数理統計 補助資料
～クラスタリング～

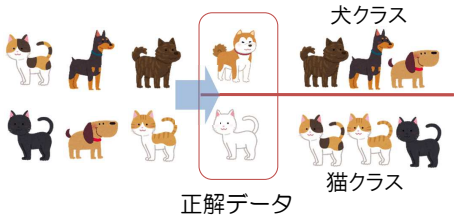
2023年度2学期:　月曜1限, 水曜3限
担当教員:　石垣　司

クラスタリング

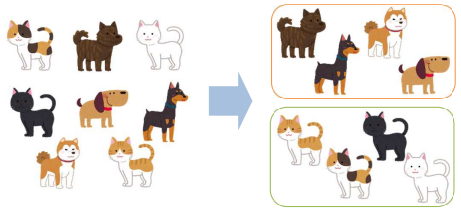
似た傾向を示すデータをまとめる方法の総称

- **データ間の類似度等を利用して, 何らかのグループやパターンを発見・分類する**

～判別分析：犬と猫に分けて～



～クラスタリング：良い感じに分けて～



本日の講義で紹介する手法

- **構造化データのクラスタリング：クラスター分析**, k-means**法**
- **テキストデータ(非構造化データ)のクラスタリング：トピックモデリング**

機械学習手法の分類の名称

教師あり学習：正解データを用いる手法の分類の名称

- **例：回帰分析, 判別分析**
  目的変数 y を正解データとして回帰直線や判別直線を学習

教師なし学習：正解データを用いない手法の分類の名称

- **例：主成分分析**
  多変量データの次元削減。削減次元の射影先の正解データはない
- **例：ベイジアンネットワーク**
  多変量間のネットワーク構造を学習。ネットワーク構造の正解データは学習に利用しない
- **例：クラスタリング**
  データ間の類似度などに基づいて何らかのグループやパターンの発見・分類。グループやパターンの正解データはない

クラスター分析

階層的クラスタリング手法

- **距離が近いデータ同士をまとめる**

デンドログラム(樹形図)による可視化

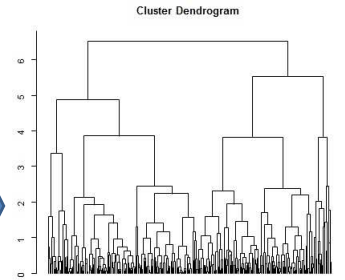
- **作成手順**
  1. **データから距離行列を作成**
  2. **距離の近い順にクラスタ化**

	age	gender	income	kids	ownHome
Aさん	47	M	49482	2	ownNo
Bさん	31	M	35546	1	ownYes
Cさん	43	M	44169	0	ownYes
Dさん	37	F	81041	1	ownNo
Eさん	40	F	79353	3	ownYes

元データ

	Aさん	Bさん	Cさん	Dさん	Eさん
Aさん	0				
Bさん	0.25	0			
Cさん	0.23	0.06	0		
Dさん	0.26	0.41	0.42	0	
Eさん	0.41	0.31	0.29	0.22	0

データ間の距離行列

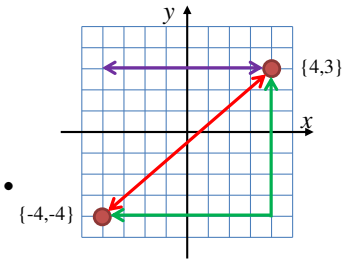


デンドログラム（N=300）

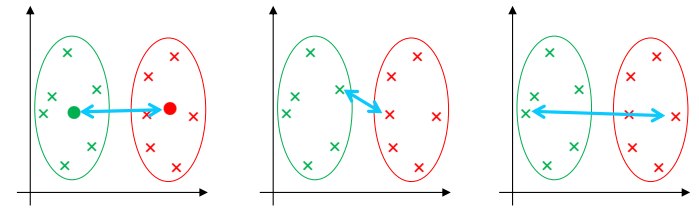
クラスター分析で用いる距離の定義

データ間の距離の一例

- ↔ユークリッド距離 (≒10.63)
- ↔最大距離 (=8)
- ↔マンハッタン距離 (=15) などなど・・・



クラスタ間の距離の一例



重心距離
2クラスタ間の平均間の距離

最短距離
最も近い対象の距離

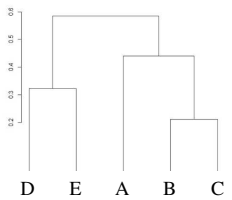
最長距離
最も遠い対象の距離

その他にも
- メディアン法
- McQuitty法
- ウォード法

などなど・・・

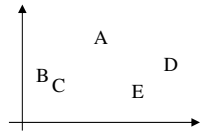
クラスター分析のアルゴリズム

1. 全てのクラスタ間距離を計算
2. 最も距離の小さい2つをクラスタ化
3. 全クラスタ数が1なら終了 (Otherwise, 1 に戻る)



