

School DX Strategy Coordinator Overview

学校 DX 戦略コーディネータ概論

【VI】

データサイエンスから見える新たな学びの未来像

岐阜女子大学
デジタルアーカイブ研究所

目 次

第 1 講	データサイエンスとは何か	2
第 2 講	データの種類と収集方法	9
第 3 講	データの前処理とクリーニング	23
第 4 講	データの可視化と探索的データ分析 (EDA)	30
第 5 講	統計学の基礎	37
第 6 講	機械学習の基本概念	53
第 7 講	回帰分析と分類モデル	59
第 8 講	クラスタリングと次元削減	68
第 9 講	データサイエンスにおけるプログラミング基礎	75
第 10 講	自治体が実施する大規模学力・学習状況調査とその分析	81
第 11 講	生成 AI と ELSI フレームワーク	88
第 12 講	データサイエンスの実践的応用例	98
第 13 講	データ可視化の高度な技術	108
第 14 講	AI と深層学習の基礎と応用	118
第 15 講	データサイエンスの未来と教育への展望	128

本テキストの活用にあたって

1. 学習を進めるにあたって

本テキストは、第1講から第15講まで、15の講義により構成され、学校DX戦略コーディネータ概論について学ぶようになっています。

本テキストを使って学習する際、次のことに留意して、学習活動を行ってください。

- ①本テキストとe-Learningは、事前にテキストと動画で学習する自律的なオンライン研修の教材です。
- ②講習の内容は、まず、テキストとe-Learningとの両方を活用して学びます。
- ③講義では、始めに各講で講義の目的と学習到達目標についての説明を行います。
- ④講義内容について、受講者による自己研修を行います。
- ⑤各講の終わりに課題を示します。自分の学習の深度に従って、考えてみましょう。

2. このテキストによる講義の特色

- 学習が進めやすいようにするテキストと、講義内容を解説する動画の視聴を併用することで、受講者の学びを確実にするとともに、受講者の便宜を図っています。
- 多忙な学習者にとって、いつでも、誰とでも、どこからでも受講者の都合で講義内容についての基礎的な学習が進められます。
- 講義の内容は、データサイエンスの基礎として必要な資質・能力について短時間で学習できるようにするものです。一度の講義による講習とちがって、テキストとe-Learning等の教材は、繰り返し視聴することができます。
- 講義の内容は、受講者にとって、その後の職場でのデジタルアーカイブの実践に有効に活用していただける内容です。また、テキストと動画を、職場の校内研修や研究会などで活用していただくことで、受講者が学んだことを多くの学習者に広めることが可能になります。

3. 本テキスト及びe-Learningの利用にあたって

- ・本テキスト及びe-Learningの著作権は、岐阜女子大学にあります。
- ・著作権や肖像権など取扱いには注意してください。

4. QRコードの利用にあたって

- ・QRコードは、タブレットPCやスマートフォンのQRコードリーダーをご利用ください。



学校DX戦略コーディネータ概論【VI】



利用の際は必ず下記サイトを確認下さい。

www.bunka.go.jp/jiyuriyo

第 1 講 データサイエンスとは何か

白水 始（国立教育政策研究所）

【学習到達目標】

1. 一般的なデータサイエンスについて理解したうえで、教育のためのデータサイエンスの在り方を説明できる。
2. データサイエンスを支える統計と機械学習という二本の柱について、具体例をもとに説明できる。
3. 本講座全体の流れを理解し、学習の見通しと動機づけを持つことができる。

本講義は、「人工知能概論【Ⅱ】データサイエンスから見える新たな学びの未来像」の第 1 講であり、「データサイエンスとは何か」をテーマとしています。一般的なデータサイエンス論ではなく、特に教育現場の質向上につながるデータサイエンスの在り方に焦点を当てます。

教育とは、一人ひとりの児童生徒の「学び」と「育ち」を支える人間的な営みです。この営みの中で、データ、分析、可視化がどのように役立つのか、そしてデータサイエンスを導入する際の留意点は何かを考察します。

1. 「教育データ」の定義と種類

2.1 そもそもデータとは何か

データとは、「人が直面する課題に対して**判断、議論、意思決定**するために用いることができる情報」全般を指します。

2.2 「教育データ」の定義

1. 本来の定義（教育場面で得られるデータ全般）

本来の「教育データ」は、教育場面で得られるデータすべてを意味します。これには、以下のものなどが含まれます。

記述回答：教師が児童のプリントを集めて学習到達度を把握するために参照する情報。

観察情報：教師が気になった児童を後で確認するために見る「表情」や「つぶやき」。

研究データ：授業改善のために研究者が記録した「ビデオデータ」など。

2. 文部科学省が定義する「教育データ」（デジタルデータ）

昨今「教育データ」と呼ばれるものは、文部科学省の定義にあるように、主に**デジタルデータ**を意味することが多くなっています。

背景：GIGA スクール構想による 1 人 1 台端末の整備。

内容：端末の利用ログ、デジタルドリルの回答時間など、紙媒体（アナログ）では得られなかった子供の学びに関するデジタルデータが利活用できるようになってきています。

2.3 「教育データ」の主な種類

教育データは複数の視点から分類されます。

1. 業務の性質による分類

- **行政系データ：**文部科学省や教育委員会など、行政組織に関わるデータ。
校務系データ：学校の運営（出欠席、成績処理など）に関わるデータ。
学習系データ：児童生徒の学びの過程や成果に関わるデータ。本講座で主に扱う対象です。

2. 形式による分類

アナログデータ：連続的に変化するデータ（例：波形、アナログ時計の動き）。

デジタルデータ：離散的に（飛び飛びで）変化するデータ（例：コンピュータで処理されるデータ）。

3. 構造による分類

- **質的データ（非構造化データ）：**文字、画像、動画情報など、そのままでは四則演算が適用しにくいデータ。

処理：後半の講義で、機械学習等により分析が可能になることが解説されます。

量的データ（構造化データ）：数値情報のデータであり、統計処理が可能になるもの。

処理：本講座で主に扱う対象です。

Ⅲ. 「教育データ」利活用のメリットと基盤

3.1 立場ごとの利活用メリット

教育データを活用することで、様々な関係者に以下のようなメリットがもたらされます。

立場	メリット
子供	自らの学びを振り返り、広げたり、伝えたりすることが可能になる。
教師	よりきめ細かい指導や支援が可能となり、自身の経験や知見と照合することで成長につながる。
保護者	子供の学校での様子を確認するなど、学校との連携が容易になる。
学校設置者	類似の自治体との比較や施策の改善が容易になる。

3.2 利活用の基盤：「学習過程の解明」

上記のメリットを実現するためのすべての基盤となるのが「学習過程の解明」です。

子供や教師が**どう学ぶのか**というプロセスが分からなければ、データを有効に活用した支援はできません。

教育データをまずは**学習過程の解明**に役立てることが最も重要です。

IV. データサイエンスと教育データサイエンスの構造

4.1 データサイエンスの構造

一般的なデータサイエンスは、以下の3つの視点（perspective）が重なる領域にあると定義されます。

Computational（計算機科学的視点）：データを処理し、アルゴリズムを実行する技術的な側面。

Statistical（統計的視点）：データを分析し、傾向や予測を行う数理的な側面。

Human（人間的視点）：課題解決や意思決定といった、データを活用する目的や文脈に関わる側面。

4.2 教育データサイエンスの定義と構造

この枠組みを教育分野に適用したものが「教育データサイエンス」です。Sanne Smith 氏のモデルに基づき、以下の3つの要素の重なりとして説明されます。

構成要素	領 域
コンピューターサイエンス (Computer Science)	技術的な知識とスキル。
統計 (Statistics)	データ分析とモデル構築の数理。
教育理論と実践・専門分野 (Education Theory and Practice / Domain Expertise)	教育特有の知識と文脈。

4.3 各領域の重なりと「教育データサイエンス」の位置づけ

上記の3つの円が重なる領域は、それぞれ以下の専門分野に対応します。

重なる領域	専門分野
コンピューターサイエンス× 教育理論	ソフトウェア開発 (Software Development)
コンピューターサイエンス× 統計	機械学習 (Machine Learning)
教育理論×統計	従来の研究 (Traditional Research)
3つすべてが重なる中心	教育データサイエンス (Education Data Science)

教育データサイエンスは、単なる教育学者やデータサイエンティストでは対応できない、専門的な知識の統合が必要な分野であり、「教育データサイエンティスト」の育成が求められます。

V. 教育における統計的視点の導入：多層型支援の例

5.1 多層型支援（RTI モデル）とは

「教育データサイエンス」を構成する要素の一つ、「統計的な視点」を教育に持ち込む具体的な例として**多層型支援**（Response to Intervention: RTI）が紹介されます。

前提：子供たちを「支援が必要／不要」と最初から二分しない。通常の学級に多様な子供がいることを前提とする。

- **第1層支援：**通常の学級における授業・学級づくり。

第2層、第3層支援：第1層支援のみでは不十分な場合に、徐々に支援を付け足す。

- 支援の有効性を**評価・改善**するプロセスを持つ。

「できない」原因を子供個人に帰属させず、**環境（指導や支援）との相互作用**と捉える。

教師チームで意思決定を行う。

5.2 「分布」という統計的な視点

ある学校長の「Cプラスの子供たちをB層に確実に引き上げられる授業をしよう」という発言は、「分布」という統計的な視点を教育現場に持ち込んだ例です。

正規分布の視点：子供たちの成績などの分布は、正規分布のような形（A層、B層、C層）になることが多い。

現場の課題：教員は、極端に進んでいる子（A層）や遅れている子（C層）に目が行きがちで、**ボリュームゾーンである中間層（B層）への意識**が薄くなることがある。

分布視点の効用：

「クラス全体（第1層）」への支援を強化し、**分布全体を引き上げる**ことを目指す。

結果的に、個別の支援が必要な子供たち（第2層・第3層）への対応も明確になる。

「分布」の視点を持つことは、教育実践に新たな気づきを与え、**クラス全体での質の向上**に貢献します。

VI. 教育における統計 × 機械学習の導入：教育付加価値評価システム

6.1 教育付加価値評価システム（EVAAS）の例

統計に「機械学習」の視点を持ち込む例として、アメリカで導入された**教育付加価値評価システム（Education Value-added Assessment System: EVAAS）が挙げられます。

目的：学校や教師が生徒の学力向上に与えた影響（付加価値）を測定するツール。

仕組み：

過去の実績データから、生徒が**将来どれくらいの成績をとるか**を予測（Predicted Performance）。

実際の成績（Current Performance）と予測値の**差分**（Value-added）を算出。

この差分（付加価値）を「教師の指導の効果」とみなす。

背景にあるモデル：生徒の学力の伸びから、家庭の経済状況や学校・地域の影響などの要因を統計的に取り除くことで、**教師個人の効果**を抽出するモデルが構築されています。

VII. 教育データサイエンスの目指すべき方向性と留意点

7.1 EVAAS の課題と教訓

EVAAS のようなシステムには、教育データサイエンス導入の難しさを示す課題があります。

現場の混乱：アメリカでは、この指標を教員の**給与や賞与に直結**させた結果、訴訟や現場の混乱を招きました。

懸念点：教育の成果を**学力テストの結果だけ**で判断することへの懸念や、教員間の**分断**を招く恐れがある。

Sanne Smith 氏が指摘するように、教育データサイエンスに対しては**慎重**であるべきであり、教育をより良くするためのツールとして機能させる必要があります。

7.2 避けるべき方向性（教育をおかしくする例）

データサイエンスを導入する際に、教育の本質を見失い、避けるべき方向性は以下の通りです。

インセンティブによる管理：データ（結果）を基にインセンティブを付けようとする（ボトムアップな管理）。

成果の限定：教育の成果を**標準的な学力**だけで見ると（資質・能力の一体性を見落とす）。

教師間の分断：教師を相対評価し、競争を促し、分断する。

7.3 目指すべき方向性（健全な機能）

教育データサイエンスを健全に機能させるための、目指すべき方向性、つまり「門番」の役割を果たすスタンスは以下の通りです。

理論に基づく実践：仮説（学習理論）を**もとに、まずは「よい授業」を創る**（トップダウンなアプローチ）。

一体的な育成：資質・能力を**一体的に育成**する視点を持つ。

協働の重視：教師の「協働」を何よりも大切にする。

目的のための活用：上記の目的のために、**テクノロジーとデータ**を活用する。

課題

- ① 教育データと教育データサイエンスの説明：本講座の例を結び付けて、教育データと教育データサイエンスとは何かを説明する。
- ② データサイエンス導入の留意点：データサイエンスを教育に導入する際の留意点を述べる。

第2講 データの種類と収集方法

尾関智恵（岐阜大学・准教授）

【学習到達目標】

- ① さまざまな種類のデータ（定量データ、定性データ、時系列データなど）を理解し、それぞれの特徴や適した分析方法について説明できる。
- ② データの収集方法（観察、アンケート、実験など）を理解し、具体的な場面に応じた適切な収集手法を選択できる。
- ③ データの種類と収集方法の違いを理解し、実際の教育現場や調査活動において適切なデータ収集計画を立てることができる。

1. はじめに

教育研究において、データは主観を客観化し、経験を共有可能な知識に変換するための重要な基盤です。統計学は、教育現場で得られたデータから客観的な根拠を提供し、一般化可能な知見を得るための強力なツールとなります。

授業研究や実践において、主観的な感覚や学習効果を示すのは難しいものです。しかし、教育の品質を上げ、改善を進めていくためには、やはり客観的な根拠が必要となります。データを収集し活用することで、印象や経験だけでなく、確か証拠に基づいた教育実践の改善が可能になります。

