第３講 データの前処理とクリーニング

1. はじめに

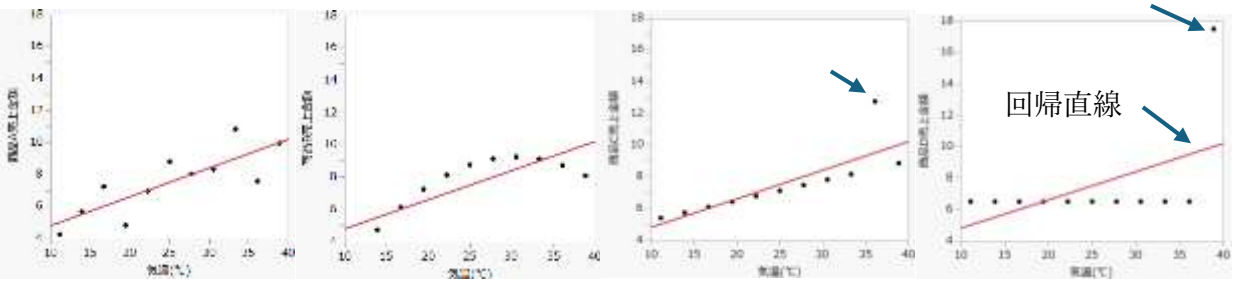
データサイエンスの最初のステップは「データの準備」です。 現実のデータには、抜けている値（欠損)、入力ミス（誤記)、同じデータの重複など、さまざまな問題が含まれています。こうした問題があるままでは、正しい分析や予測ができません。そこで必要になるのが「前処理」です。前処理とは、データを整えて「使える状態」にする作業のことです。これは、分析の正確さや信頼性に大きく影響する、非常に重要なステップです。実際、データサイエンスの工程の中で、前処理にかかる時間は全体の約８割を占めるとも言われています。そのため、データを手に入れたら、すぐに集計や分析、数理モデルの構築に進むのではなく、まず前処理を行ってデータをきれいに整えることが大切です。

２．前処理を行わなかった場合

なぜ前処理が必要なのか、それは前処理を行わないと解析結果がかわってしまうことが多いためです。例えば、気温と４つの商品 A,B,C,D の売り上げ金額の相関関係をグラフ（散布図）に描いた例で説明します。４つのグラフは回帰直線※の傾きはすべて同じになっています。

売上金額＝３．０（切片）＋１．８×気温（℃）

※回帰直線については第７講の「回帰分析」にて詳細は説明




商品 A　　商品 B　　商品 C　　商品 D

商品 A は、気温が上がると売上金額が増加しています。回帰直線からは多少ばらつきはあります。次に商品 B ですが、気温が上がり始めると売上金額は増加して行きますが、途中から下がっています。商品 C は回帰直線から外れてはいますが、１つの点が傾きを上げるように影響しています。商品 D は、温度が上がっても売上金額は変わらないのですが、温度が一番高いところだけが売上金額が大きい状態となっています。このような場合、例えば商品 C と商品 D の外れ値を除去などする前処理を行って回帰直線を求めるのが良いでしょう。

もし、前処理しなかった場合、４つの商品の回帰直線が同様な式になってしまい、傾向が全く違うのにも関わらず、売上予測が同じ結果となってしまうので注意が必要です。

## ３．前処理の全体像

前処理は以下のようなステップに分かれます。

・欠損値の処理
・異常値と外れ値の検出と対応
・データ型の変換
・重複データの削除
・変数のスケーリングや正規化
・カテゴリ変数のダミー化

これらは分析の目的や手法に応じて柔軟に選択されます。

欠損値とは、データの一部が記録されていない状態を指します。例えば、アンケートで「年齢」の欄が空白だった場合、その項目は欠損値です。空白であることからゼロをイメージして、欠測値＝０とすることは誤りとなります。欠損値があると統計分析や機械学習に悪影響を与えます。主な対処法は以下の通りです。

| 対応法 | どのような場合に対応できるのか、注意点など |
|---|---|
| 欠損行の削除 | データ量が十分な場合に対応できる。行自体を削除するので、データ総数が減ってしまう。<br>（例えば、変数の数よりサンプルの数が３倍以上ある） |
| データ全体で平均、中央値、最頻値による補完 | 欠測値を平均で埋めると、“だいたいの傾向”は見えるようになるが、“ちょっと変わった特徴”が見えにくくなることがある。 |
| 前後の値で補完 | 時系列データ※などで欠測値の前後のデータがあり、前後の平均値で補完する。 |
| モデルによる補完 | 高度な手法であり、アプリケーションが必要。 |

※時系列データ：「時間の流れにそって記録されたデータ」のこと。例えば、毎日の気温（8 月 1 日：35℃、8 月 2 日：34℃…）

## ４．異常値と外れ値

### 4.1　異常値と外れ値の違い

異常値とは、通常とは異なる挙動やパターンで発生するデータです。システム的・意味的におかしい値で、その原因は入力ミスや測定エラーなど様々です。例えば、岐阜市の夏の気温のデータの

