クラスタリング入門

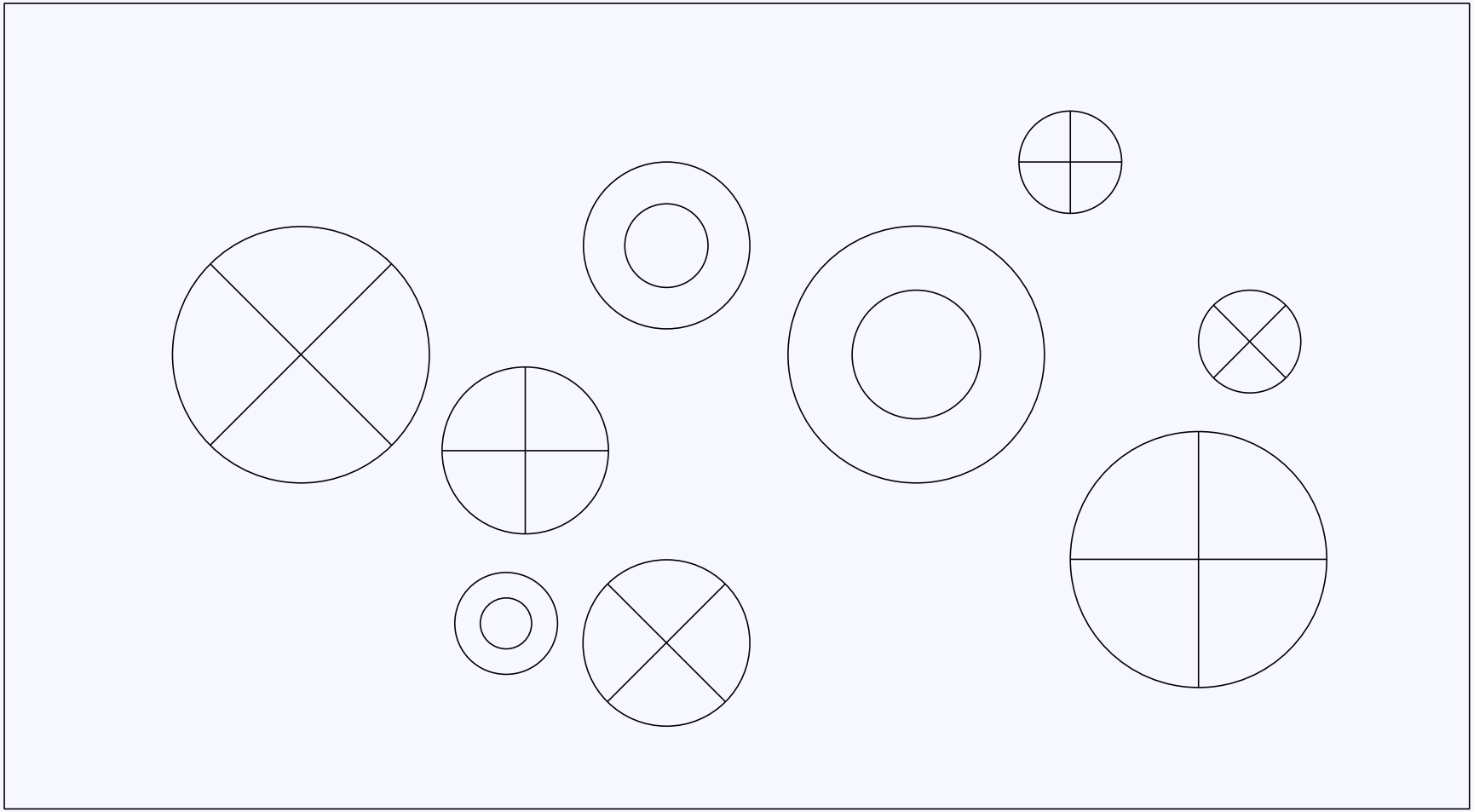
クラスタリングの概念
結果に対する評価基準

４月１５日（火）
田中洸一

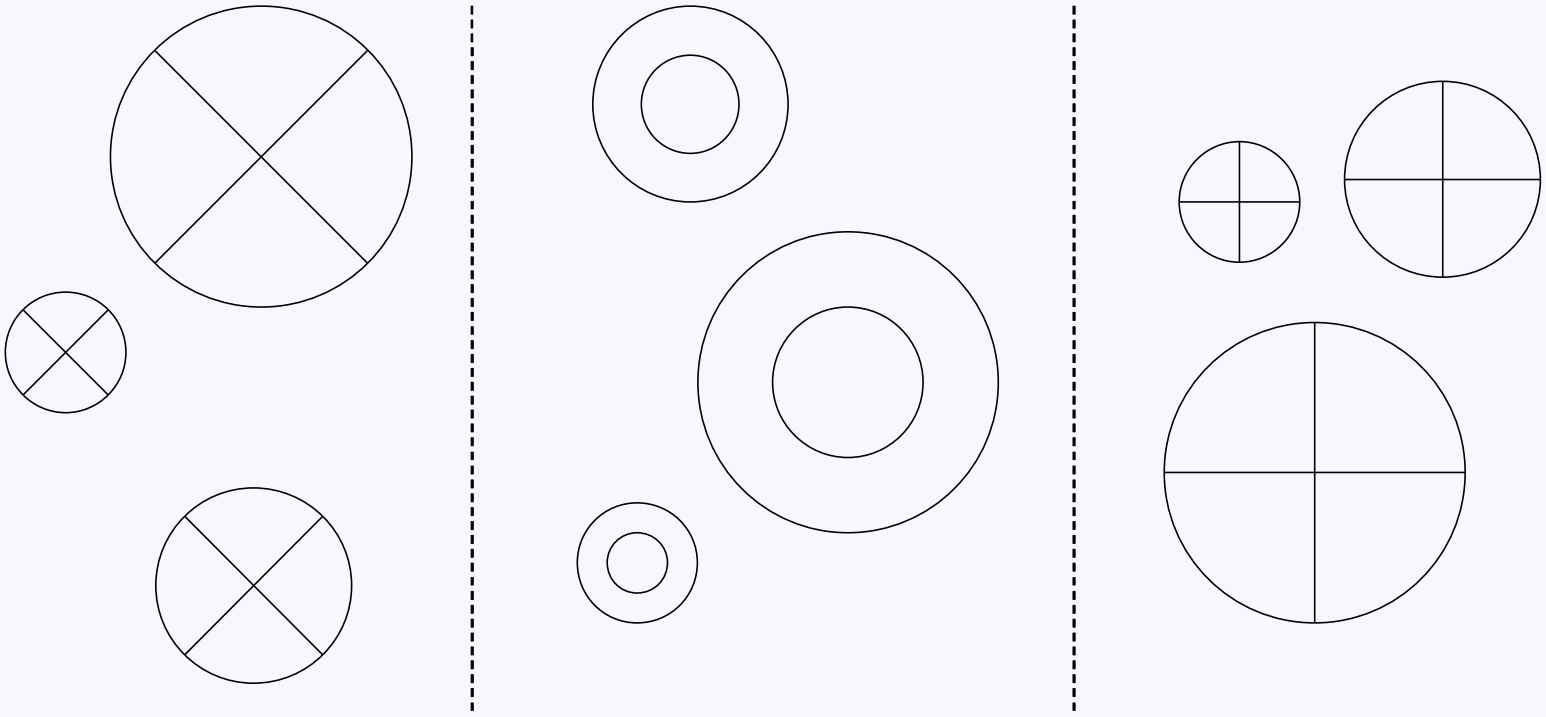
クラスタリングとは

- データの集まりをデータ間の類似度（あるいは秘類似度）に従って、いつかのグループに分けること
- どのような観点で類似度を設定するかでクラスタリングの結果は異なるので、クラスタリングには厳密な正解はない