本講では、データの基本的な種類と特徴、そして教育現場における適切な収集方法について学びます。データを収集することは大変な作業ですが、それだけの価値があるということをまず押さえておいていただきたいと思います。

2. 教育研究におけるデータと統計の役割

2.1 教育研究におけるデータとは

教育研究におけるデータには、以下の3つの重要な役割があります。

(1) 主観を客観化する

教員の「感覚」や「印象」を数値や記録として表現できます。主観的な効果を客観的に見られるようにすることで、教育実践の効果を明確に示すことができます。

(2) 経験を共有可能な知識に変換する

個人の経験を他者と共有できる形にします。経験則を重ねていく中で、その経験を有益なものとして共有可能な知識に変換するためには、見直しや確認が必要です。データはそのための重要なツールとなります。

(3) 教育実践の改善サイクルを回す基盤

こうした蓄積がなされていくと、教育全体の改善、実践の改善をサイクルとして回せるようになります。データはそのための基盤を作る重要な要素です。

2.2 教育研究における統計の役割

統計学は教育研究において以下の3つの重要な役割を果たします。データが集まってくると統計的な分析が可能になり、様々な知見を得ることができます。

(1) 客観的な根拠を提供する

「この学習活動は効果がありそう」という印象を、数値で検証できます。個人の経験や直感だけでなく、データに基づいた判断ができるようになります。数値に振り回されることは避けたいところですが、数値的な評価が出てくると、客観的な根拠として使えるものになります。

(2) 一般化可能な知見を得る

目の前の30人のクラスでの結果が、他のクラスや学校でも当てはまるかを推測できます。特定の文脈での観察を超えて、より広い適用可能性を議論できます。統計処理をした数値化されたデータがあれば、他の授業や他の応用でも使える知見として育てることができます。

(3) 複雑な現象を整理・理解する

多くの要因が絡み合う教育現場で、何が重要な要因かを見極めることができます。データから意味のあるパターンを抽出したり、要因の相対的な影響力を比較したりできます。統計処理や解析ができるということは、推定をするといったようなことが可能になり、複雑な現象を整理・理解するための解析をかけやすくなります。

3. データの種類と特徴

3.1 情報とデータの関係

世の中は「物質」「エネルギー」「情報」の3つで構成されています。情報学では、この3つの組み合わせで世界を理解します。物があり、それはエネルギーがないと動きません。エネルギーを入れることによって動くことができますが、しっかりとした情報がないと非常に危険な動き、もしくは予想外の動きになってしまい制御できません。したがって、きちんと情報を入れて動かしていくということが一般的な考え方となります。この制御や正しく動かすという際に、情報データというものが非常に重要になってきます。

情報とは、文字・数字などの記号やシンボルの媒体によって伝達され、受け手に状況に対する知識や適切な判断を生じさせるものです。これについてデータ化してまとめていくというのが、今回の話題の中心になります。

3.2 記号化とデータ化の意味

たとえば、花を見たとき、人はそれを「パターン」として認識し、「花」という記号で表現し、さらに「種子植物の生殖器官」という知識として理解します。人間の知的活動には、こういった情報を記号化して、他者や未来の子供たちに伝えるといったまとめ方をしているという前提があります。情報を長期間保存し、他者と共有するためには、このような記号化（デジタル化）が必要となります。従来は口伝といったような形で人から人へ情報を伝えてきましたが、IT 革命、情報化の波が押し寄せた結果、人が口伝するよりももっと確実にその事実を長期に残せるような技術が手に入っています。授業や教育活動の中でも、こういった ICT の力を借りるのは非常に当たり前になってきています。

3.3 デジタルとアナログ

情報データというのは、いくつかの特徴によって保存のされ方が変わってきます。

(1) デジタル

- ・ 離散的に整数値で表現すること。デジタルコンピュータでは 2 進数（0 と 1）で表す
- ・ 正確な記録と再現が可能
- ・ 長期的に残しやすい
- ・ 離散的なため、中間の数値を抜いてしまうという特徴がある
- ・ 連続した滑らかな波ではなく、離散状態になり、情報が落ちる可能性がある

(2) アナログ

- ・ 連続した量を他の連続した量で表示すること
- ・ 時計や温度計などがその例で、あいまいな表現が可能
- ・ 中間のデータ、情報は全部残っている
- ・ 長期保存については、残すべきか残さないべきかという判断が必要
- ・ 冗長な表現や、場合によっては不要かもしれない情報が含まれる

このトレードオフをどう扱うかというところが情報学やコンピュータサイエンスでよく話されているところです。データ分析では、連続値（アナログ）を離散値（デジタル）に変換することで、コンピュータでの処理が可能になります。

3.4 量的データと質的データ

教育研究で扱うデータは、大きく「量的データ」と「質的データ」に分けられます。データの形式の話もありますが、その内容や特徴についての分け方もあります。

(1) 量的データ（数値化できるデータ）

量的データには以下のようなものがあります。

- ・テストスコア、出席率、課題提出数
- ・アンケートの評定値（5段階評価など）
- ・学習時間、発言回数

特徴：

- ・統計分析が可能
- ・客観的な評価ができる
- ・大規模調査に向いている
- ・数えることができる
- ・客観性が高く、調査しやすい

注意点：テストの点が取れていれば全てを理解できているかというのはイコールではありません。量的データだけでは捉えきれない側面があることを認識しておく必要があります。

(2) 質的データ（数値化が難しいデータ）

質的データには以下のようなものがあります。

- ・学生の自由記述回答
- ・インタビューの発話内容
- ・授業観察のフィールドノート

特徴：

- ・文脈や意味の深い理解が可能
- ・少数のデータでも深い洞察が得られる
- ・「なぜ」という問いに答えられる
- ・学習者の内的な状態や、教員がそのときに何を考えて話をしたかといった文脈、判断した深い意味の理解を表現できる

重要性：質的データを収集することはできても、それをどのように解釈するのか、分析するのは非常に難しいものです。しかし、学習研究では、こういったデータを解釈として添えておくことが非常に重要になります。

(3) 両方のアプローチの重要性

現状としては、統計的な処理をした数値化されたデータと、質的なデータの両方を収集できるようになってきています。この両面から解釈するというのが、研究の中でごく自然になされています。

今後、収集するデータの種類が増えてくると、複数のアプローチから学習者の状況を推し量ることがますます増えてきます。そのため、データを収集する、蓄積する際に、こういったものを蓄積しておくべきなのかということは、ますます重要な議論になってきます。

3.5 データ駆動型処理と概念駆動型処理

人間の情報処理には2つのアプローチがあります。

- ・データ駆動型処理（客観科学的）：入力された情報の特徴を抽出し、パターンを認識する論理的な処理
 - ・概念駆動型処理（主観科学的）：既有知識・経験・期待を基にした確認処理
- 教育研究では、これら2つのアプローチが相互に影響し合いながら、解釈、理解、学習が進んでいきます。

4. 尺度の種類

データを分析する際、データがどのような尺度（測定の水準）で測られているかを理解することが重要です。データを収集するときには、尺度と呼ばれるものがあり、代表的な4つが挙げられます。

4.1 名義尺度（カテゴリー）

- ・単なる分類やラベルを表します
- ・どんなラベリングをするのかということが重要
- ・例：性別、専攻、出身地、学習スタイル
- ・使用できる統計：度数、割合、カイ二乗検定

4.2 順序尺度（順位）

- ・順序に意味がありますが、間隔は等しくありません
- ・順番があるものについては、きちんと順番も記録する必要があります

- ・例：成績（A・B・C）、理解度（理解できた～理解できない）、満足度（とても満足～全く満足しない）
- ・使用できる統計：中央値、順位相関

4.3 間隔尺度（等間隔）

- ・間隔が等しいですが、絶対的な 0 点がありません
- ・偏差値やばらつきというものを解釈するための感覚尺度
- ・例：偏差値、IQ スコア、温度（摂氏）
- ・使用できる統計：平均値、標準偏差、相関係数、t 検定

4.4 比率尺度（絶対 0 点あり）

- ・間隔が等しく、絶対的な 0 点があります（0 = 何もない状態）
- ・全体を推し量るために、比較する率で考える
- ・例：テスト得点、学習時間、年齢
- ・使用できる統計：すべての統計手法が使用可能、比率の計算も可能

これらの尺度をどう扱うかということについても、議論をしていく必要があります。

5. データの収集方法

教育研究においては、学習過程、どういった学び方をしているのか、何がつまづいているのか、もしくは早い段階で何か新しい推定に成功しているのかといった学習の内的な状況を見るために収集することが主となります。

学習過程や教材の提供タイミング、説明、学習活動が行われている際に発生する情報は、より質の高い学習を構築・検討するために非常に重要な情報となります。教育研究では、目的に応じて適切なデータ収集方法を選択することが重要です。

5.1 学習成果物・質問紙

収集対象：

- ・学習者自身が書き込んでいるもの
- ・読んでいるときにメモをした本
- ・レポート、課題、テスト
- ・プレゼンテーションの内容
- ・アンケート調査
- ・ポートフォリオ

特徴：

- ・学習過程を捉えるために有効
- ・学習成果を直接確認できる
- ・課題として提出する方法で収集可能
- ・最近ではタブレット教材も多く、タブレット上で作成されたログも含めて収集可能

注意点：

- ・ループリックを使った評価など、解釈方法が難しい場合がある
- ・質的データの場合、数値として推し量ることが難しい
- ・授業デザイン通りにしっかり活用ができているかといった評価方法が必要

5.2 観察**収集対象：**

- ・授業中の学習者の様子
- ・グループワークでの相互作用
- ・教室環境の影響

特徴：

- ・実際の行動を直接捉えられる
- ・学習者の実際の行動を直接見ることができる
- ・無意識の行動や相互作用を把握できる
- ・自己報告では得られない情報が得られる
- ・学習者自身が気づいていないけれども無意識に行っている学習活動や他者との議論の相互作用の状況を見ることができる
- ・自分では自認していないので報告はしないけれども、学習活動において重要な活動、動き、判断が見て取れる

注意点：

- ・観察者の主観が入りやすい
- ・時間と労力がかかる
- ・観察されることによる行動の変化（ホーソン効果）
- ・観察の方法や学習研究に経験がある人でないと、なかなか拾うことができない
- ・あまりじっと見ているとやりにくいので、カメラで撮る、気づかないうちに学習者の学習を阻害しないように収集するといった配慮が必要

5.3 インタビュー

収集対象：

- ・ 学習者の考え方や感情の理解
- ・ 教育実践の背景にある意図の把握
- ・ 困難さや課題の詳細な把握

特徴：

- ・ 深い理解が得られる
- ・ 予想外の発見がある
- ・ 「なぜ」を探れる
- ・ 観察よりも直接学習者本人にいろいろ問うことができる
- ・ 半構造化インタビューなど、質問をすることはある程度決めているけれども、話の流れとして盛り上がった方についてもう少し深掘りをするといった臨機応変な対応が可能
- ・ 学習者が何を考えていたかという深い理解が得られる
- ・ 教員側が持っていたこと以外のことが発生していたという予想外の発見がある

注意点：

- ・ 時間がかかる
- ・ 分析が複雑
- ・ インタビュアーのスキルに依存
- ・ インタビューについての経験ややり方を知っておく必要がある
- ・ 満足にインタビューについてデータが取れたかは、人によって、場合によってムラができてしまう可能性がある
- ・ データとして収集はできるけれども、扱いは慎重にしないといけない

5.4 その他の既存データ

収集対象：

- ・ 成績データベース、出席記録
- ・ LMS（学習管理システム）のログデータ
- ・ 図書館の利用記録

特徴：

- ・ 授業以外でも収集している情報が使える場合がある
- ・ 生徒の属性を知るためには必要

注意点：

- ・ 研究目的で収集していないものは注意が必要（研究倫理）

- ・データの目的外使用に関する倫理的配慮
- ・個人情報保護への配慮
- ・データの質と完全性の確認
- ・目的として研究目的で収集していない場合もあるので、研究倫理の配慮が必要
- ・取り扱いは簡単にできず、注意をして確認の上で使う必要がある
- ・場合によっては同意を得る必要がある

6. データ収集の実践例

いろいろな手法でデータを収集して、研究のため、よりよい授業のために分析・改善をしていきます。ここでは、実際の研究実践の例として、収集方法の具体例を示します。

6.1 協調学習場面における観察ツール（ROG）

協調学習場面における観察者の解釈を電子的に記録・共有する支援ツール

（ROG）の例があります。このシステムでは、観察者用タブレット PC から赤外線通信でキャプチャデータを送信し、リポジトリに保存します。

これにより、学習過程を詳細に分析・解釈した経験から学生の状態を把握し、最適な学習を作る形成的評価の実現が可能になります。蓄積しておけば、後からこういった解釈・分析をすることもできます。

●大学授業における包括的データ収集の例

大学生を対象とした約 80 人から 100 名の授業に対して、学習活動、グループ活動を記録した例があります。当時（20 年ほど前）は、パソコンを授業中に使うというのは結構珍しい状況でしたが、以下のような取り組みを行いました。

背景：

- ・当時は紙や黒板での授業が当たり前だった時代
- ・ICT の力を借りて、学習者が必要なときに知識にアクセスし、他の人が考えているものについてもオンライン越しでアクセスして議論を深めていくという斬新な新しい教育方法を目的として実施

収集したデータ：

- ・教案（word データ）：教員が事前に用意している講義案、時間配分・目標・話す内容などで構成
- ・配布物（word データ）：学生に配布する資料、作業用紙
- ・観察メモ（紙面）：講義中に教員・TA が作成する記録
- ・学生ノート（PDF）：講義中に学生が記入したノート