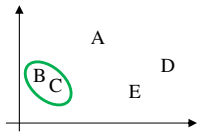
初期状態

	A	B	C	D	E
A	0				
B	0.25	0			
C	0.23	0.06	0		
D	0.26	0.41	0.42	0	
E	0.41	0.31	0.29	0.22	0



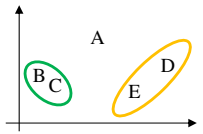
BとCのクラスタ化

	A	B	C	D	E
A	0				
B	0.25	0			
C	0.23	0.06	0		
D	0.26	0.41	0.42	0	
E	0.41	0.31	0.29	0.22	0



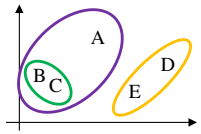
DとEのクラスタ化

	A	{B,C}	D	E
A	0			
{B,C}	0.24	0		
D	0.26	0.41	0	
E	0.41	0.30	0.22	0



Aと{B,C}のクラスタ化

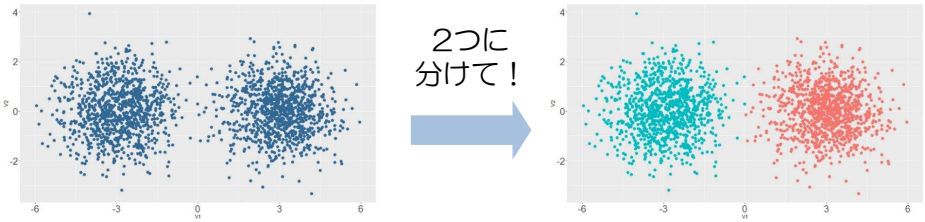
	A	{B,C}	{D,E}
A	0		
{B,C}	0.24	0	
{D,E}	0.33	0.35	0



k-means法

非階層的クラスタリング

- 距離に基づくクラスタリング
- 入力：クラスタ数。出力：各データのクラス
- ハードクラスタリング
  - 各データが所属するクラスを与える
  - ソフトクラスタリング：各データが各クラスに所属する確率を与える

