて，$s(G_{index}, G_i)$（$d(G_{index}, G_i)$）を再計算し，ステップ2へ．

- 併合後の類似度（非類似度）を併合されたクラスタ内の事例と，他のクラスタ内の事例との類似度の最大値（非類似度の最小値）とする．
- 併合された新しいクラスタをG_{index}とすると，併合されて除去された番号以外の$i < index$なるクラスタについて

$$\mathrm{s}(G_{index}, G_i) = \max_{x \in G_{index}, y \in G_i} s(x, y)$$

$$\mathrm{d}(G_{index}, G_i) = \min_{x \in G_{index}, y \in G_i} d(x, y)$$

最短距離法の例

	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

データ

	G_1	G_2	G_3	G_4	G_5
G_1	0	2	32	9	16
G_2	2	0	18	5	10
G_3	32	18	0	16	16
G_4	9	5	16	0	1
G_5	16	10	16	1	0

非類似度行例

$s_{ij} = \left({x_i}^a - {x_j}^a\right)^2 + \left({x_i}^b - {x_j}^b\right)^2$

ただし，${x_i}^a$はx_iの属性aの値を表す．

最短距離法の例

- 最小非類似度を示すG_4とG_5が併合され，$G_6 = G_4 \cup G_5$を作る．

$$\mathrm{d}(G_6, G_i) = \min_{x \in G_6, y \in G_i} d(x, y)\ (\mathrm{i} < 6, \mathrm{i} \neq 4,5)$$

	G_1	G_2	G_3	G_6
G_1	0	2	32	9
G_2	2	0	18	5
G_3	32	18	0	16
G_6	9	5	16	0

最短距離法の例

- 最小非類似度を示すG_1とG_2が併合され，$G_7 = G_1 \cup G_2$を作る．

$$\mathrm{d}(G_7, G_i) = \min_{x \in G_7, y \in G_i} d(x, y)\ (\mathrm{i} < 7, \mathrm{i} \neq 1,2,4,5)$$

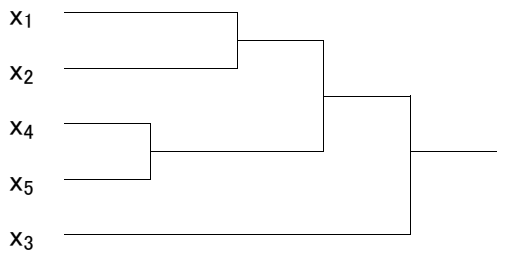
	G_7	G_3	G_6
G_7	0	18	5
G_3	18	0	16
G_6	5	16	0

最短距離法の例

- 最小非類似度を示すG_6とG_7が併合され，
  $G_8 = G_6 \cup G_7$を作る．
  $\mathrm{d}(G_8, G_i) = \min_{x \in G_8, y \in G_i} d(x, y)$ (i < 8, i ≠ 1,2,4,5,6,7)

	G_8	G_3
G_8	0	16
G_3	16	0

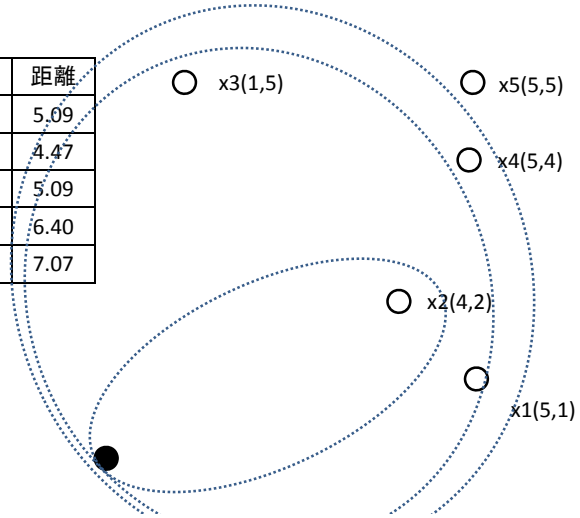
(dendrogram)