- ・ ReCoNote 利用履歴（テキストデータ）：講義中に利用した学習システム履歴
- ・ 音声データ（mp3 など）：講義中の教員・学生の発話データ（全学生に IC レコーダーを装着し、誰がどの時点で何の話をしていたかを全て記録）
- ・ 映像データ（QT など）：講義風景の映像データ

システムの特徴：

- ・ パソコンを間にして学生たちが資料を読んで説明し合う活動
- ・ パソコン上でどういう動きがあったかを自動的に収集するシステム
- ・ その場で教員が観察した結果をメモする機能
- ・ これらすべてをポータルで見られるようにした

従来はパソコンでの蓄積がなかなかできなかったものもデジタルデータとして収集できるようになりました。

6.2 学習環境と協調学習の効果の検討

インタラクティブな学習環境（MILaIS）と通常教室での協調学習の効果を比較した研究例では、以下のような知見が得られています。

- ・ 協調活動に適した学習環境の要件の解明
- ・ アイデアや知識の外化→共有→再吟味の活動設計プロセスの詳細分析
- ・ 学習成果物の質的評価（文章としての成立、業務日誌としての体裁、書き方のポイント、自分なりの工夫）

普通の教室で議論をする場合よりも、アクティブラーニング型の机や話す体制でやった方が議論が活発化するということを可視化しました。こういった活動で最終的な成果物にも変化が出てきたりといったことで、活動プロセスというものをどうデータ化するかという例として参考になります。

これらのデータ収集と分析により、教育実践の改善に向けた具体的な示唆が得られます。

7. データの信頼性と妥当性

データを集めて、それに対して何らかの統計処理をする、もしくは解釈を与えるということは当然できます。データも昔よりは収集しやすくなり、気軽にデジタルツールを使って集めることができます。

しかし、その集めたデータ自体が正しいものなのか、その学習者を本当に表すものか、学習者の内的状況を推し量るために妥当なデータであるかどうかということとは重要になってきます。

7.1 信頼性 (Reliability)

データの一貫性や再現性を示す指標です。

- ・再テスト信頼性：同じ測定を繰り返したときに同じ結果が得られるか
- ・一貫性：測定項目間で一貫した結果が得られるか
- ・評価者間信頼性：複数の評価者が同じ結果を出すか

7.2 妥当性 (Validity)

データが測定したいものを正しく測定しているかを示す指標です。

- ・内容的妥当性：測定内容が目的に適しているか
- ・基準関連妥当性：他の基準と適切に関連しているか
- ・構成概念妥当性：理論的な概念を適切に測定しているか

重要な問い：学習者の状況や過程を捉えるために適切なデータであるか？

7.3 データの質の確保

数値化されたものは特に、計算や統計的な処理をすれば、ある程度の評価値を出すことができます。しかし、その数値化されたものがどの程度信用できるものか、その学習者を表すのに妥当なものであるかというロジックがしっかりしていませんと、結果を出したとしてもあまり意味がない分析となってしまいます。学習者の状況や過程を捉えるために適切かどうかということは、収集段階から考えられる内容です。そういった意味でも、授業の設計・デザインをしっかりする必要がありますということに立ち返っていただけるといいと思います。

7.4 倫理的配慮

データの種類のところでも述べましたが、もともとその授業をより良くする、学習研究を目的にしていなくて集めているデータを使うことがある場合があります。その場合は、やはり個人情報等を含む内容ですので、適切な処理が必要です。

具体的には：

- ・同意を得る
- ・この情報は収集したら授業改善に使う可能性があるという断りを入れる
- ・データを使うときには倫理的な配慮が必要

今後ますます重要になってきますので、データを収集するときには、こういった同意を得る、使用者に対して何に使うかということをしっかり説明するということをご心掛けていただきたいと思います。

8. 良いデータ収集のポイント

教育研究において良質なデータを収集するためには、以下のポイントを押さえることが重要です。せっかく集めたデータを分析なり解釈なりしていくわけですが、やはり集めたデータで何か算出するというのは当然ですが、教育研究においては以下の点が重要です。

8.1 研究目的に合った方法を選ぶ

- ・ 研究の問いは何か？
- ・ どのような情報が必要か？
- ・ 量的データと質的データのどちらが適切か？
- ・ その目的に合ったデータが適切な方法で収集されているか

8.2 複数の方法を組み合わせる

- ・ 異なる視点からデータを収集
- ・ 結果の信頼性を高める
- ・ より豊かな理解を得る

8.3 限界を認識し、解釈に反映させる

- ・ どのようなデータ収集方法にも限界がある
- ・ バイアスの可能性を認識する
- ・ 結果の解釈に慎重さを持つ
- ・ 学習過程を推し量るには、データを集めたところで、その人の学習者の内部で起こっていることをすべて書き出すというのはほとんど難しい
- ・ 収集したデータでは、この内容はわかるけれども、ここが限界だということで、限界点をしっかり押さえた上で、誠実に解釈をして、その学習成果を報告していくという考え方が重要

8.4 実現可能性を考慮する

- ・ 時間：データ収集と分析にかかる時間
- ・ コスト：必要な費用
- ・ 人員：必要な人的リソース
- ・ 倫理：研究倫理審査の必要性

9. まとめ

本講では、教育研究に主に焦点を当て、データがどのような使われ方をすべきかという話から始め、その種類、特徴、集めたときにどういった整理方法があるか、信頼性と妥当性、またデータを使うときには倫理的な配慮が必要になってくることについて触れました。

重要ポイント：

- ・ データの種類 ： デジタル・アナログ、量的データ・質的データの違い
- ・ 尺度の理解 ： 名義尺度、順序尺度、間隔尺度、比率尺度の特徴と統計手法
- ・ データ収集方法：学習成果物・質問紙、観察、インタビュー、既存データなど、目的と用途に応じた選択が重要
- ・ 信頼性と妥当性：データの質を保证するための重要な概念です
- ・ 倫理的配慮 ： 研究倫理と個人情報保護への配慮が必要です

教育研究においてデータを適切に収集・活用することで、主観的な経験を客観的な知識に変換し、教育実践の改善につなげることができます。データを収集することは大変な作業ですが、やる価値があるということを押さえておいていただきたいと思います。

次講以降では、収集したデータをどのように前処理し、分析していくかについて学んでいきます。

参考文献

- ・ 総務省統計局, レベル別テキスト,
<https://www.stat.go.jp/dss/getting/textbook.html>
- ・ 総務省統計局, なるほど統計学園,
<https://www.stat.go.jp/naruhodo/index.html>
- ・ ソコスト, <https://soco-st.com/>
- ・ 中溝幸夫, 箱田裕司, & 近藤倫明. (1983). リンゼイ/ノーマン情報処理心理学入門 I—感覚と知覚. P.8-16
- ・ 平成 11 年度 CREST「高度メディア社会の生活情報技術」研究領域 三宅なほみ「高度メディア社会のための協調的学習支援システム」

課題

- ① 次のデータの種類を分類し、それぞれの特徴と適した分析例を述べなさい。
 - a) 生徒の身長測定値
 - b) 生徒の好きな教科（国語、数学、英語など）
 - c) 1 週間の気温の変化（時系列データ）
- ② 以下の状況に適したデータ収集方法を選び、その理由を説明しなさい。
 - a) 学校の授業改善のために生徒の意見を集めたい。
 - b) 校内の運動会の参加者数を正確に把握したい。
 - c) 地域の気候変動を長期的に観察したい。
- ③ あなたが教員として、クラスの学習状況を把握するためのデータ収集計画を立てるとします。どのようなデータを収集し、どの方法で行うかを具体的に記述しなさい。
- ④ 教育研究においてデータを収集し客観化することにはどのような役割や意義があるか。
- ⑤ 質的データと量的データの特徴と使い分けを説明せよ。
- ⑥ データの信頼性と妥当性の違いを簡潔に述べよ。

第3講 データの前処理とクリーニング

笹山和明（株式会社 村田製作所・情報科学アーキテクト）

【学習到達目標】

- ① データ前処理とクリーニングの基本的な目的と重要性を理解できる。
- ② 欠損値や異常値の検出と適切な処理方法を説明できる。
- ③ データの整形や正規化の手法を理解し、実際に適用できる。

1. はじめに

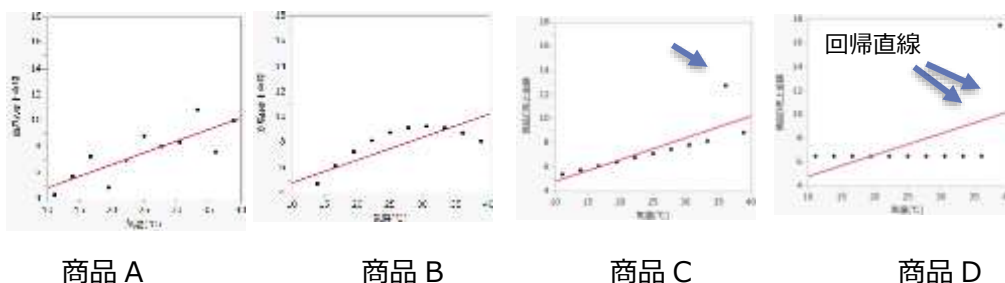
データサイエンスの最初のステップは「データの準備」です。現実のデータには、抜けている値（欠損）、入力ミス（誤記）、同じデータの重複など、さまざまな問題が含まれています。こうした問題があるままでは、正しい分析や予測ができません。そこで必要になるのが「前処理」です。前処理とは、データを整えて「使える状態」にする作業のことです。これは、分析の正確さや信頼性に大きく影響する、非常に重要なステップです。実際、データサイエンスの工程の中で、前処理にかかる時間は全体の約8割を占めるとも言われています。そのため、データを手に入れたら、すぐに集計や分析、数理モデルの構築に進むのではなく、まず前処理を行ってデータ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

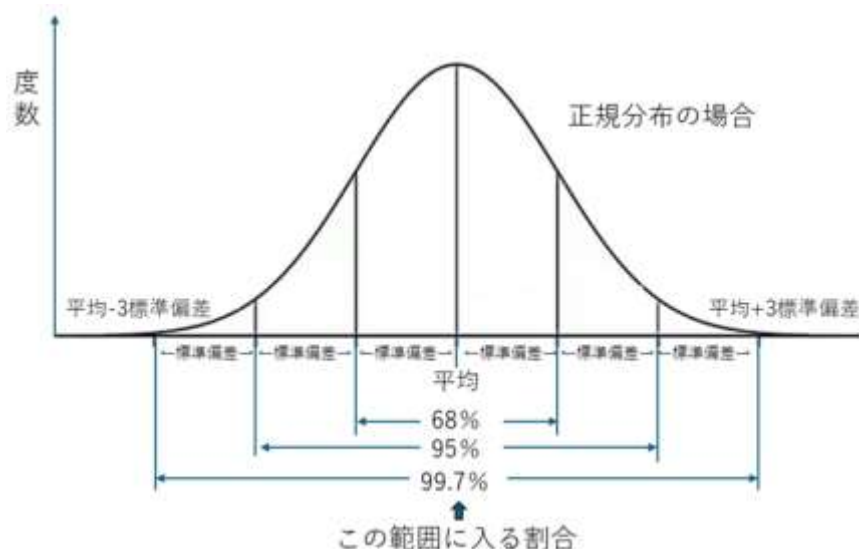
異常値とは、通常とは異なる挙動やパターンで発生するデータです。システマ的・意味的におかしい値で、その原因は入力ミスや測定エラーなど様々です。例えば、岐阜市の夏の気温のデータの中に、 -30°C や 0°C が含まれていた場合、実際にはありえない値であり、意味的におかしく、 -30°C は符号誤り、 0°C は誤って入力した、などが考えられます。

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く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

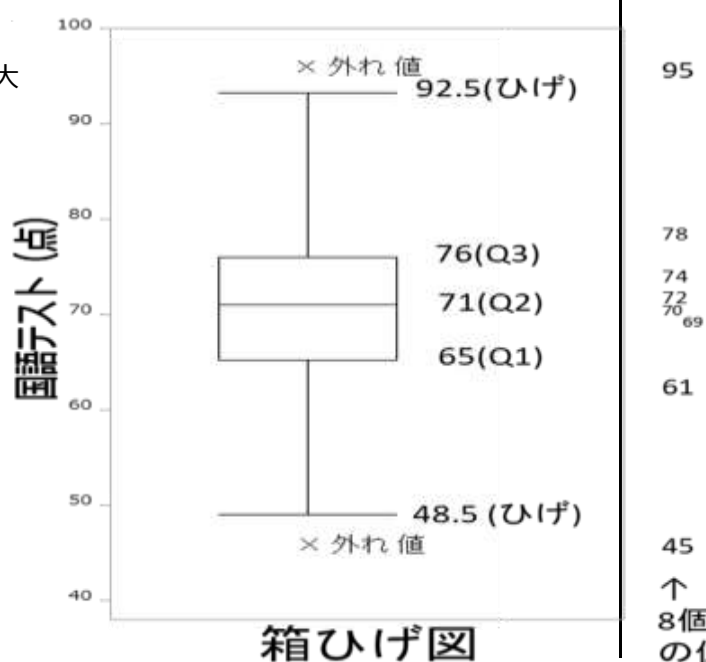
ひげの上端 : 第 3 四分位数

(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、要因が分かっているにもかかわらず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が、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	▼	:	✕	✓	<i>f</i> x	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 1 5 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 数値で統一する
YYYY/MM/DD

