データ処理・レポート

氏名：山北倫太郎
学籍番号：1423107

2025 年 7 月 6 日

Contents

1　クラスター分析とは

クラスター分析とは、ある集団の中から、互いに性質が似ているものを集めてグループ（クラスター）に分けるための統計的な分析手法です。データに内在する構造やパターンを明らかにすることを目的とした、教師なし学習の一種です。

分析の手順

一般的に、以下の手順で分析を進めます。

1. **変数の選定とデータ準備:** 分析目的に合わせて変数を決定し、必要に応じて尺度を揃えるための標準化などを行います。
2. **距離（非類似度）の計算:** 各データ点がどの程度似ていないかを定量的に示す「距離」を計算します。
3. **クラスターの形成:** 算出した距離に基づき、データをグループ分けします。本課題で用いる階層的手法などがこれにあたります。
4. **クラスター数の決定と結果の解釈:** 最適なクラスター数を決定し、各クラスターがどのような特徴を持つかを分析・解釈します。

距離の取り方

データ間の距離尺度には様々ありますが、本課題では**ユークリッド距離**を用います。これは、多次元空間における 2 点間の直線距離を示す、最も一般的な尺度です。

$$d(p,q) = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2}$$

2　USArrests データセットの分析

課題の指示に従い、R に組み込まれた `USArrests` データセットの「Murder」「Assault」「UrbanPop」の 3 変数を用いて分析を行いました。

2.1　(a) 階層的分類法によるクラスター分析と結論の考察

分析方法

- **データ間の距離:** 課題の指定に基づき、**ユークリッド距離**を用いました。
- **クラスター間の距離:** 複数のクラスターを併合する際の手法として、**ウォード法（Ward's method）**を選択しました。ウォード法は、併合によってクラスター内の情報の損失（平方和の増加量）が最小になるようにグループを形成する手法です。
- **データの前処理:** 各変数は測定単位が異なるため、分析前に全変数の値を平均 0、標準偏差 1 となるよう**標準化**を行いました。

結論と考察

分析の結果、アメリカ 50 州は犯罪率と都市化の度合いに基づき、性質の異なる 4 つのグループに明確に分類されました。

- **クラスター 1（南部・殺人型）:** 殺人発生率が最も高いグループ。アメリカ南部の州が多く含まれます。

- **クラスター 2（大都市・暴行型）**：暴行発生率が突出して高く、都市人口比率も高いグループ。大都市を抱える州が多く見られます。
- **クラスター 3（全米平均型）**：各指標が全米の平均値に近く、標準的な特徴を持つグループ。
- **クラスター 4（地方・平和型）**：殺人・暴行発生率が共に著しく低く、最も安全な州のグループ。

この結果から、各州の治安状況は単に「良い・悪い」だけでは測れず、「どのような犯罪が多いか」という質的な違いが存在することが示唆されます。クラスター分析を用いることで、こうしたデータに隠れた複雑な構造を客観的に明らかにすることができました。

2.2 (b) デンドログラムの作成

階層的クラスター分析の結果を可視化するために、以下のプログラムを用いてデンドログラム（樹形図）を作成しました。

プログラム

```r
# データの準備（読み込み、変数選択、標準化）
library(clustrd)
data(USArrests)
USArrests_selected <- USArrests[, c("Murder", "Assault", "UrbanPop")]
USArrests_scaled <- scale(USArrests_selected)

# 距離行列の計算とクラスタリングの実行
d <- dist(USArrests_scaled, method = "euclidean")
hc <- hclust(d, method = "ward.D2")

# デンドログラムの描画
plot(hc, hang = -1, cex = 0.6, main = "Dendrogram of USArrests")
rect.hclust(hc, k = 4, border = "red")
```