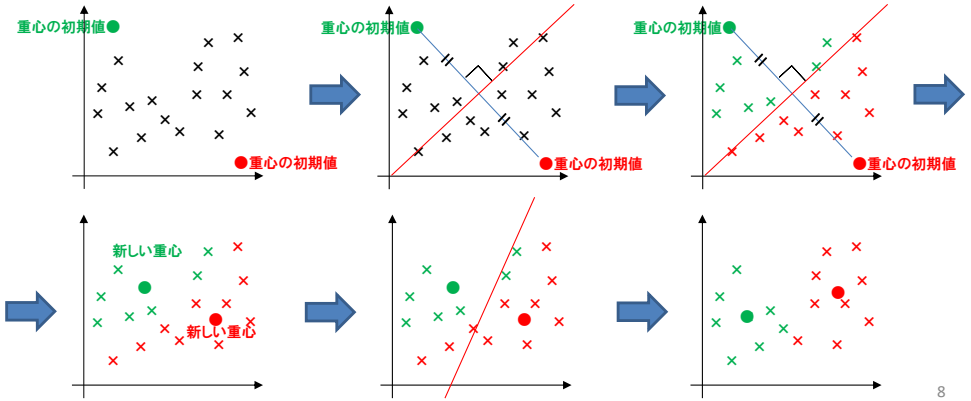


k-means法のアルゴリズム

初期設定：クラスタ数, 各クラスタの重心の初期値

結果が収束するまで以下を繰り返し

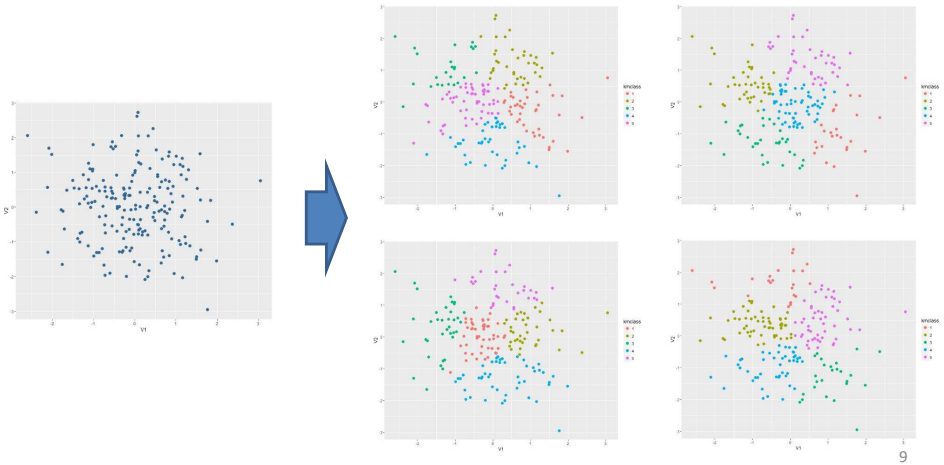
1. 各データ i を, 最も近い重心のクラスタへ分類
2. 各クラスタに含まれたデータから新しい重心を計算



k-means法の注意事項 #1

初期値により結果が(比較的大きく)変化

- **例: クラスタ数5の設定で異なる初期値でクラスタリング**



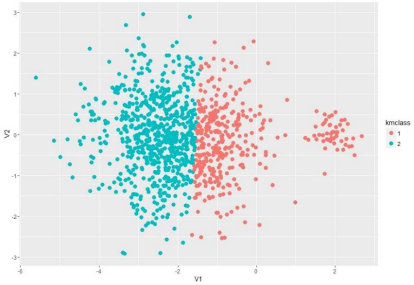
k-means法の注意事項 #2

暗黙の仮定1:「各クラスタは円状(超球状)に分布」

暗黙の仮定2:「各クラスタ内のデータ数は大体同じ」

失敗例1
「各クラスタは球状に分布」

失敗例2
「各クラスタ内のデータ数は大体同じ」



2つのクラスタリング法の特徴

簡単, 単純, アルゴリズムを理解しやすい

古典的かつ単純なクラスタリング法

- **良い意味でも単純だが, 悪い意味でも単純なアルゴリズム。実応用の場面では, 十分な結果が得られないことも多い**
- **適切なクラスタリング数の検証が難しい**
  学習のキーワード: 有限混合モデル, ノンパラメトリックベイズモデル
- **クラスター分析の注意点**
  距離の定義で結果が変化。*N* が大なら計算量が大きくなる($O(N^2)$)
- ***k*-means 法の注意点**
  初期値の影響で結果が変化。データ内容の理解や正規化が必要

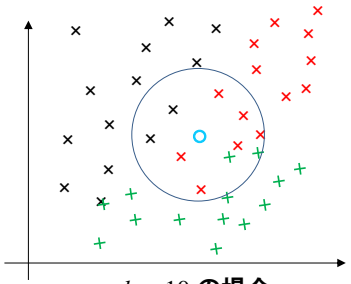
補足: *k*-近傍法 #1

非線形・多群の判別・分類の手法

- ***k*-nearest neighbor, *k*-NN**
  ***k*-means(*k*-平均法)と混同されやすいが別手法**
- **教師あり学習(簡単, 単純, 理解しやすい, 判別精度が良い)**

アルゴリズム

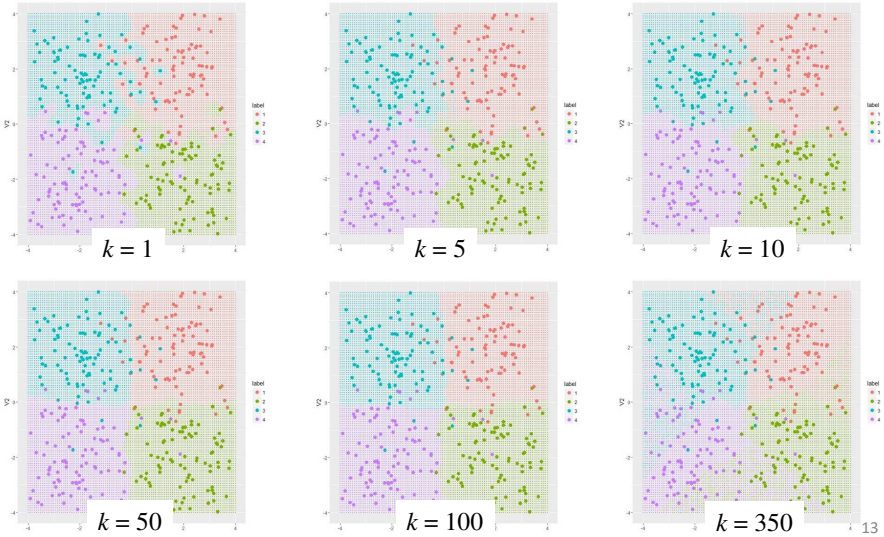
- **目的: 未知データ○の判別**
- **入力: *k* の値**
- **点○の近傍 *k* 個のデータのラベルを数え, 最も多いラベルを判別結果とする**