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。

課題

- ① 欠損値が含まれるデータセットに対して、どのような処理方法が考えられるか説明してください。
- ② 異常値を検出するための方法を2つ挙げ、それぞれの特徴を説明してください。
- ③ データの正規化と標準化の違いについて説明し、それぞれのメリットを述べてください。
- ④ データ前処理が分析結果の正確性や信頼性に与える影響と重要性は何ですか。
- ⑤ 異なる形式のデータを分析可能にするための変換や正規化の役割は何ですか。

第4講 データの可視化と探索的データ分析 (EDA)

荒木貴之（日本経済大学/ 社会構想大学院大学・教授）

【学習到達目標】

- ① 「探索的データ分析（EDA）」の概念を理解する：仮説検証型の分析に入る前に、データの構造や特徴を直感的に把握するプロセスの重要性を理解する。
- ② 基本統計量の限界と可視化の効用を知る：平均値や分散などの数値要約だけでは見落としてしまうデータの「真の姿」を、グラフ化によって発見できることを学ぶ。
- ③ 多角的な視点でデータを解釈する能力を養う：ヒストグラム、箱ひげ図、散布図などの適切な使い分けを習得し、シンプソンのパラドックスなどの統計的な落とし穴を回避する視座を持つ。

本講義では、AI時代に人間が担うべき「データの読み解き」に焦点を当てます。統計学者のジョン・テューキー（John Tukey）が提唱した「探索的データ分析（EDA）」を軸に、数値の羅列であるデータをグラフ化し、そこに潜むパターンや特異点を発見するための実践的技法を学びます。

1. なぜ私たちはデータを可視化するのか

1-1 データと対話する技術

「データ分析」と聞くと、高度な数式やAIによる自動予測を思い浮かべるかもしれません。しかし、最も重要なステップは、計算の前に人間が自分の目でデータをじっくりと観察することです。これを体系化したのが、米国の統計学者ジョン・テューキー（John Tukey）が1970年代に提唱した「探索的データ分析（Exploratory Data Analysis: EDA）」です。

従来の統計学が「立てた仮説が正しいか」を確認する「確証的データ分析（CDA）」を重視していたのに対し、EDAは「そもそもデータの中に何があるのか」を探る、いわば刑事の捜査のようなプロセスです。

デジタルアーカイブに眠る膨大なメタデータやログデータも、まずは EDA によってその性質を理解しなければ、有効な活用はできません。

1-2 アンスコム の例 (Anscombe's quartet) が教える「平均値の嘘」

私たちが普段業務で使う「平均値」は、時に重大な事実を隠蔽します。これを劇的に示したのが「アンスコム の例 (Anscombe's quartet)」です。

いま、4 つの異なるデータセット I から IV があり、それぞれの「平均値」「分散」「相関係数」は完全に一致しています。数値だけで判断すれば、これらは「同じ性質のデータ」です。しかし、これをグラフ (散布図) にすると、その姿は全く異なります。

表 1 基本統計量 <アンスコム の例 (Anscombe's quartet) >

アンスコム の例 (Anscombe's Quartet)				
特性	データセット I	データセット II	データセット III	データセット IV
平均値 (x)	9.0	9.0	9.0	9.0
分散 (x)	11.0	11.0	11.0	11.0
平均値 (y)	7.50	7.50	7.50	7.50
分散 (y)	4.12	4.12	4.12	4.12
相関係数	0.816	0.816	0.816	0.816
回帰直線	$y = 3.0 + 0.5x$	$y = 3.0 + 0.5x$	$y = 3.0 + 0.5x$	$y = 3.0 + 0.5x$

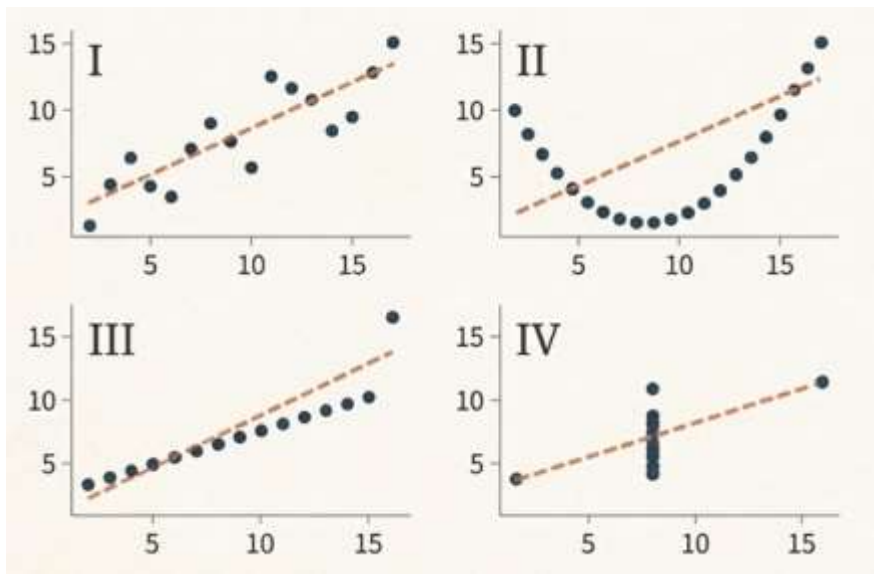


図 4-1 散布図 <アンスコム の例 (Anscombe's quartet) >

- I 綺麗な右肩上がりの直線
- II 放物線を描く曲線
- III 直線上に乗っているが、一つだけ極端な外れ値があるもの
- IV Xの値が全て同じで、一つだけ離れた点があるもの

このように、数値要約はデータを「丸める」行為であり、可視化はデータの「個性を復元する」行為です。特に文化情報や地域データのような複雑な対象を扱う場合、可視化を省略した分析は極めて危険です。

2. 分布の形状を読む（1変量データの可視化）

2-1 ヒストグラムと「ビンの魔術」

データの全体像をつかむ基本は「ヒストグラム（度数分布図）」です。横軸に階級（区間）、縦軸に度数（件数）をとります。ここで重要なのは、分布の「山」がいくつあるかです。

例えば、ある観光地の来訪者年齢層の平均が「40歳」だったとします。しかしヒストグラムを描くと、20代の山と60代の山の「二峰性（bimodal）」になっているかもしれません。この場合、「平均40歳向けの施策」は、若者にもシニアにも響かない残念な手となります。

また、ヒストグラムは階級の幅（ビンサイズ）を変えるだけで印象が激変します。意図的に印象操作を行わないためにも、複数のビン設定を試す姿勢が必要です。

データとの対話をはじめよう：EDAの基本ツール

ヒストグラム (Histogram)



データの全体像を把握し、「分布の形」を見る。例えば、観光客の満足度アンケートで平均3点でも、実は「5点(大満足)」と「1点(大不満)」の二つの山（二峰性）に分かれているかもしれません。平均点にはない物語を発見します。

箱ひげ図 (Box Plot)



複数のグループの分布を比較し、「外れ値」を発見する。統計学ではノイズとして扱われがちな外れ値は、文化研究においてはイノベーションの源泉です。

図4-2 EDAの基本ツール<ヒストグラムと箱ひげ図>

2-2 箱ひげ図で「外れ値」を愛する

複数のグループ（例：月ごとの来館者数）を比較する際に強力な武器となるのが「箱ひげ図（Box Plot）」です。

- 箱（Box）：データの中央 50%（第 1 四分位数から第 3 四分位数まで）が含まれ、データの「実力」を示します。
- 中央値（Median）：箱の中の線。平均値より外れ値の影響を受けにくい指標です。
- ひげ（Whiskers）：通常のデータの範囲を示します。
- 外れ値（Outliers）：ひげの外にある点。

統計処理において外れ値は「ノイズ」として削除されることもありますが、EDA ではこの外れ値こそが「宝の山」です。「なぜこの日だけアクセスが急増したのか？」「なぜこの地域だけ数値が突出しているのか？」。外れ値の背後には、必ず固有の文脈（コンテキスト）やストーリーが存在します。

3. 関係性の罫を見抜く（2 変量・多変量の可視化）

3-1 散布図と相関関係

2 つの変数の関係を見るには「散布図」が最適です。右上がりなら「正の相関」、右下がりなら「負の相関」です。しかし、ここで陥りやすいのが「相関関係を因果関係と混同する」ことです。

有名な例として「アイスクリームの売上と水難事故の件数には正の相関がある」という話があります。アイスクリームが事故の原因ではありません。ここには「気温」という第 3 の変数（交絡因子）が隠れています。可視化によって相関を見つけた後は、必ず人間のドメイン知識（現場の知見）でその背景を解釈しなければなりません。



図 4-3 因果関係を見抜く（擬似相関）

3-2 シンプソンのパラドックス

データを分析する際、全体で見た傾向と、グループ別に分けた傾向が逆転する現象を「シンプソンのパラドックス」と呼びます。

例えば、デジタルアーカイブの利用調査で、「グループA：観光客（一般）」と「グループB：研究者（専門家）」の2つのグループにおいて、全体では「（デジタルアーカイブの）閲覧数が多いほど幸せではない」という結果が出たとします。しかし、これを「観光客」と「研究者」とに層別化（ドリルダウン）して散布図を描くと、それぞれの層の中では「閲覧数が高いほど満足度が高い」という逆の傾向が見えることがあります。

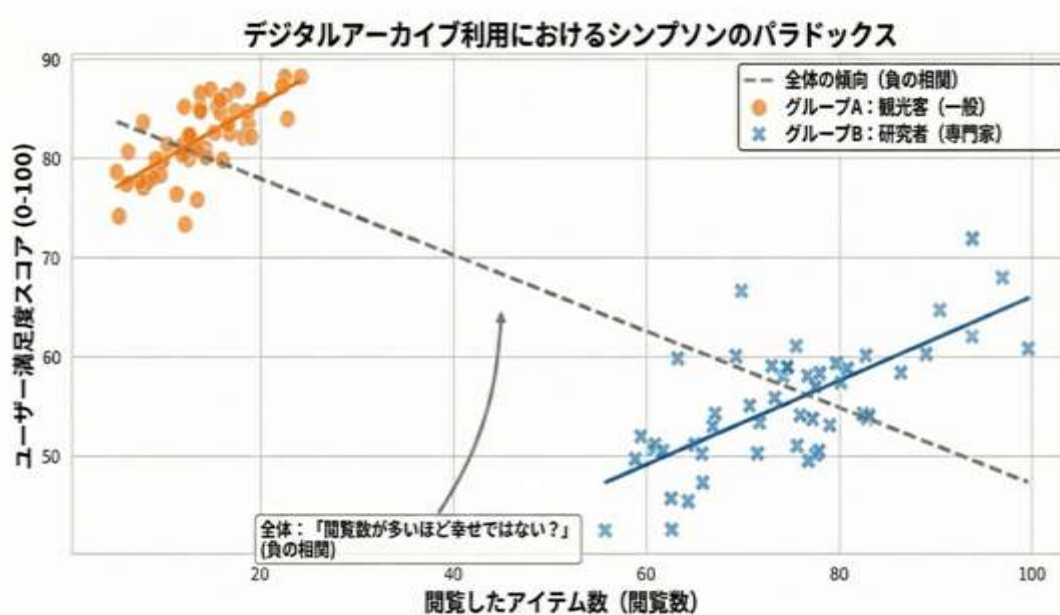


図4-4 シンプソンのパラドックス

これは、観光客（一般）の多くがユーザー満足度の出やすい利用者であり、研究者（専門家）は観光客に比べてユーザー満足度が出にくい傾向があるというような場合に起こる現象です。データを「混ぜるな、分けろ」は、EDAの鉄則です。

4. まとめ

本講義では、AI や高度な分析手法を使う前の基礎体力として、探索的データ分析（EDA）の重要性を学びました。

1. 数値は嘘をつく：平均値などの要約統計量を過信せず、必ずグラフを描いて分布を確認すること。

2. 分布の形を見る：ヒストグラムで山の数を、箱ひげ図で外れ値を確認し、データの多様性を把握すること。

3. 層別化して見る：全体の相関だけで判断せず、属性ごとにデータを分割（層別化）して、隠れた構造（シンプソンのパラドックスなど）をあぶり出すこと。

データ可視化は、単なる発表資料作成のスキルではありません。それは、データという無機質な信号の中から、人間臭い事実や、地域固有の課題を発見するための「レンズ」なのです。本講の発展としての第 13 講「データの可視化の高度な技術」では、より複雑な「つながり」や「空間」を可視化する技術へと進みます。

課題

① 外れ値のケーススタディ

ご自身の職場や身近なデータ（なければ公開されているオープンデータ）において、「外れ値」と思われるデータを探してください。そして、その外れ値が「単なるエラー（ノイズ）」なのか、それとも「重要な意味を持つ特異点（インサイト）」なのか、その背景を調査して記述してください。

② 「平均値」の再考

ニュースや業務報告で使われている「平均値」を一つ取り上げ、それが実態をミスリードしている可能性がないか考察してください。「もしヒストグラムを描いたら、どのような形になっていると推測されるか」を図示して説明してください。

③ シンプソンのパラドックスの構築

「全体で見ると A の傾向があるが、層別化すると逆の傾向になる」という架空の、あるいは実際のシナリオを一つ作成してください。（例：病院の手術成功率、学校のテストの平均点など、身近な例で構いません）。