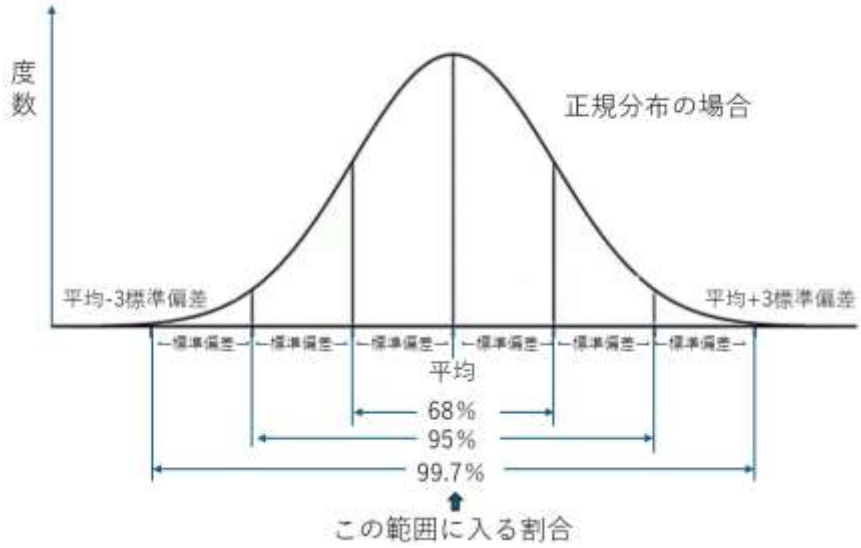
中に、-30℃や 0℃が含まれていた場合、実際にはありえない値であり、意味的におかしく、-30℃は符号誤り、0℃は誤って入力した、などが考えられます。

一方、外れ値は統計的に他と大きく異なる値です。例えば平均から大きく離れた値ある場合です。外れ値の検出には、標準偏差を用いた統計的検出、箱ひげ図（IQR）などがあります。

4.2 標準偏差を用いた統計的検出

標準偏差とは、データが平均からどれぐらい散らばっているかを示す指標です。標準偏差を用いることで、外れ値かどうか判定することができます。

よくある判定方法は、標準偏差の値から 3 倍以上離れているかどうかを基準とし、3 倍より外れたものを、外れ値と見なす方法です。ただし、標準偏差を用いて外れ値を判定する場合は、極端な外れ値に引っ張られる可能性に注意しなければなりません。




※正規分布：正規分布とは、平均を中心に左右対称に広がる「山の形」をした分布のことです。 たとえばテストの点数では、平均点の近くに多くの生徒が集まり、平均から離れるにつれて生徒の数は急激に減っていく傾向があります。

上記の図は中央が平均であり、平均を中心に「標準偏差の±1 倍、±2 倍、±3 倍」の範囲を考えます。 グラフの縦軸は、各値がどれくらいの頻度で現れるか（度数）を示しています。 範囲が広がるにつれて、そこに含まれるデータの割合も増え、

±1 標準偏差の範囲には約 68%

±2 標準偏差には約 95%

±3 標準偏差には約 99.7%

のデータが含まれます。 たとえば、データが 1000 個ある場合、約 997 個が「平均±3 標準偏差」の範囲内に収まり、残りの約 3 個はその外に出ます。 このように、平均から大きく離れた値（±3 標準偏差より外）を「外れ値」として扱います。