Listing 1: デンドログラム作成用 R コード

作成されたデンドログラム

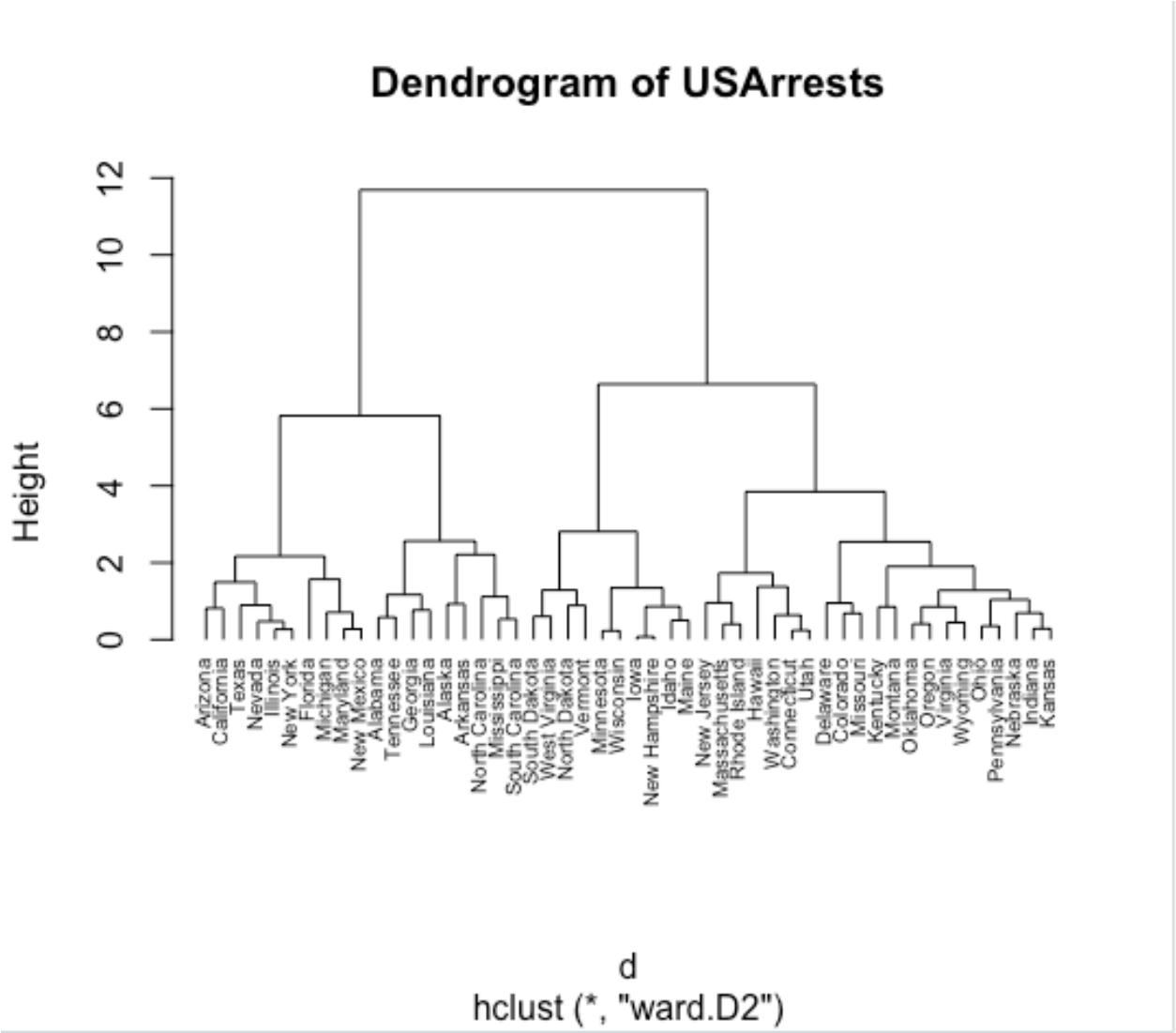

Figure 1: USArrests データセットのデンドログラム

2.3 (c) クラスター数 4 の場合の所属州と使用プログラム

クラスターの数を 4 つとした場合に、どの州がどのクラスターに属するかを調べるためのプログラムと、その実行結果を以下に示します。

プログラム

```r
# (b)で実行したクラスタリング結果(hc)を4つのクラスターに分割
clusters <- cutree(hc, k = 4)

# クラスター番号ごとに州の名前を一覧表示する
for (i in 1:4) {
```

```r
  cat("Cluster", i, ":\n")
  print(names(clusters[clusters == i]))
  cat("\n")
}
```

Listing 2: クラスター所属特定用 R コード

各クラスターの所属州

Table 1: クラスター分類結果

クラスター番号	所属する州
Cluster 1	Alabama, Arkansas, Georgia, Louisiana, Mississippi, North Carolina, South Carolina, Tennessee
Cluster 2	Alaska, Arizona, California, Florida, Illinois, Maryland, Michigan, Nevada, New Mexico, New York, Texas
Cluster 3	Colorado, Connecticut, Delaware, Hawaii, Indiana, Kansas, Kentucky, Massachusetts, Missouri, Montana, Nebraska, New Jersey, Ohio, Oklahoma, Oregon, Pennsylvania, Rhode Island, Utah, Virginia, Washington, Wyoming
Cluster 4	Idaho, Iowa, Maine, Minnesota, New Hampshire, North Dakota, South Dakota, Vermont, West Virginia, Wisconsin

3　分析の感想

今回の分析を通して、一見するとただの数値の羅列である統計データから、クラスター分析という手法を用いることで、意味のあるパターンや構造を浮かび上がらせることができる点を実感しました。特に、事前に何の正解も与えていないにもかかわらず、アメリカの州が我々の持つ地理的なイメージ（例えば「南部」や「大都市圏」）に近い形で自然にグループ分けされたことには、純粋な驚きがありました。

また、「犯罪率が高い」という単純な括りの中でも、殺人が多い「クラスター 1」と暴行が多い「クラスター 2」のように、その内実が異なるグループが形成されたことは、データ分析の奥深さを示しているように感じます。今回の分析は 3 つの変数のみで行いましたが、ここに所得水準や失業率、教育レベルといった新たな変数を加えることで、また違った側面から各州の姿が見えてくるのではないかと、さらなる分析への興味が湧きました。

4　参考文献

References

[1] 山田剛史, 杉澤正人 (2019). *R* による多変量解析. 株式会社オーム社.

[2] 脇本和昌 (2014). 多変量解析入門. 株式会社朝倉書店.