2グループに分けるなら：{x_3},{x_1,x_2,x_4,x_5}
3グループに分けるなら：{x_3},{x_1,x_2},{x_4,x_5}
4グループに分けるなら：{x_3},{x_1},{x_2},{x_4,x_5}

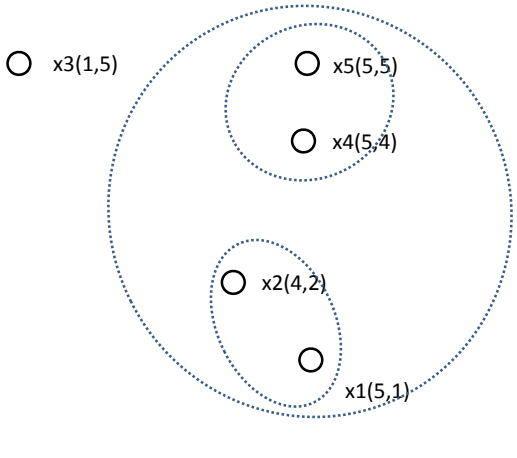
ソーティング

	属性a	属性b	距離
x_1	5	1	5.09
x_2	4	2	4.47
x_3	1	5	5.09
x_4	5	4	6.40
x_5	5	5	7.07



階層併合的クラスタリング

	属性a	属性b	距離
x_1	5	1	5.09
x_2	4	2	4.47
x_3	1	5	5.09
x_4	5	4	6.40
x_5	5	5	7.07



- 併合後の類似度（非類似度）を併合されたクラスタ内の事例と，他のクラスタ内の事例との類似度の最小値（非類似度の最大値）とする．
- 併合された新しいクラスタをG_{index}とすると，併合されて除去された番号以外の$i < index$なるクラスタについて

$$\mathrm{s}(G_{index}, G_i) = \min_{x \in G_{index}, y \in G_i} s(x, y)$$

$$\mathrm{d}(G_{index}, G_i) = \max_{x \in G_{index}, y \in G_i} d(x, y)$$

- 併合後の類似度（非類似度）を併合されたクラスタ内の事例と，他のクラスタ内の事例との類似度（非類似度）の平均値とする．
- 併合された新しいクラスタをG_{index}とすると，併合されて除去された番号以外の$i < index$なるクラスタについて

$$\mathrm{s}(G_{index}, G_i) = \frac{1}{|G_{index}||G_i|} \sum_{x \in G_{index}, y \in G_i} s(x, y)$$

$$\mathrm{d}(G_{index}, G_i) = \frac{1}{|G_{index}||G_i|} \sum_{x \in G_{index}, y \in G_i} d(x, y)$$

- あらかじめクラスタの数を指定し，事例を与えられた基準に基づいて，クラスタ内に割り当てる方法である．

1. 初期値として，クラスタ数と初期の事例の分割（クラスタ）を与える．
2. データの分割に基づいて各分割の重心を求め，重心と各事例との距離を求める．
3. 各事例を最も近いクラスタに割り当てる．
4. もし上記割り当てが前回の割り当てと同じであれば終了．更新があれば，ステップ2へ．

- 初期分割によって結果は変わることがある．

K-means法の例1

	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2，初期データ分割をC_1={x_1,x_2,x_3}, C_2={x_4,x_5}とする．