④ なぜ数値要約だけではデータの真の姿を正確に把握することができないのでしょうか。

⑤ 探索的データ分析（EDA）と確証的データ分析の違いを述べよ。

第5講 統計学の基礎

尾関智恵（岐阜大学・准教授）

【学習到達目標】

- ① 基本的な統計量（平均値、中央値、最頻値、分散、標準偏差など）の意味と計算方法を理解し、適切に使い分けられる。
- ② データの分布や傾向を表すための代表的な統計的手法（ヒストグラム、箱ひげ図など）を理解し、実際に作成・解釈できる。
- ③ 確率の基本概念と、その応用例を理解し、日常や教育現場でのデータ解釈に役立てられる。

1. はじめに

統計学は、教育研究において客観的な根拠を提供し、教育実践の改善を支える重要なツールです。本講では、統計学の基礎的な考え方と、教育研究における統計の役割、標本と母集団の関係、仮説検定の考え方について学びます。

基本的には考え方の話になりますので、数式や具体的な計算といったものは出てきませんが、最終的に論文にまとめる、授業の学習効果や試みについて検討する際に注意すべき点についてお話しします。

2. 統計とは

2.1 統計の定義

統計とは「統（す）べて計（はか）る」ことです。

- ・「統べる」は、「多くのものを一つにまとめる」という意味
- ・「計る」は「ある基準をもとにして物の度合いを調べる」という意味

つまり、統計は必要な情報の全体を捉え、まとめる方法を提供します。多くのものを一つにまとめる、要約するということと、ある基準をもとに度合いを調べる、評価するということを行うのが統計です。

2.2 教育研究における統計の必要性

私たちが取り組んでいる教育研究に関して、統計の役割は近年非常に高まっています。

従来の実践研究では、授業をどのように実践したかということが主でしたが、そこで行われたことに具体的に、客観的にどのような効果があったかを示すために、数値で評価を表したり、その数値がどのように変容したかを捉えるために、統計処理がますます使われるようになってきています。

2.3 統計学の基礎知識

統計学は、データを収集・整理・分析し、そこから意味のある情報を引き出す学問です。教員が教育や調査の場面でデータを理解し、適切な判断を下すために不可欠な基礎知識です。

(1) 記述統計の基本概念（前提知識となるので動画では詳細説明を省略しています）

記述統計とは、データの特徴を要約し、分かりやすく表現する手法です。データの中心や散らばりを表す統計量について理解しましょう。

① 代表値（データの中心を表す指標）

平均値

データの合計をデータ数で割った値で、データの一般的な傾向を示します。最も一般的に使用される代表値ですが、極端な値（外れ値）の影響を受けやすいという特徴があります。

教育現場での活用例：

- ・ クラスのテスト平均点を算出し、全体的な理解度を把握する
- ・ 学習時間の平均を計算し、適切な課題量を検討する

注意点：

一人だけが極端に高い点数（または低い点数）を取った場合、平均値が実際のクラスの状況を適切に表さないことがあります。

中央値

データを小さい順に並べたときの中央の値です。外れ値に影響されにくい特徴があります。データ数が奇数の場合は真ん中の値、偶数の場合は真ん中の2つの値の平均を用います。

教育現場での活用例：

- ・ 極端な得点がある場合に、クラスの「典型的な」得点を知りたいとき
- ・ 所得や学習時間など、偏りが大きいデータの代表値として

最頻値

最も頻繁に出現する値で、カテゴリーデータや離散データの代表値として用いられます。

教育現場での活用例：

- ・ クラスで最も多い出身地を知りたいとき
- ・ アンケートで最も選ばれた選択肢を特定するとき
- ・ 「多くの生徒が選んだ」傾向を把握したいとき

② 散らばりを表す指標

分散

各データと平均値との差の二乗平均であり、データのばらつきの大きさを示します。分散が大きいほど、データは平均値から離れて散らばっていることを意味します。

教育現場での重要性：

- ・ クラス内の学力のばらつきを数値化できる
- ・ 個別指導の必要性を判断する材料になる
- ・ 習熟度別指導の編成に活用できる

標準偏差

分散の平方根で、データの散らばり具合を元のデータと同じ単位で表します。平均値と標準偏差を組み合わせることで、データの分布をより詳細に理解することができます。

教育現場での活用例：

- ・ テストの平均点と同じ 2 つのクラスでも、標準偏差が異なれば学力のばらつきが違ってくる
- ・ 標準偏差が大きい→個人差が大きい→個別対応が必要
- ・ 標準偏差が小さい→比較的均質→一斉指導が効果的

(2) データの可視化

データの分布や偏りを理解するために、視覚的な表現方法が用いられます。

ヒストグラム

データを一定の範囲（ビン）に分け、その範囲内のデータ数を棒グラフで表すものです。

ヒストグラムで把握できること：

- ・ データの分布の形状（正規分布、偏った分布など）
- ・ データの偏り
- ・ 外れ値の有無
- ・ データの集中傾向

教育現場での活用例：

- ・ テストの得点分布を可視化し、問題の難易度が適切だったかを判断

- ・ 高得点と低得点に二極化していないかを確認
- ・ 特定の得点帯に多くの生徒が集中していないかをチェック

箱ひげ図（ボックスプロット）

データの最小値、第一四分位数（Q1）、中央値（Q2）、第三四分位数（Q3）、最大値を箱とひげで表します。

箱ひげ図の構成要素：

- ・ 箱の部分：データの中央 50%が含まれる範囲（四分位範囲）
- ・ 箱の中の線：中央値
- ・ ひげの部分：外れ値を除いたデータの範囲
- ・ 外れ値：ひげの外にある点

箱ひげ図の利点：

- ・ データの散らばりや偏り、外れ値を一目で理解できる
- ・ 複数のクラスやテストの結果を並べて比較しやすい
- ・ 経年変化を視覚的に把握しやすい

教育現場での活用例：

- ・ 複数クラス間の成績を比較する
- ・ 同じクラスの前期と後期の変化を見る
- ・ 特に配慮が必要な生徒（外れ値）を発見する

(3) 確率の基本概念

確率は、ある事象が起こる可能性を数値で表したもので、0 から 1 の範囲で示されます。確率 0 は「絶対に起こらない」、確率 1 は「必ず起こる」ことを意味します。例えば、コインを投げたときに表が出る確率は 0.5（50%）です。

確率の基本的なルール

① 排反事象の確率の和

同時には起こり得ない事象（排反事象）の確率は、それぞれの確率の和に等しくなります。

例：サイコロで 1 または 2 の目が出る確率 = 1 の目が出る確率 + 2 の目が出る確率

$$= 1/6 + 1/6 = 2/6 = 1/3$$

教育現場での応用：

- ・ クラスで成績が A 評価または B 評価の生徒の割合を求める
- ・ 複数の選択肢のうちどれか一つを選ぶ確率を計算する

② 独立事象の同時確率

互いに影響を与えない事象（独立事象）が同時に起こる確率は、それぞれの確率の積に等しくなります。

例：コインを 2 回投げて両方とも表が出る確率 = 1 回目に表が出る確率 × 2 回目に表が出る確率

$$= 0.5 \times 0.5 = 0.25$$

教育現場での応用：

- ・ 2 つの異なるテストで両方とも高得点を取る確率を推定する
- ・ 複数の条件を同時に満たす生徒の割合を予測する

これらのルールは、教育現場や調査結果の解釈においても重要です。

(4) 統計的手法の教育現場での活用

これらの統計的手法や確率の知識は、データの正しい解釈や意思決定に役立ちます。

テストデータの分析

- ・ テストの平均点だけでなく、標準偏差も確認することで、クラス内の学力のばらつきを把握できます
- ・ ヒストグラムを作成することで、得点分布の偏りや二極化などの傾向を視覚的に理解できます
- ・ 箱ひげ図を用いることで、クラス間の比較や経年変化を効率的に把握できます

偏差値の理解

平均値と標準偏差を用いて、個々の生徒の相対的な位置づけを理解できます。偏差値は、平均を 50、標準偏差を 10 とした標準化された得点です。

偏差値の計算の考え方：

- ・ 平均点を取った生徒の偏差値は 50
- ・ 平均より 1 標準偏差高い得点の偏差値は 60
- ・ 平均より 1 標準偏差低い得点の偏差値は 40

異常値・外れ値の発見

箱ひげ図やヒストグラムを用いることで、通常のパターンから外れたデータを見つけることができます。

外れ値が示すもの：

- ・ 特別な配慮が必要な生徒の発見
- ・ 測定上の問題（記入ミス、集計ミスなど）
- ・ 特異な学習パターンの把握

調査結果の確率的解釈

アンケート調査などの結果について、確率の考え方をを用いて解釈することで、より適切な判断ができます。

例：

- ・ある選択肢を選んだ生徒の割合から、母集団全体の傾向を推測する
- ・複数の要因が組み合わさって起こる事象の確率を計算する
- ・ランダムサンプリングで得られた結果の信頼性を評価する

これらの知識は、教育の質向上や改善策の立案に直結します。データを適切に理解し、可視化し、解釈することで、エビデンスに基づいた教育実践が可能になります。

3. 教育研究における統計の役割

3.1 統計活用における注意点

教育研究において統計を使用する際には、いくつかの重要な注意点があります。

(1) 数値に振り回されない

例えば、従来であれば、教員もしくは学習者の感覚的なところで「学力を上げるためには国語力があつた方がいいよね、だから読書の時間を増やすことが学習の効果を上げるのではないだろうか」というような仮説、授業の考え方をもとに授業デザインを行っていきます。

それを客観的に示すために、例えば読んだ冊数が多い人には評価を与えようというようなことをしがちでしたが、これらの数値を集計してまとめたとしても、本当にそれが学習効果があつたのかどうか、因果関係も含めて説明するというのはまだまだ難しいのです。

(2) 学習者の主体性への配慮

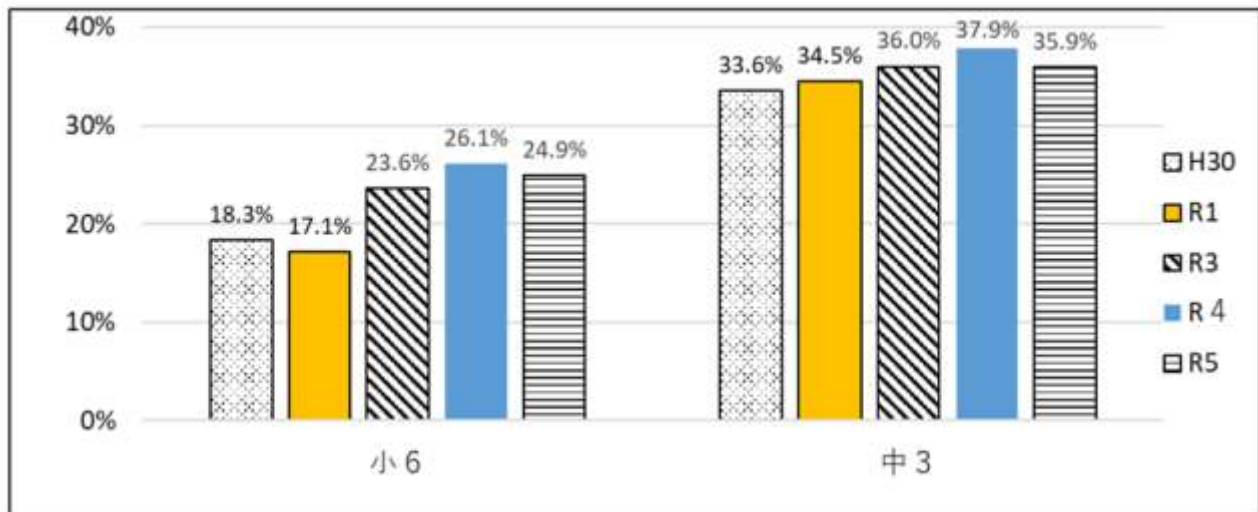
主体的に学習者が学びを構築していくということが、アクティブラーニングや最近の個別最適化の学習においては重視されています。しかし、数値にとらわれてしまうと、例えば学習者自身があまり面白くないとか、その読書自体の魅力というものについて離れていってしまうという危険性もあります。

教育研究における統計学の使用というのは慎重になる、使いどころをよく考えなくてはいけないということが注意点です。

3.2 データから学習者を深く知る

岐阜県の「1日に読書を全くしない児童生徒の割合」のデータを例に考えてみましょう。

【図表 11 岐阜県の「1日に読書をまったくしない児童生徒」の割合】



令和5年度 全国学力・学習状況調査（岐阜県の回答集計結果）

小学校6年生は大体20%から25%の間で、中学校3年生になりますと30%以上、40%までいかない割合が読書をしていないという事実を確認することができます。この結果から、読書の環境を整えよう、読書をもっと増やすように対策しなければと考えがちになってしまいますが、教育研究においては、この数値から見て取れる学習者の状態を考えることの方が重要です。

例えば：

- ・小学校6年生と中学校3年生で10%ほど差がありますが、なぜこれが起こるのか
- ・平成30年から令和5年までの調査で、年によっても特に小学校6年生は変動していますが、こういった要因で差ができているのか

数値的な結果に対して学習者をさらに深く知るためのツールとして使うべきであって、この数値が足りないから何とかしなくてはいけないというような性急な考え方をしないようにというのが注意点です。

また、中学校3年生は他の活動や受験勉強など時間的な問題もあり、読まないという要因があると思いますが、この時期に読書ができないということが、その後の子どもたちにとってどういった影響を及ぼすのか。これらは学習者それぞれ一人一人の状況にもよりますが、中学校3年生として一つのグループとしての傾向がもし見られるのだとしたら、それに対してどう対応するかという新しい議論を作るための根拠としてこういったデータを使うべきです。