4.3 箱ひげ図を用いた外れ値の判定

8 人の国語テストの結果があった場合で説明します。

まずデータを小さいものから順番に並べて 4 等分します。

| 45　61 | 69　70 | 72　74 | 78　95 |
|---|---|---|---|

最小値：45

第 1 四分位数（Q1）：65（61 と 69 の平均）

中央値（Q2）：71（70 と 72 の平均）

第 3 四分位数（Q3）：76（74 と 78 の平均）

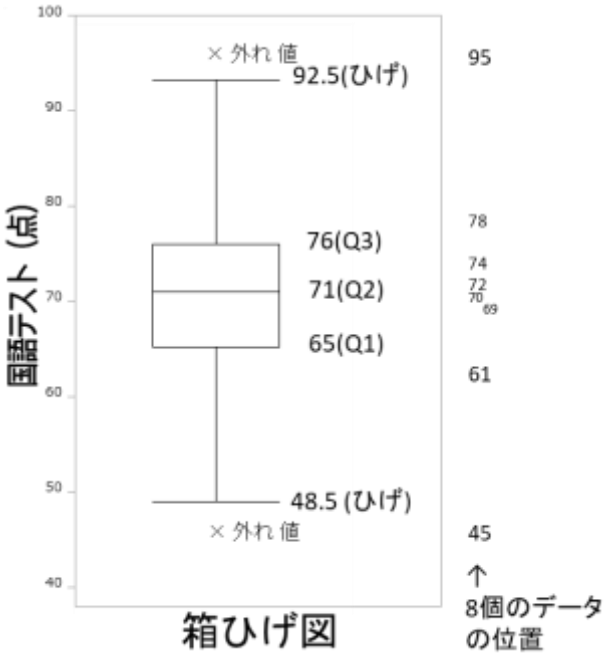
最大値：95

四分位範囲（IQR）＝第 3 四分位数（Q3） － 第 1 四分位数（Q1）＝76-65=11

ひげの上端：第 3 四分位数（Q3）+1.5×IQR より小さい最大値＝76+1.5×11=92.5

ひげの下端：第 1 四分位数（Q1）－1.5×IQR より大きい最小値＝65-1.5×11=48.5

ひげの範囲から外れた値が外れ値であり、95 と 45 が該当します。



## 4.4　外れ値の対応

対応は、削除・修正・別変数として扱うなど、目的に応じて選択します。

対処法は大きく 3 通りです。

・正しい値に修正する

データの入力ミスやシステムのエラーなど、外れ値が生じた要因が判明している場合は正しい値に修正します。

・行ごと除外する

要因が不明な場合や、要因が分かっていてもあまりに外れ具合が大きいデータは行ごと除外することを検討します。

・そのまま使用する

外れ値の要因が判明しており、かつ分析結果に大きな影響がなさそうな場合は、そのまま使用することもあります。

外れ値が残ったままデータ分析を実行すると、ほとんどのケースで全体の分析結果がゆがんでしまいます。極端に大きな値や小さな値を分析データに含めることで、分析結果が外れ値に引っ張られてしまうからです。

## 4.5　事例による説明

ある生徒の通学時間（分）に関するデータを 100 件取得し、スプレッドシートに右記の通り入力しました。通学時間は通常、約 60 分かかりますが、ばらつきがあります。

このデータの中で、2025/4/7 の-58 分は、マイナスとなっており、確認が必要です。また、2025/4/8 のゼロも同様に時間がゼロということはありえませんので、確認が必要です。一方、2025/4/11 は病欠となっており、通学していないので、欠損値として処理が必要です。

C10　病欠

| | A | B | C |
|---|---|---|---|
| 1 | Data No | 年月日 | 通学時間（分） |
| 2 | 1 | 2025/4/1 | 65 |
| 3 | 2 | 2025/4/2 | 58 |
| 4 | 3 | 2025/4/3 | 62 |
| 5 | 4 | 2025/4/4 | 101 |
| 6 | 5 | 2025/4/7 | -58 |
| 7 | 6 | 2025/4/8 | 0 |
| 8 | 7 | 2025/4/9 | 60 |
| 9 | 8 | 2025/4/10 | 61 |
| 10 | 9 | 2025/4/11 | 病欠 |
| 11 | 10 | 2025/4/14 | 62 |
| 12 | 11 | 2025/4/15 | 75 |
| 13 | 12 | 2025/4/16 | 67 |

：　省略

| | | | |
|---|---|---|---|
| 100 | 99 | 2025/8/22 | 59 |
| 101 | 100 | 2025/8/25 | 60 |

５．データ型の変換と整形

同じ列（変数）の 中で、数値・文字列・日付などの形式がバラバラだと、処理ができません。例えば、下表の「誕生日」について年号や表示形式が混在しています。「クラス」については、数字と文字が混在しています。

| 行　　列 | 名前 | 誕生日 | クラス番号 |
|---|---|---|---|
| 1 | A | 平成１５年５月１５日 | 3 |
| 2 | B | 2003 年 8 月 1 日 | 1 |
| 3 | C | 31/12/2003 | 01 |
| ： | | | |

| 誕生日 | クラス番号 |
|---|---|
| 2003/05/15 | 3 |
| 2003/08/01 | 1 |
| 2003/12/31 | 1 |
| 日付型 YYYY/MM/DD に統一する | 数値で統一する |

データを扱うとき、「正規化」という前処理が必要になることがあります。これは、データの値の範囲（スケール）を整えることで、異なる単位や桁のデータを比較しやすくするための方法です。代表的な正規化の方法には、次のようなものがあります。

Min-Max 正規化：データを 0〜1 の範囲に収めます。 例：身長や体重など、値の幅が大きく異なるときに使います。

Z スコア正規化：データを平均 0、標準偏差 1 に変換します。 例：テストの点数など、ばらつきを分析したいときに使います。

６．データ変換・カテゴリ変数

「性別」「地域」などのカテゴリ変数は、機械学習モデルで扱うためにデータ変換、数値化が必要です。代表的な方法は「ダミー変数化」で、例えば「性別」が「男性」「女性」の場合、male = 1、female = 0 のように変換します。

７．まとめ

前処理は単なる技術ではなく、「データの意味を理解する力」を養う学びでもあります。学校現場では、生徒の成績データやアンケート結果など、身近なデータを題材にすることで、実践的な理解が深まります。教員自身が前処理を体験することで、データ活用の可能性を広げる第一歩となるでしょう。