C_1とC_2の重心はそれぞれ

$$\left(\frac{5+4+1}{3}, \frac{1+2+5}{3}\right) = (3.33, 2.67)$$

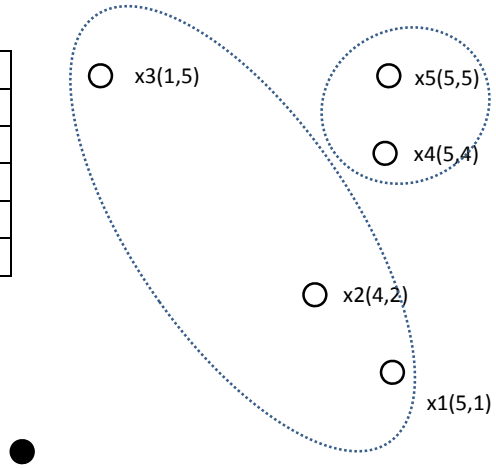
$$\left(\frac{5+5}{2}, \frac{4+5}{2}\right) = (5, 4.5)$$

	C_1	C_2
x1	2.36	3.5
x_2	0.94	2.69
x_3	3.29	4.03
x_4	2.13	0.5
x_5	2.86	0.5

移動する事例がないので，この分割で終了となる．

K-means

	属性a	属性b	距離
x_1	5	1	5.09
x_2	4	2	4.47
x_3	1	5	5.09
x_4	5	4	6.40
x_5	5	5	7.07



K-means法の例2(1/3)

	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2，初期データ分割を
C_1={x_1,x_3}, C_2={x_2,x_4,x_5}とする.

C_1とC_2の重心はそれぞれ

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

$$\left(\frac{4+5+5}{3},\frac{2+4+5}{3}\right)=(4.67,3.67)$$

	C_1	C_2
x1	2.82	2.69
x_2	1.41	1.79
x_3	2.82	3.90
x_4	2.23	0.46
x_5	2.82	1.37

x_1はC_2へ，x_2はC_1へ移動させる.

K-means法の例2(2/3)

	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

データ分割はC_1={x_2,x_3}, C_2={x_1,x_4,x_5}となる.

C_1とC_2の重心はそれぞれ

$$\left(\frac{4+1}{2},\frac{2+5}{2}\right)=(2.5,3.5)$$

$$\left(\frac{5+5+5}{3},\frac{1+4+5}{3}\right)=(5,3.33)$$

	C_1	C_2
x1	5.44	12.5
x_2	2.78	4.5
x_3	18.78	4.5
x_4	0.44	6.5
x_5	2.78	8.5

x_4 ,x_5はC_1へ，x_3はC_2へ移動させる.

K-means法の例2(3/3)

	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

データ分割はC_1={x_1, x_2,x_4,x_5}, C_2={x_3}となる.

C_1とC_2の重心はそれぞれ

$$\left(\frac{5+4+5+5}{4},\frac{1+2+4+5}{4}\right)=(4.75,3)$$

(1,5)

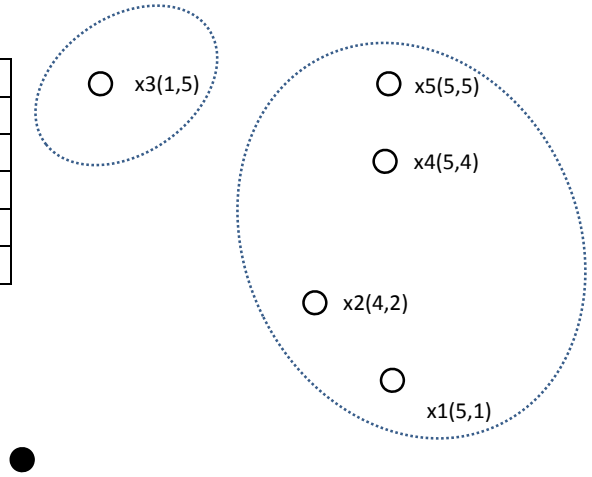
	C_1	C_2
x1	2.01	5.65
x_2	1.25	4.24
x_3	4.25	0
x_4	1.03	4.12
x_5	2.01	4

移動する事例がないので，これで終了する.

初期データの分割によって，結果が異なる.

K-means

	属性a	属性b	距離
x_1	5	1	5.09
x_2	4	2	4.47
x_3	1	5	5.09
x_4	5	4	6.40
x_5	5	5	7.07



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.