3.3 教育研究における統計の3つの役割

教育研究において統計は、ただ単に実際行った授業や学習実践に対して良い悪いというのを決めるものではありません。よりよく授業を構築していく、刷新していくために何が必要なのか、学習者にとって必要なものは何なのかということを決めるために使うべきです。

(1) 客観的な根拠を提供する

「この学習活動は効果がありそう」という印象を、数値で検証できます。個人の経験や直感だけでなく、データに基づいた判断ができます。

読書は学力に役に立つのではないかなというような曖昧な経験や直感だけではなく、データとして読書の時間を確保してあげた方がいいというようなことを示すことができます（ただし、先程のデータでは説明しきれいでないので、調べるのであれば、そういったところを関連付けてわかるようなデータを集める必要があります）。

(2) 一般化可能な知見を得る

目の前の30人のクラスでの結果が、他のクラスや学校でも当てはまるかを推測できます。特定の文脈での観察を超えて、より広い適用可能性を議論できます。学習効果に直結するようなデータを集めるのは難しいですが、関連がありそうな要因をまとめた調査をまとめていくと、一般化が可能な知見を得ることができます。30人の一つのクラスで効果があったのであれば、他のクラスや他の市町村の学校でも役に立つ学習方法として伝えることができ、学習理論の一つを作ることができる可能性があります。

(3) 複雑な現象を整理・理解する

多くの要因が絡み合う教育現場で、何が重要な要因かを見極めることができます。データから意味のあるパターンを抽出したり、要因の相対的な影響力を比較したりできます。

学習現場は特に多くの要因が絡み合う場所です。学習者本人の状況はもちろん、学習環境、教室、教育リソース、コンテンツ、教員ができる時間やリソース、そういったものも全て絡んできます。そういったことを整理・理解するためにデータ収集し、まとめていくことは重要です。

ただし、数値が出た、その数値に振り回されて「あまり良くない」とか「課題点はこれだ」というふうに悲観的に捉えるものではありません。現状を把握して、よりよい学習を構築するために使うべきです。

それは教員がすごく負担を強いるということではなく、教員にとっても、時間的もしくは効率的ではあるけれども効果がある教授手法を見つけるために、こういったデータの蓄積はこれからますます重要になってきます。

4. 標本と母集団の関係

統計学とその解釈で学習の質を上げていくためには、基本的な用語、考え方をやはり押さえておく必要があります。

4.1 基本用語の整理

(1) 母集団

特徴や傾向などを知りたいと考える集団全体を母集団といいます。私たちが学習者と言っている場合には、先生方の担当する学年や学校の単位など、いろいろな規模がありますが、それを全体を指し示して言うために、全体調査と言われる全てを調査する必要があります。

(2) 標本調査

母集団の特徴を知るためにその一部を選び出し、調査を行う方法のことを標本調査といいます。授業を実施しながら、自分のクラス以外の人達、例えば校内の人達を全部調査するというのは基本的には難しいので、大体標本調査と呼ばれる手法をとることが多いです。

例えば、ある高校の中の一つのクラス、普通科のクラスのみに調査を実施し、これがその高校の普通科の傾向であるというふうに捉えるといったことを一般的にします。

(3) 標本（サンプル）

実際に調査を実施する母集団の一部を標本といいます。選び出すことを標本抽出といいます。

(4) 標本の大きさ

標本として選び出される個体数を標本の大きさといいます。

4.2 標本と母集団の関係

標本調査から母集団の性質を正しく推計するためには、母集団の情報が標本に正しく反映される必要があります。つまり、標本が母集団の「縮図」になっていることが望ましいのです。例えば、国内のある高校、中学校、小学校の生徒・児童に対しての傾向を見たいという場合、全国調査する場合もありますが、先生が考えられた実践授業や研究授業に関しては、全国の人達にするのは実際難しいので、標本として自分の学校のあるクラスを取り上げます。

小学校 3 年生に実施した標本調査に対してどのような効果があったかどうかを検証するという使い方をします。



4.3 標本選択の注意点

小学校、中学校、高校といったような枠組みであれば、母集団と標本との差というのは大きく分かれるということはあまり考えにくいところです。しかし、自分たちが知りたい学習者の全体像が、実際に授業を施すクラス（標本対象）の縮図になっているというのが研究の中では望ましいものとなっています。

実践授業を行い、ある一定の効果が現れ、それを例えば県内で展開していくといったときに、その効果を保証するために、標本と母集団の関係が著しく外れているとなりますと、保証ができなくなりますので、その部分を気をつけなくてはなりません。

普通科の学生や、他の職業科、工業科の学生では、それぞれクラスによって特性が違ってきますので、実際に授業するクラスをどこにするのかということを慎重に決めなくてはならないのです。

4.4 データの分布とばらつき

統計では、分析の対象となるデータがどんな値をとり、どのように、どの程度ばらついているかを知ることにより、データの全体的な姿を捉えることができます。このデータのばらつきの様子を「分布」と言います。

実際に標本として適切であろうクラスに対して授業をやったとき、良い成果が出た場合、例えば平均点をとって論ずる際に注視すべきことがこのばらつきです。このばらつきがあるかないかというところが、出てきた統計量（平均値や標準偏差といった代表値）の取り扱いに非常に影響してきます。

ばらつきの重要性（具体例）

例えば、平均が 37.7 点の確認テストでした。ここでやった授業が効果があったかなかったかと言いますと、一人の子は 90 点という非常に良い点数を出していますが、他の子はあまり良い点数でなかったという場合を考えてみましょう。

この場合、この実践授業が役に立つのか立たないのかということを論ずるときには、このばらつきがあるかどうかを慎重に確認する必要があります。

残念ながら、この場合はクラス全体に対して学習効果を等しく与えることができていない可能性が考えられます。ただし、一人の学生に対しては学習機会として有用であったというような見方もできます。できた、できてない、良い悪いという考え方ではなくて、何が要因でその結果になったかというところをよく深く考えていただく必要があります。



平均点：37.7点

5. 研究アプローチ：観察と介入

教育研究においては2つのアプローチが考えられます。特に学習科学やデザインアプローチの研究では、以下のような方法があります。

5.1 観察するアプローチ

実際に授業をやった、新しい方法でやったときに、生徒さんたちがいろいろな反応をします。この反応をまず冷静に観察するという手法です。

5.2 介入するアプローチ

デザインとして、例えば学習活動を半分の子たちには実際にやってもらい、半分の子たちにはその学習活動を入れずに効果を見るというような介入するアプローチです。

5.3 両アプローチに共通する重要点

いずれにせよ、学習活動の効果を見るためにこういった差をつけるわけです。いずれにしても、やはり先ほどから述べている母集団と標本、自分たちが知りたい学習者全体、それを代表する人たちとして、この実践授業をやるのに値するふさわしいサンプリングがなされているかどうかは、実施前に授業を設計するのと同様に考える必要があります。

6. 仮説検定の考え方

実践授業や研究授業において授業の設計をし、その狙い通りに授業の成果が出るかどうかを検討する際に、例えばテストの点数を使う、満足度調査を行うなど、様々な手法がありますが、そのときにはこの仮説検定の考え方を少し背景に入れておく必要があります。

6.1 仮説検定とは

仮説検定とは、統計的仮説の有意性の検定のことをいいます。仮説のもとで期待した値と観測した結果との違いを、これらの差が単に偶然によって起こったものかどうかという観点から、確率の基準で評価します。

統計において仮説検定は、仮説のもとに期待した値と観測した結果の違いが偶然に起こっているのか、実際に施した介入もしくは授業の内容によって起こっているものなのか、これを判定するために使うものです。

仮説検定の目的は、母集団について仮定された命題を標本に基づいて検証することです。

6.2 教育研究における仮説検定の注意点

薬品を使ったような実験や心理学のような刺激を一定数与える群と与えない群を比較する場合には、仮説検定をよく使いますが、教育研究のような学習者に対して施す場合には、少し気をつけなくてはならない点があります。

学習活動のある一つのグループには与えて、もう一つに与えないというのは、学習の平等機会を損なうことになります。この点に関しては、例えば最初に学習活動を与えるグループを先に実施し、その後に学習活動をしていないグループは後半で学習活動を体験するといったように、全員が等しく学習ができるように、その新しい指導方法を体験できるようにするという配慮が必要です。

6.3 仮説検定の具体例

ある学習理論に基づいた学習活動を組み込んだ方が、従来の授業よりも学習効果があるかを検討する場合を考えてみましょう。

- ・ 学習活動を組み込んだクラス：確認テスト平均 80 点
- ・ 従来の授業方法のクラス：確認テスト平均 75 点

この平均点の「差」は学習活動による効果か？ それともたまたまの結果なのか？

仮説検定は、この平均点の差が学習活動による効果なのか、それとも偶然に起こったことなのかを判定するためのツールです。

したがって検定で有意差が出たからといって、それが強い証拠になるということではないというところが、注意点です。

6.4 仮説の設定

授業研究においては、対立仮説と帰無仮説の2つを作ります。

(1) 対立仮説（証明したいこと）

- ・「学習活動には効果がある」
- ・「2つのグループに差がある」

(2) 帰無仮説（反対の仮定）

- ・「学習活動には効果がない」
- ・「2つのグループに差はない（差は偶然）」

この帰無仮説を棄却する、帰無仮説が認められないということが検定として出せるかどうかというところが仮説検定の考え方です。

6.5 有意水準の決定

有意水準とは、帰無仮説が間違っていると判断する（帰無仮説を棄却する）基準となる確率のことです。

- ・有意水準 0.05 に設定した場合、5%以下の確率で生じる現象は、非常にまれなことであるとしします
- ・有意水準は、0.05 や 0.01 が多く使われています

ある授業をやって、確認テストがある正規分布を出しています。裾野にあるような高い点数、低い点数が出る確率が5%以内、もしくは1%以内であるというように、この赤いところ（分布の両端）が起こる確率が非常に低いということを導くのが仮説検定の計算の考え方です。

6.6 なぜ背理法的アプローチを使うのか

なぜこんな回りくどいことをするのでしょうか？

- ・「効果がない」と仮定して、その仮定が疑わしいことを示す
- ・直接「効果がある」を証明するより確実

効果がないと仮説を設定して、その仮説が疑わしいということをまず示します。その疑わしいことを標本の数値をもって限定して、直接効果があるというようなことを言い張るよりも、逆説的に説明を組み立てていきますので、信憑性が高いというような理解でよく使われている手法です。

差があるということに意味があるのか、やった教授法が効いているということに説得性を持たせるために、このような回りくどいことをしているのです。

6.7 p 値の理解

統計検定をすると必ず出てくる数値に p 値があります。

p 値 = 「偶然でこの結果が起こる確率」

帰無仮説「学習活動を組み込むことによって効果（差）はない」が正しいと仮定し、その仮定のもとで、今回の結果（5 点差）が偶然起こる確率を計算したのが p 値にあたります。

p 値が低い：偶然起こる可能性は低い

p 値が 0 に近いほど、帰無仮説で起こることが偶然として起こる可能性は低いということを示すため、この値が低い、つまり帰無仮説が棄却される、だから授業中にやった学習活動は意味があるんだというような論理立てで使います。

7. 解釈の際の注意点

7.1 授業デザインの重要性

仮説検定では、あくまでも授業デザインがしっかりできていることが前提となります。ある学びを促進する、補強するためにはこういった教材を入れるべきである、こういった学習活動を入れることが学習理論に適っているというような設計がしっかりされた上で、これを入れたからこそ効果が上がった、なかったというような仮説検定ができるわけです。授業デザインをしっかりしておくことが効果検討には必須です。

7.2 多要因の慎重な検討

多くの要因が絡み合う教育現場で、何が重要な要因かを見極める際には、慎重にデータを検討していく必要があります。

そうはいつてもその教材が効いている、その学習活動が効いているということを言いたいところですが、学習活動・教室の中には多くの要因が絡み合います。学習者自体も既習の知識を持っている、知識や状態も違うというのは何回も申し上げている通りですが、出てきた結果、統計検定で例えば有意な差が出たとしても、その結果は一つの結果として、他に要因が絡んでいるのは明確です。

慎重にデータを検討して解釈を導く必要があるというのが、教育研究においては非常に注意すべき点です。

8. 統計的有意 ≠ 実践的重要性

8.1 重要な区別

p 値は「差が偶然かどうか」を教えてください。しかし「その差が重要かどうか」は別問題です。統計的に有意な差はあったということは、この授業に非常に価値があるということで、飛びつきやすい、わかりやすい数値ではあります。

しかし、それは授業実践で他の学校、他の教室、他の道具立てでやったときに本当に効果があるのかという一般化に関わってくるところには影響がありません。実践的な重要性和統計的なサンプル・標本でやったときの結果を安易に結びつけるということはしないようにしていただきたいと思います。

8.2 適切な根拠の選択

それはあなたが検討したい効果を示す根拠として適切ですか？

例えば、収集したデータが授業の満足度なのだとしたら、それは学習の効果ではなくて、学習者の自己満足度が高くなったということで、本当に何か学んだのかというようなところに直結しません。

あなたが調べたいところが根拠として適切かどうかというところをきちんと踏まえた上で、収集するデータの選定からしなくてはいけないというところをもう一度よく考えておいていただきたいと思います。

9. まとめ

統計は必要な情報の全体を捉え、まとめる方法を提供する便利なツールです。教育研究においては、効果や影響を説明するために以下の役割を果たします：

- ・客観的な根拠を提供する
- ・一般化可能な知見を得る
- ・複雑な現象を整理・理解する

また、記述統計の基本概念として、データの中心を表す代表値（平均値、中央値、最頻値）や散らばりを表す指標（分散、標準偏差）、データの可視化手法（ヒストグラム、箱ひげ図）、確率の基本概念を学びました。これらの知識は、教育の質向上や改善策の立案に直結します。ただし、そういった便利なツールを使うためには、以下の点に注意が必要です：