***k* = 10 の場合**
データ○はクラス赤×へ分類

補足：k-近傍法 #2

k の値で結果が変化



テキストデータのクラスタリング

現在のビジネス環境での応用例：
大量のテキストデータの中からユーザーの声をクラスタリング&可視化し，情報を抽出

- **「大量」：小量のデータ（数百件程度）では人間で処理可能。しかし，数千，数万件のデータの処理は人間には限界あり**
- **「テキストデータ」：非構造化データでありテキストデータ独自の処理が必要**

“大量”の“テキストデータ”　　**ユーザーの声の可視化**

構造化データと非構造化データ #1

構造化データ

- **スプレッドシートに格納された数値や文字列の各行や各列の要素の比較や計数に意味があるデータ**
- **多変量解析（線形回帰，線形判別，主成分分析，ベイジアンネットワークなど）で利用するデータは構造化データ**

構造化データ

生徒 No.	国語	数学	理科	社会	英語	音楽	体育
1	83	60	55	81	90	50	93
2	70	80	78	80	55	44	59
3	50	90	95	70	80	80	49
4	60	44	44	99	78	73	30
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

中学生の成績データを格納したスプレッドシート

- No.1の理科と英語のスコアの差は45点である
- No.1とNo.4の理科のスコアの差は40点である
- No1のスコアの平均点は約73点である
- 国語の全生徒のスコアの平均点は83点である

このような操作が意味をもつ

構造化データと非構造化データ #2

非構造化データ

- **構造化データ以外のデータの総称**
- **具体例：テキスト，画像，動画，音声，センサーデータ**
- **特徴：スプレッドシートに格納された数値や文字列の各行や各列の要素の比較や計数に多くの場合で意味がない**

例：テキストデータ

題名	1	2	3	4	5	6	…
雪国	国境	の	長い	トンネル	を	抜ける	…
細雪	こいさん、	頼む	わ。	鏡	の	中	…
人間失格	私	は、	その	男	の	写真	…
羅生門	ある	日	の	暮方	の	事	…
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

- 雪国の3番目の単語と5番目の単語を比較する
- 雪国と細雪の3番目の単語を比較する
- 雪国の単語の平均値を計算する
- 全作品の1番目の単語を比較する

このような操作に意味がない

各データに特有の分析手法を理解するが必要ある

分析の前に～形態素分析

日本語の文章を各単語に分解する作業

- **日本語のテキスト分析には必須となる処理**
- **英語のテキスト分析では不必要な処理**
- **例:「すももももももものうち」**

```
すもも  名詞,一般,*,*,*,*,すもも,スモモ,スモモ
も  助詞,係助詞,*,*,*,*,も,モ,モ
もも  名詞,一般,*,*,*,*,もも,モモ,モモ
も  助詞,係助詞,*,*,*,*,も,モ,モ
もも  名詞,一般,*,*,*,*,もも,モモ,モモ
も  助詞,係助詞,*,*,*,*,も,モ,モ
の  助詞,連体化,*,*,*,*,の,ノ,ノ
うち  名詞,非自立,副詞可能,*,*,*,うち,ウチ,ウチ
```


分析の前に～WordCloud

文書内の単語の出現頻度に応じて, 単語の大きさを変化させて可視化する手法

「坊ちゃん」
（夏目漱石）

「細雪（上巻）」
（谷崎潤一郎）

トピックモデリング

大量のテキストデータ群の単語や文書をクラスタリングすることで複数のトピックを抽出する手法の総称

- **自然言語処理の分野で発展。レビュー, ニュース, 論文データ分析, 購買行動分析, マーケティング, 遺伝子分析などへ応用**

潜在ディリクレ配分法(LDA: Latent Dirichlet Allocation)

- **従来のPLSI法をベイズモデルへ拡張。文書-トピック-単語の間の相関関係や付加的な情報の付与などへの拡張が可能な柔軟なモデル**
- **2003年に提案されて以降, 様々な亜種が提案されている**
- **トピックの数の設定と抽出されたトピックの解釈は人間が行う**

