

G検定対策講座 第8回

教師なし学習

正解を教わずに、隠れた構造を見つける



今日学ぶこと

1

教師なし学習とは

特徴量だけで、隠れた構造を見つける

2

クラスタリング

k-means法・ウォード法

3

次元削減

PCA・SVD・MDS・t-SNE

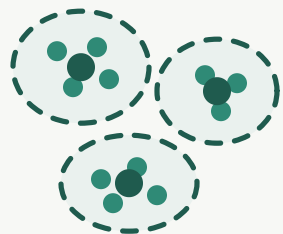
4

トピックモデルと推薦システム

LDA・協調フィルタリング・コールドスタート問題

正解のないデータから どう学ぶ？

世の中には、ラベルの付いていないデータのほうが圧倒的に多い



ラベルなし

教師なし学習とは

正解ラベルがない状態で、
パターンや構造を発見する方法



必要なのは特徴量だけ
教師あり学習との大きな違い

具体例

主な用途

クラスタリング

似たものどうしを
グループに分ける

次元削減

たくさんの情報を
少ない軸に圧縮する

応用

トピック抽出・異常検知
推薦システム

POINT

「教師なし学習は特徴量のみ。正解は与えない」——ここがよく問われる

教師あり と 教師なし

人間が正解を用意するかどうか、決定的な違い

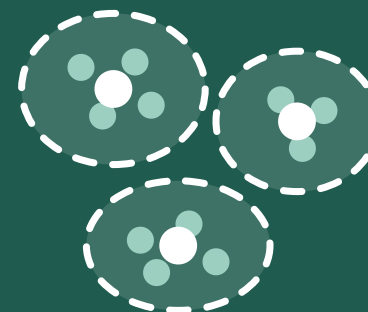


教師あり学習

特徴量 + 正解ラベルのペアで学ぶ

猫の画像に「猫」というラベルを付けて見せる

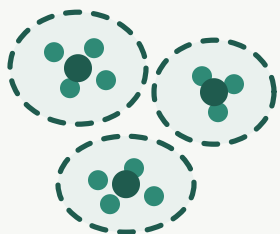
≠



教師なし学習

特徴量だけで学ぶ

データの隠れた構造を、自分の力で見つける



1967年 マックイーン

k-means法

重心をもとに、データをグループ分けする



グループの数 k は
人間が事前に決める

具体例 手順（重心を置き直しながら収束させる）

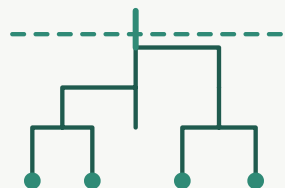
- 1 グループの数 k を人間が決める
- 2 各グループの重心を仮に置く
- 3 各データを近い重心に割り当てる
- 4 重心を置き直してくり返す



似たものが
グループにまとまる

POINT ビジネスでは顧客セグメンテーションによく使われる

1963年 ウォード



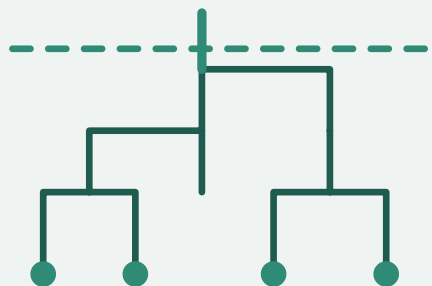
ウォード法

近いものから段階的にまとめ上げる
階層クラスタリング



クラスタの中のばらつきが
なるべく小さくなるように併合

具体例 結果はデンドログラム（樹形図）で表される



どこで切るかで
グループ数が決まる

k k-means法

重心ベース

グループ数は事前に指定

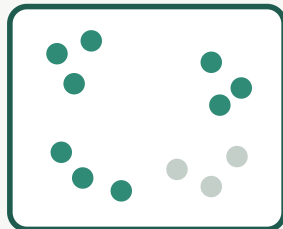
W ウォード法

階層的・デンドログラム

グループ数は

あとで決める

POINT この対比が頻出。k-means=重心ベース／ウォード=階層的



線形

主成分分析 (PCA)

情報をなるべく保ったまま、
少ない軸に圧縮する

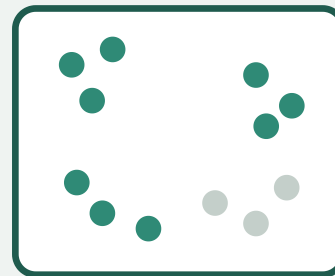


分散が最大になる方向を
新しい軸に選ぶ

具体例 5教科の成績を2つの軸にまとめる

国語・英語・社会は
いっしょに動きやすい

数学・理科も
セットで上下しやすい

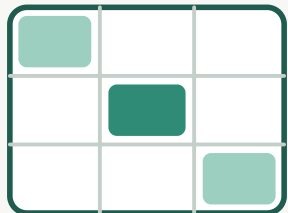


文系の軸と理系の軸
の2つに作り直す

5次元のデータが
2次元の散布図になり、全体
の傾向が見える

POINT

「PCAは線形の次元削減。分散が最大になる新しい軸にとる」と覚える



線形

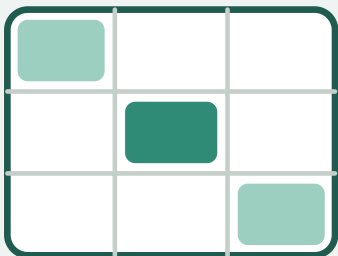
特異値分解 (SVD)