(1) 標本と母集団の関係

どう実践授業をするクラスで得た結果が、すぐ他のクラス、他の学校で使えるかというところというわけではありません。では、そこで得られたデータが学習の効果や品質、質を上げるためにどう役立ったかを十分慎重に検討するために、こういった客観データは非常に重要です。標本と母集団の関係、統計手法によって得られる結果を慎重に検討する必要があります。

(2) 仮説検定のアプローチ

仮説検定は非常に便利です。このアプローチを使うためには、学習理論に基づいた授業デザイン、狙いがあって、どういう道具立てでやるかという設計理念がしっかりしている場合に有用に使えます。

仮説検定のアプローチを行うには学習理論に基づいた授業デザインが必須です
(目標とそれを実現する要因・方法を明確化)。

(3) 結果の慎重な解釈

導出された結果に有意差があったから効果があったと判断するのは要注意です。
人の学習には複雑な要因が絡んできます。一つの結果としては解釈するけれど
も、他の要因も慎重に検討する必要があります。導出された結果と解釈は慎重に
検討する必要があります。学習者自身をさらに深く知る、授業というものの醍醐
味をもっと深めるために、この統計ツールは使われるべきです。

参考文献

- ・ 総務省統計局, レベル別テキスト,
<https://www.stat.go.jp/dss/getting/textbook.html>
- ・ 岐阜県, 岐阜県子どもの読書活動推進計画 (第五次) ~読もう! つなごう! ひ
ろげよう読書活動~, <https://www.pref.gifu.lg.jp/page/2083.html>
- ・ 総務省統計局, なるほど統計学園,
<https://www.stat.go.jp/naruhodo/index.html>
- ・ ソコスト, <https://soco-st.com/>

課 題

- ① データの平均値、中央値、最頻値の違いと、それぞれの特徴について説明し
てください。
- ② 以下のデータセット (例: 5, 7, 8, 8, 9, 10, 12) について、分散と標準偏
差を計算し、その意味を説明してください。
- ③ コインを 10 回投げたときに表が出る確率は 0.5 です。このとき、実際に表
が 7 回以上出る確率について二項分布を用いて計算し、その結果から何がわかる
か説明してください。
- ④ 教育研究において統計を活用する際、数値に振り回されず学習者を深く理解
する方法は何か。
- ⑤ 仮説検定において帰無仮説を立てる目的を説明せよ。
- ⑥ 仮説検定において、なぜ背理法的アプローチを用いるのか。

第6講 機械学習の基本概念

澤井進（岐阜女子大学・特任教授）

【学習到達目標】

- ① 機械学習の基本的な仕組みと種類（教師あり学習、教師なし学習、強化学習）を理解し、それぞれの特徴と適用例を説明できる。
- ② モデルの学習過程（訓練、検証、テストの流れ）と、その目的や重要性を理解し、適切なモデル評価指標（正確率、精度、再現率など）を選択できる。
- ③ 過学習やバイアス・バリエーションのトレードオフについて理解し、モデルの汎化性能を向上させるための基本的な対策を説明できる。

1. 機械学習とは

機械学習とは何か？データからパターンを学び、予測や分類に活用する技術です。教育、医療、金融など幅広い分野で活用可能です。

機械学習を用いて、脳活動から心の中でイメージした内容を画像化することもできます。漫画「ドラえもん」で、眠っている人の頭の中で何を見ているかというようなこと見られる機械が登場しましたが、機械学習を用いて脳活動から心の中でイメージした内容を分類し画像化することができるようになったわけです。

機械学習の仕組みはどうなっているかというと、人間の脳の神経細胞が相互に接続した構造を模擬したのになります。機械学習には、「教師あり学習」、「教師なし学習」、「強化学習」の3つの仕組みがあります。

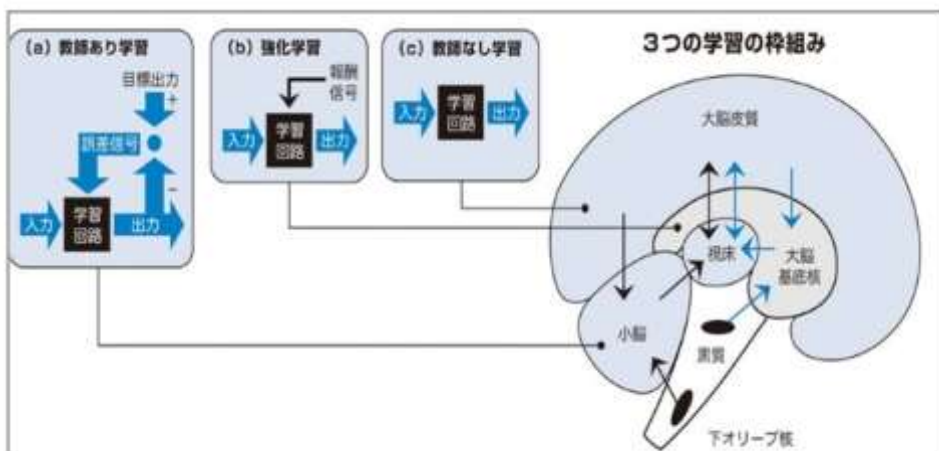


図1 3つの機械学習の仕組み

2. 教師あり学習

教師あり学習は、正解データで学習する手法です。例えば、テスト結果から合否を予測したり、画像から猫か犬かといった正解を教えて学習させたりします。機能的には小脳が担う学習機能です。視覚野で神経活動が活発になっている部位を「fMRI」で測定して、特徴量予測デコーダー（解析器）で予測パターンを認識させると、事前に用意した大規模画像との比較から動物の「ヒョウ」だと分かります。fMRI（磁気共鳴機能画像法）は、医療用のMRI装置を使って無害に脳活動を調べる方法です。

教師あり学習では、図1に示すように正しい入出力の組合せを与え、「目標出力」との誤差信号を学習することで、新規の入力に対し、適切な出力が出せるようになります。

教師あり学習の代表的手法は、誤差信号を入力と逆に伝搬させるので、誤差逆伝播法(Back Propagation)と言われます。正解、若しくは誤りを入力として、未経験入力に対する意志を決定する決定木(Decision Tree)の作成なども使われます。

教育の場合、教師あり学習を用いて、数例の模範回答で記述式問題を採点することができます。

図2の教師あり学習の例では、数例の模範解答で採点をします。機械学習の立場から「フュー・スロット学習」と言います。ごく少ないデータからAIがモデルを学習して高い精度で予測が行われるようにする機械学習の手法です。例えばプロンプトとして「あなたは有能なAI教師です。記述式問題「超AI時代における教師の役割を説明しなさい。模範回答は以下の通りです」ということで、模範回答を事前に作っておいて教え込みます。

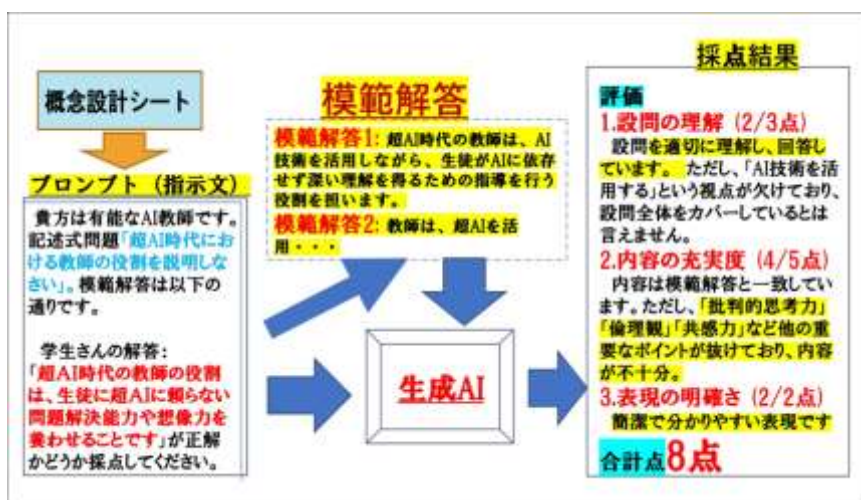


図2 教師あり学習の例：数例の模範解答で採点

学生さんの回答が「超 AI 時代の教師の役割は、生徒の超 AI に頼らない問題解決能力や想像力を養うことです」とします。それに対して採点してくれ、総合点は 8 点となります。内訳は、「設問の理解」については 3 点中の 2 点、「内容の充実度」については 5 点中 4 点、「表現の明確さ」は 2 点中 2 点と、採点し評価理由を付けてくれます。

3. 教師なし学習

教師なし学習は、正解ラベルなしでデータの構造を抽出する手法です。例えば、学習スタイルに基づく生徒のグループ分け（クラスタリング）などが挙げられます。

教師なし学習は、大脳皮質が担う学習機能で、人間が何も教えなくても機械が自律的に学習していきます。教師なし学習では、図 1 に示すように新規の入力に対し、機械が自律的に適切な出力を出します。

教師なし学習の代表的手法は、クラスタリングの K-means 法（k 平均法）です。

k-means は、指定した k 個のクラスタ（集団）にデータを分類していくアルゴリズムです。重心を求め、似た特徴の画像を近くに置くことで、グループ分けする方法です。グループ分けの結果から新しいアイディアを得ることができます。

4. 強化学習

強化学習は、環境との相互作用を通じて報酬を最大化する行動を学習する手法です。学習アプリが自動的に難易度を調整したり、ロボット制御や将棋のようなゲームに使われたりします。強化学習は、大脳基底核が担う学習機能です。

強化学習では、図 1 に示すように正しい入出力の組合せに対する報酬信号を学習することで、新規の入力に対し、適切な出力が出せるようになります。

強化学習を用いるロボット犬「A i b o」では、最終結果若しくは、途中経過に対して、程度良かったかを示す「報酬信号」に基づき、例えば「ける」ことにより報酬をなるべく大きくするように探索をします。

5. 学習過程と評価指標

学習は、訓練、検証、テストという流れで進められます。

- 1) 訓練: データでモデルを学習します。
- 2) 検証: パラメータ調整やモデルの選択を行います。
- 3) テスト: 新しいデータで性能評価を行います。

図3に示すように、学習過程の評価指標には、正当率（Accuracy）、適合率（Precision）、再現率（Recall）、そしてF値（バランス指標）といった4つの重要な指標があります。

それぞれ以下のような意味を持ちます、

- 1) 正当率: 全体のデータのうち、正しく正解を検索（分類）できた割合です。
- 2) 適合率: 正解として検索されたデータのうち、実際に正解だった割合です。
- 3) 再現率: 実際に正解のデータのうち、正しく正解と検索できた割合です。
- 4) F値（バランス指標）: 適合率と再現率のバランスを示す指標です。

これらの評価指標は、図3のような数式で表現できます。

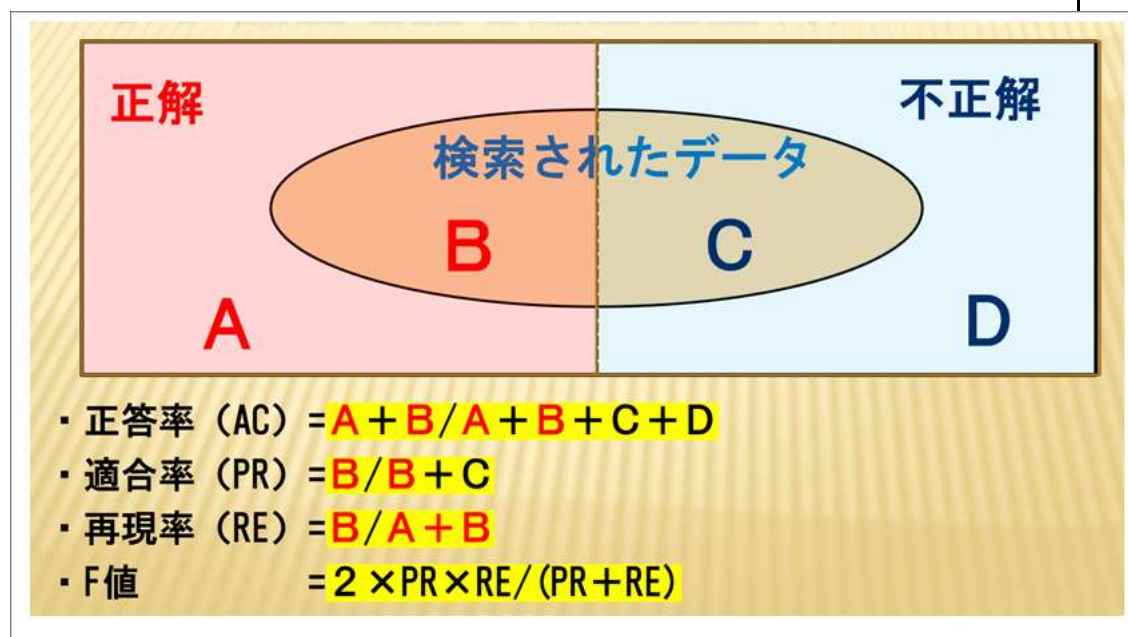


図3 学習過程の4つの評価指標

6. 過学習

過学習とは、モデルが訓練データに過剰に適合し、新しいデータに対する性能が低下する現象です。模試だけ強くて実力テストでは点が取れない生徒に例えることができます。

このような過学習を防ぐには、以下のことが挙げられます：

- 1) データを増やす：学習不足が解消します
- 2) 正則化：バリエーションが過剰に大きくなりすぎない程度に、バイアス（偏り）を抑制します
- 3) 交差検証：モデルの性能を過小適合と過剰適合の観点から測定します。