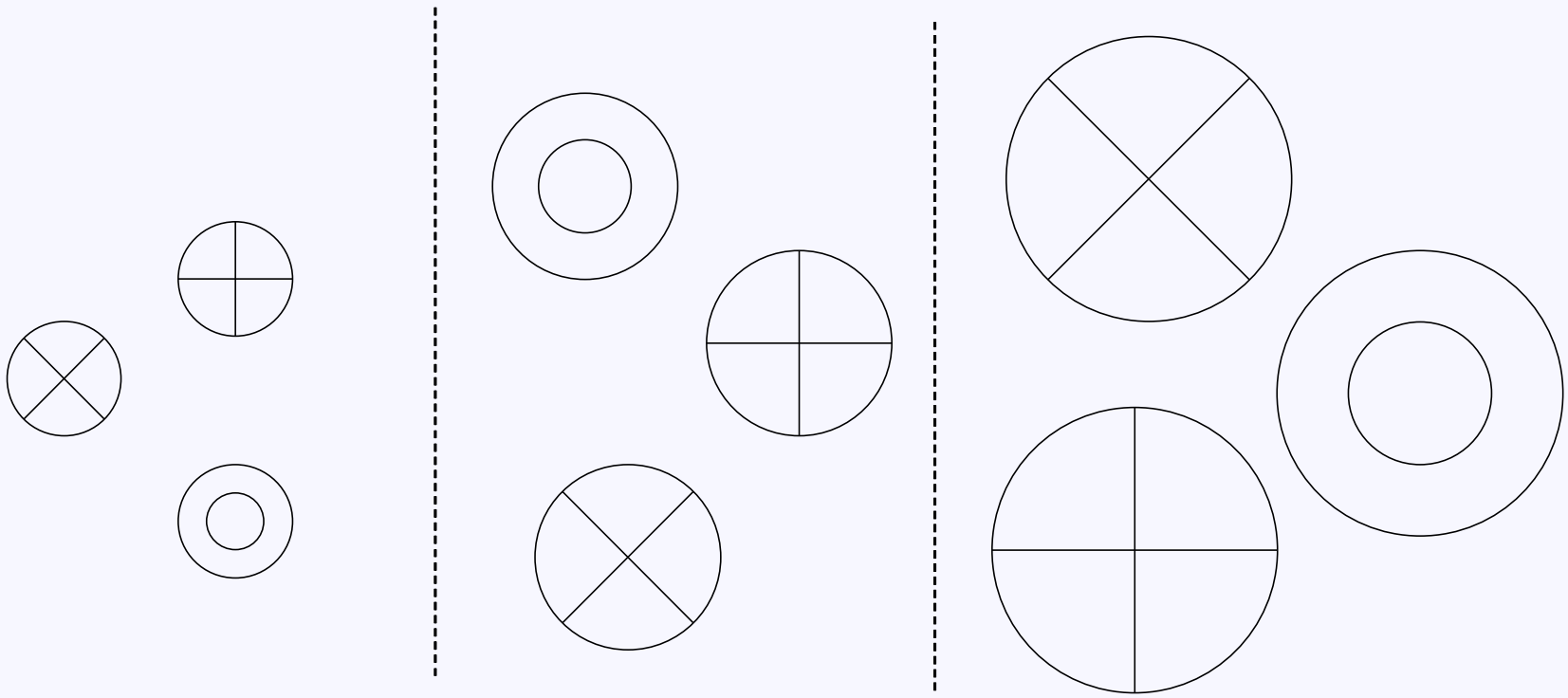
クラスタリングの例



模様による分類



大きさによる分類



クラスタリング手法の概要

- クラスタリングの手法は、２つの観点から分類される
- １つめの観点は、階層的手法と非階層的手法
- ２つめの観点は、ソフトクラスタリングとハードクラスタリング

階層的手法とは

- 各データを1つのクラスタとして、クラスタ間の距離や類似度に基づいてクラスタを逐次的に併合していく手法
- 通常は1つのクラスタになるまで併合を繰り返す
- データの階層構造が得られ、デンドログラムで表現される

階層的手法の種類

- 単連結法
- 完全連結法
- 群平均法
- ウォード法
- 重心法
- メディアン法

非階層的手法とは

- データの分割の良さを表すある評価関数によって、最適解を探索する手法
- 階層的手法ではデータが多いと階層構造が複雑になってしまうため、非階層的手法のほうが実用的
- k-meansが代表的な手法