大きな表（行列）を、
重要な成分だけに圧縮する技術



実はPCAの数学的な基礎に
なっている

具体例 ユーザー × 映画の評価表を分解する



ユーザー × 映画
の巨大な表

ユーザーの好みの型

映画の型

データの圧縮・ノイズの除去

推薦システム・潜在意味解析
(LSA)

POINT 行列を圧縮する技術で、PCAの基礎——ここまで覚えれば十分。計算の手順は問われない

1952年 トーガソン



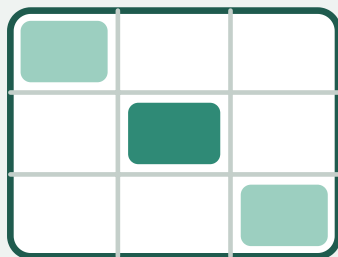
多次元尺度構成法 (MDS)

データどうしの距離の関係を保ったまま、低い次元に写す

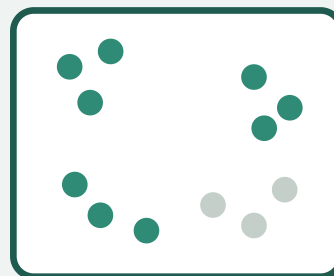


似ているものは近くに
違うものは遠くに配置される

具体例 入力は「データどうしの距離」



距離をまとめた
距離行列



位置関係を保って
2次元に配置



位置関係がそのまま残るので
可視化にぴったり

POINT PCAは分散の最大化、MDSは距離の保存。この対比を押さえる

2008年



t-SNE (ティー・エス・エヌ・イー)

非線形の次元削減で、可視化に特化した方法



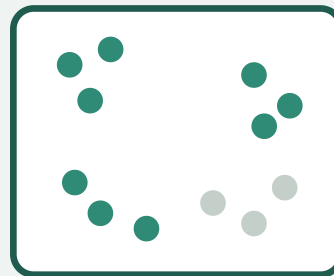
ファン・デル・マーテンと
ヒントンが発表

具体例

近所どうしの関係を保ったまま2次元へ

高次元データのローカルな構造
(ご近所どうしの関係)を保つ

2次元・3次元に落とし込む



Word2Vec で作った
単語のベクトル
を地図にする

ディープラーニングの
埋め込みの可視化で活躍

POINT PCAは線形で全体をとらえる／t-SNEは非線形でローカル——この対比が頻出

次元削減の4つの手法

名前と、ひとつのキーワードの対応だけ押さえれば十分

1

PCA

線形

分散が最大の方へ
前処理・要約に

2

SVD

線形

行列を分解して圧縮
PCAの基礎

3

MDS

距離を保存
位置関係を保つ
おもに可視化

4

t-SNE

非線形

ローカルな構造を保つ
DLの可視化で活躍

2003年



トピックモデル (LDA)

文書の集まりから、**潜在的なテーマ**を抽出する

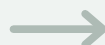


潜在的ディリクレ配分法
ブレイらが2003年に発表

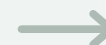
具体例 1つの文書は、複数トピックの混ざり合い



ニュース記事の
山



LDA



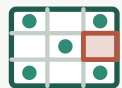
スポーツ

政治

料理

… 自動で発見

POINT LDA=トピックモデルの代表。ただし線形判別分析（同じLDA）は教師あり学習。混同しない



1992年～

推薦システム

教師なし学習の代表的な応用



協調フィルタリングは
1992年にゴールドバーグが提案

対比

2つのやり方

協 協調フィルタリング

ユーザーの類似を使う

「この商品を買った人はこちらも」

C コンテンツベース

アイテムの特徴を使う

「アクション映画が好きな人に別のアクション」

POINT 実際は両方を組み合わせたハイブリッド型がよく使われる



推薦の弱点

コールドスタート問題

履歴がないと推薦できないという問題



とくに深刻なのは
協調フィルタリング

具体例

2つの場面と、その対策

新 新規ユーザー

履歴がないので
似たユーザーを見つけられ
ない

新 新規アイテム

まだ誰も評価しておらず
推薦されない

人気アイテムを出す
簡単なアンケートをとる

コンテンツベースとの併用で緩和
できる

POINT コールドスタート=協調フィルタリングの弱点。セットで覚える