LDAの数理モデル #1

Notation

- D: **文書の数**, K: **トピックの数**, W: **語彙の数**
- w_{di}: **文書** d **の** i **番目の単語(データ:** $w_{di} = \{1, \cdots, W\}$**)**
- z_{di}: **文書** d **の** i **番目の単語の割当トピック(**$z_{di} = \{1, \cdots, K\}$**)**

LDAのモデル

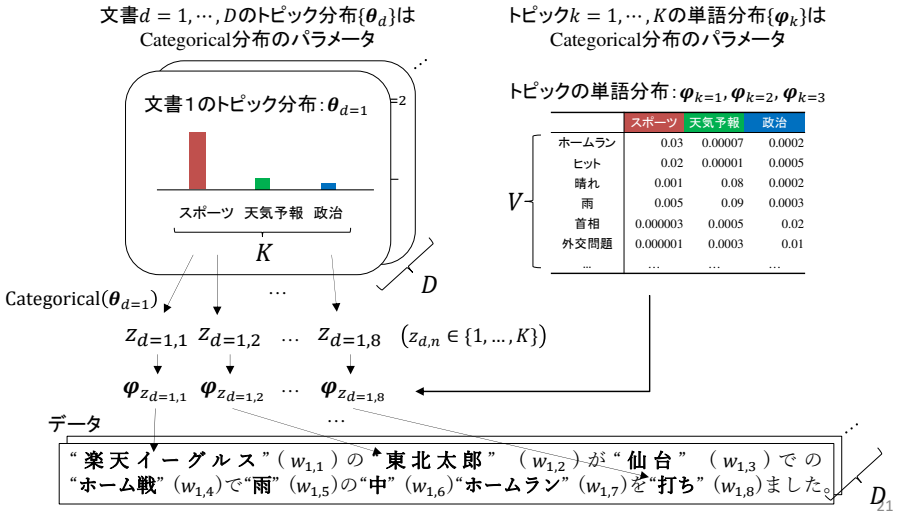
1. **文書** d **は** $\boldsymbol{\theta}_d = \{\theta_1, \cdots, \theta_K\}$ **のトピック確率を持つ**
2. **トピック** k **は** $\boldsymbol{\varphi}_k = \{\varphi_1, \cdots, \varphi_W\}$ **の単語確率を持つ**
3. **単語** w_{di} **のトピック** z_{di} **はトピック確率** $\boldsymbol{\theta}_d$ **に従い決まる**
4. **文書** d **の単語** w_{di} **は単語確率** $\boldsymbol{\varphi}_{k=z_{di}}$ **に従い決まる**

単語データ $\{w_{di}\}$ **からパラメータ** $\{\boldsymbol{\theta}_d, \boldsymbol{\varphi}_k, z_{di}\}$ **をベイズ推定**

#メモ 数理的内容は学部2年生の範囲を大きく超えるため詳細は説明できない

LDAの数理モデル #2

LDAのモデルの概要（パラメータの推定法の詳細は割愛）



トピックモデルの結果の例 #1

使用データ

- TIS株式会社により無償公開されている有価証券報告書のデータセット「chABSA-dataset」
  - 上場企業 230 社の2016年の有価証券報告書の一部のテキストデータ
  - 1社につき 6 ~ 252 の文章数
  - 全文章中から70文字以上の2,570文書を抽出
- テキストデータの具体例：

「当連結会計年度におけるわが国経済は、企業収益や雇用環境など底堅く推移しているものの、英国のEU離脱等・・・」

「当社グループの所属する映像関連業界におきましては、技術革新に伴う映像メディアの変化や映像制作工程の変化・・・」

などなど

chABSA-dataset (Creative Commons Attribution 4.0 License ~ CC BY 4.0 LEGAL CODE)
https://www.tis.co.jp/news/2018/tis_news/20180410_1.html, https://github.com/chakki-works/chABSA-dataset

トピックモデルの結果の例 #2

トピック数を 6 とした LDA によりトピックを作成

ビッグデータ分析の有用性と注意点

ビッグデータ分析の有用性

- 大量データから小量データでは見えにくい何らかの情報やパターンを抽出する

テキストビッグデータ分析の注意点

- バイアスを含まない標本抽出は多くの場合で満たされない。分析の目的を意識した使い分けが必要

ほとんどのユーザーは商品購入時にレビューをしない（Amazon は非公開だが, Amazon.com での review rate は1~2% [Landing Cube 2021] ）

レビューサイト上の意見を集約して消費者全体の感想を把握することができるか？

商品のレビューサイト上の意見から, その商品への不満点を把握することができるか？

演習問題

次の *k*-means 法を用いたクラスタリングが失敗している理由を説明しなさい