非階層的手法の種類

- k-means
- 混合分布モデル
- スペクトラルクラスタリング
- pLSI
- NMF
- Fuzzy c-means

ソフトクラスタリングとは

- データが複数のクラスタに属することを許すクラスタリング手法
- データが各クラスタに属する度合い（帰属度）が出力となる

ソフトクラスタリングの種類

- Fuzzy c-means
- 混合分布モデル
- pLSI
- NMF

ハードクラスタリングとは

- データがある1つのクラスタに属する形で出力されるクラスタリングを手法
- 一般的にクラスタリングといえばこちらのほうをさす

ハードクラスタリングの種類

- 単連結法
- 完全連結法
- 群平均法
- ウォード法
- 重心法
- メディアン法
- k-means
- スペクトラルクラスタリング

クラスタリング結果の評価

- 正解集合が用意されているか、クラスタの数が与えられているかなどの条件によって、評価方法がいくつかある
- どの方法にも長所と短所があり、標準的な評価方法は確立されていない

評価方法とクロス表

- 今回は正解集合が存在し、クラスタ数が既知である場合の評価方法（F値、エントロピー、純度、精度）とこれらの評価値を算出するためのクロス表を示す

クロス表（1）

- クラスタリングの結果C

  $C = \{C_1, C_2, \cdots, C_k\}$

- 正解となるクラスタリングA

  $A = \{A_1, A_2, \cdots, A_k\}$

- X_{ij}を$|C_i \cap A_j|$つまりCiとAjに共通に属するデータの個数とし、クロス表を作成する

クロス表(2)

	A_1	A_2	…	A_j	…	A_k
C_1	X_{11}	X_{12}	…	X_{1j}	…	X_{1k}
C_2	X_{21}	X_{22}	…	X_{2j}	…	X_{2k}
…	…	…	…	…	…	…
C_i	X_{i1}	X_{i2}	…	X_{ij}	…	X_{ik}
…	…	…	…	…	…	…
C_k	X_{k1}	X_{k2}	…	X_{kj}	…	X_{kk}

エントロピー（１）

- エントロピーは最も標準的に用いられている評価尺度
- 各クラスタC_iに対するエントロピーE_i

$$E_i = -\sum_{h=1}^{k} P(A_h \mid C_i) \log P(A_h \mid C_i)$$

エントロピー（２）

- クラスタのデータ数による重み付き平均によって全体のエントロピーを定義（Nはデータ数）
- この値は0から1の値をとり、値が低いほどクラスタリング結果がよい

$$\sum_{i=1}^{k} \frac{|C_i|}{N} E_i = \sum_{i=1}^{k} \frac{\sum_{j=1}^{k} x_{ij}}{N} E_i$$

純度（1）

- ある正解のクラスタのデータをどの程度含むかという指標

$$P_i = \frac{1}{|C_i|} \max_h |C_i \cap A_h|$$

純度（2）

- 各クラスタのデータ数による重み付き平均をとることで純度を定義

$$\sum_{i=1}^{k} \frac{|C_i|}{N} P_i = \frac{1}{N} \sum_{i=1}^{k} \max_h |C_i \cap A_h|$$

- 0から1の値をとり、値が高いほどクラスタリングの結果がよい

F尺度（1）

- 再現率R_{hk}と精度P_{hk}

$$R_{hk} = \frac{|A_h \cap C_k|}{|A_h|} \qquad P_{hk} = \frac{|A_h \cap C_k|}{|C_k|}$$

- A_hとC_kに対するF尺度F_{hk}はR_{hk}とP_{hk}の調和平均

$$F_{hk} = \frac{2 R_{hk} P_{hk}}{R_{hk} + P_{hk}}$$

F尺度（2）

- クラスタリング結果に対するF尺度Fは、A_hに対して、F_{hk}が最大になるようなkを求めてF_{hk}を算出し、各hに対して重み付き平均をとったもの

$$F = \sum_{h=1}^{K} \frac{|A_h|}{N} \max_k F_{hk}$$

精度（1）

- 精度はもっとも厳格な評価尺度だが、クラスタとクラスタのラベルを対応させる必要がある
- クラスタの番号をクラスタのラベルに対応させる関数をfとすると、各クラスタC_iの精度AC_iを求める

$$AC_i = \frac{|A_{f(i)} \bigcap C_i|}{|C_{f(i)}|} = \frac{X_{if(i)}}{\sum_{j=1}^{K} X_{f(i)j}}$$

精度（2）

- 各クラスタに対する精度のデータ数による重み付き平均をとり、全体の精度を定義

$$\sum_{i=1}^{K} \frac{|C_i|}{N} AC_i$$