図4は「最適解」がバイアス・バリエーション曲線の交わる点であること、バイアス・バリエーションのトレードオフがあることを示しています。

学習不足では偏り誤差が多くバイアスが高くなり、学習量が多すぎると「過学習」となり、ばらつき誤差が増大し、バリエーションも高くなります。

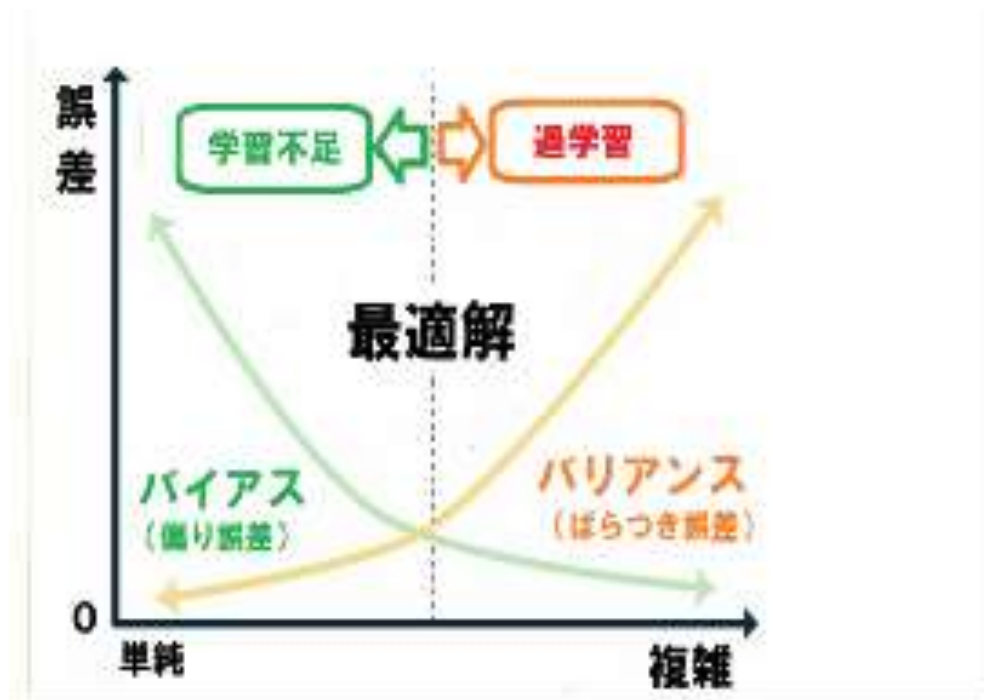


図4 バイアス・バリエーションのトレードオフ

これらのことから、

- 1) 単純すぎるモデルは、バイアスが高く、偏り誤差がでます。
- 2) 複雑すぎるモデルは、バリエーションが高く、ばらつきが多く、誤差や過学習リスクが出ます。
- 3) 適切なバランスが最適解となります。

6. まとめ

まとめますと、今後

- 1) データサイエンスは社会・教育の両面で重要性が増すので、
- 2) 教育者は未来を見据えた教育を実践する必要があります。
- 3) 未来の教育は、単なる知識の伝達だけでなく、データを活用した問題解決能力や倫理観を育むことが求められるだろうと言えます。

課題

- ① 機械学習の三つの主要な種類（教師あり学習、教師なし学習、強化学習）について、それぞれの特徴と代表的な応用例を説明してください。
- ② 過学習とは何かを説明し、過学習を防ぐための一般的な方法を2つ挙げてください。
- ③ 機械学習モデルの評価指標にはさまざまなものがありますが、正解率（Accuracy）と再現率（Recall）の違いについて具体的な例を用いて説明してください。
- ④ 学習モデルの性能を客観的に評価するために用いられる主要な指標は何ですか。
- ⑤ 教師あり学習における誤差逆伝播法の仕組みを記述しなさい。
- ⑥ 過学習の定義と、それを防ぐための具体的な対策を挙げなさい。

第7講

回帰分析と分類モデル

笹山和明（株式会社 村田製作所・情報科学アーキテクト）

【学習到達目標】

- ① 回帰分析と分類モデルの基本的な概念と違いを理解し、適切な場面で使い分けられるようになる。
- ② 回帰分析における代表的な手法（例：線形回帰）の仕組みと、その結果の解釈方法を説明できる。
- ③ 分類モデル（例：ロジスティック回帰や決定木）の仕組みと、その評価指標（例：正解率、再現率）について理解し、モデルの性能を適切に評価できるようになる。

1. データで未来・未知を予測する

データからパターンを抽出し、未来や未知のデータを予測する手法として、回帰分析と分類モデルが挙げられます。これらは、データサイエンスにおいて最も基本的かつ重要な予測技術です。回帰分析と分類モデルを理解することで、ビジネスにおける売上予測から迷惑メールの判別に至るまで、身近な課題をデータによって解決するための大きな一歩を踏み出すことが可能になります。

まず、回帰分析と分類の違いについて説明します。回帰分析は数値（連続値）を予測する手法で、分類はデータをカテゴリーに分ける手法です。目的が「数値」か「ラベル」かが主な違いです。下記に詳細な違いを示します。

項目	回帰分析 (Regression)	分類 (Classification)
目的	数値を予測する	カテゴリー（ラベル）を予測する
出力の例	価格、気温、売上などの連続値	スパム/非スパム、合格/不合格などのクラス
モデルの例	線形回帰、重回帰、リッジ回帰など	ロジスティック回帰、決定木、SVM など
評価指標	MSE（平均二乗誤差）、MAE、 R^2 など	Accuracy、Precision、Recall、F1 スコア など
数学的性質	出力は連続値（実数）	出力は離散値（カテゴリー）
使う場面	売上予測、気温予測、年収予測など	顧客の属性分類、病気の診断、画像認識など

2. 回帰分析

2.1 回帰の語源

「回帰」という言葉は、もともと「一周して元に戻る」という意味があります。そのため、たとえば後ほど紹介する単回帰分析の式 $y = \beta_0 + \beta_1 x$ を見ても、「どこが“元に戻る”のだろう？」と不思議に思われるかもしれません。回帰分析という言葉が使われるようになったのは、ある統計学者が親子の身長を分析し、極端な身長の子どもは平均身長に近づく傾向があることを発見しました。必ずしも遺伝せず、先祖返りのように平均値に戻っていく現象は「平均への収束→回帰」と呼ばれるようになりました。

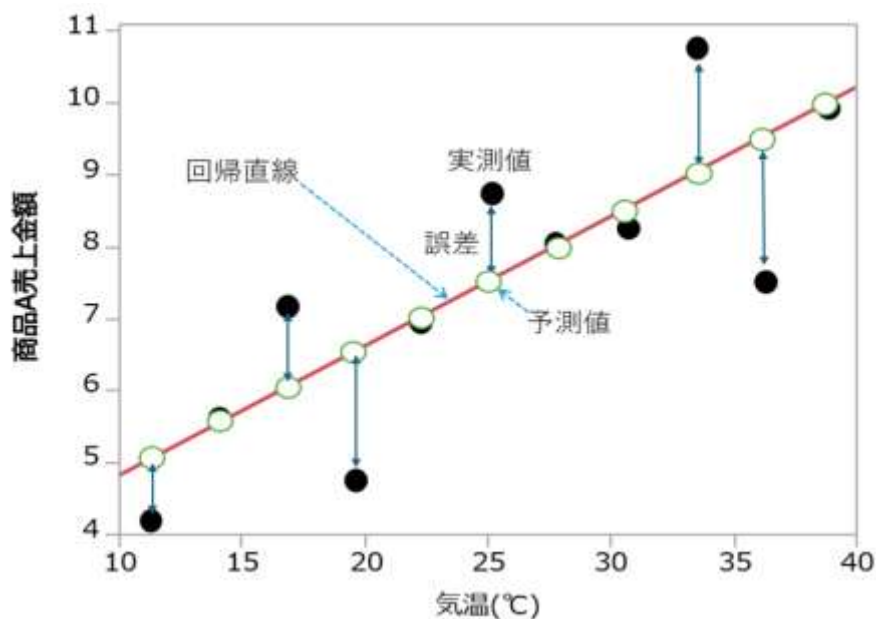
2.2 単回帰分析

最もシンプルな回帰分析の単回帰分析の特徴としては、説明変数 x が1つの場合で、次の式が回帰式（以下、モデル）となります。

$$y = \beta_0 + \beta_1 x$$

β_0 は方程式の切片とも呼び、 β_1 は回帰係数です。

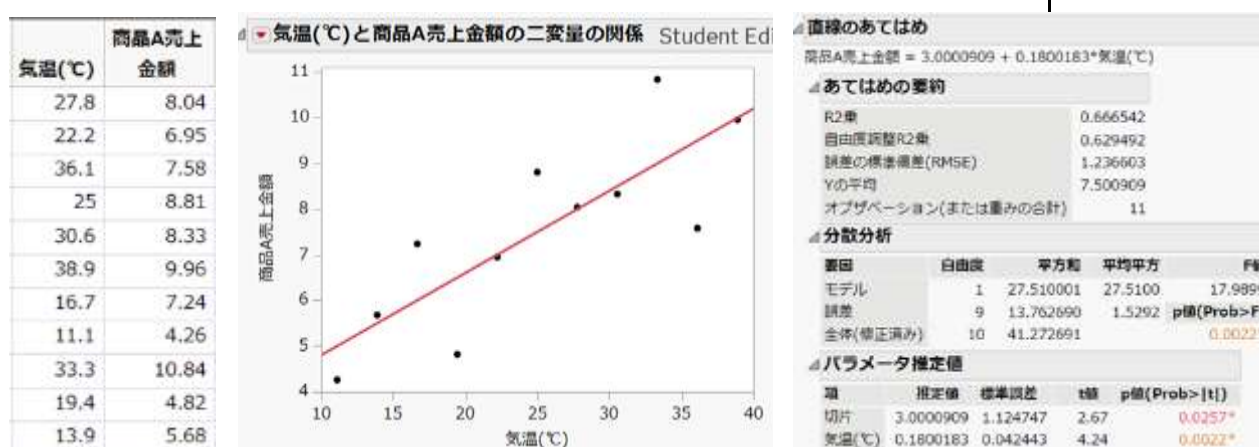
この β_0 と β_1 を推計する方法として使われているのが最小二乗法です。最小二乗法は、観測値と予測値の差である「残差」の二乗和が最小になるように β_0 と β_1 を推計する方法です。例えば、図の黒丸（●）は実際に測定された実測値で、その中に描かれている近似線（赤）は実測値との残差を最も小さくするために推計された予測値である白丸（○）を繋げた近似線、つまり回帰直線です。残差は実測値－予測値となります。



$$\beta_0 = (y \text{ の平均値}) - \beta_1 \times (x \text{ の平均値}) \quad \beta_1 = S_{xy} / S_{xx}$$

S_{xy} : x と y の偏差の積の総和 S_{xx} : x の偏差平方和 で計算することができます。

例として、下図（左）のデータ表において、「気温（℃）」から「商品Aの売上金額（百万円）」を予測します。データよりアプリケーションソフト（SAS社のJMP®）を使って計算した結果を下図に示します。このデータの場合、モデルは、商品A売上金額 = $3.00 + 0.18 \times \text{気温}$ となります。つまり、気温が1℃上昇すると、売上金額が18万円増えます。



尚、回帰分析においては、分散分析を行って、モデルが統計的に有意かどうか、残差の検定を行います。

2.3 重回帰分析

重回帰分析の特徴は複数の説明変数があることです。例えば、「面積」「築年数」「駅からの距離」などから「住宅価格」を予測したい場合、

モデルは、 $\text{価格} = \beta_0 + \beta_1 \times \text{面積} + \beta_2 \times \text{築年数} + \beta_3 \times \text{距離}$ となります。このように、複数の要因が絡む現実的な予測に向いています。

2.4 多項式回帰

多項式回帰の特徴は、説明変数のべき乗を使って、曲線的な関係を表現します。

例えば、面積のべき乗に比例して価格が変化する場合、

モデルは、 $\text{価格} = \beta_0 + \beta_1 \times \text{面積} + \beta_2 \times \text{面積}^2$ となります。直線では表せない複雑な関係に対応できます。

2. 5 リッジ回帰・ラッソ回帰

この2つの手法は、学習したデータに対しての精度は高いものの未知のデータに対しては同様の精度が出せない問題である「過学習」が起こりにくいように工夫されています。

2. 6 線形以外の回帰

決定木回帰、ランダムフォレスト回帰、サポートベクター回帰など、様々な回帰分析手法があります。線形では表せない複雑な関係に対応でき、精度は高いのですが仕組みは複雑です。

3. 分類モデルとは

3. 1 分類の種類

分類は、与えられたデータを事前に定義されたカテゴリーに分けることを指します。例えば、メールが「スパム」か「スパムでないか」を判別するスパムフィルターや、画像が「犬」か「猫」かを判別するモデルが分類の例です。代表的な方法にはロジスティック回帰と決定木があります。ロジスティック回帰は、確率を使って分類します（例：スパムの確率が70%ならスパムと判定）。決定木は、特徴量に基づいて「分岐」を繰り返し、最終的に分類を決める木のような構造のモデルです。直感的でわかりやすく、特徴量の重要度も見やすいのが魅力です。その他、サポートベクターマシン、決定木、ランダムフォレストなど、多くの手法があります。

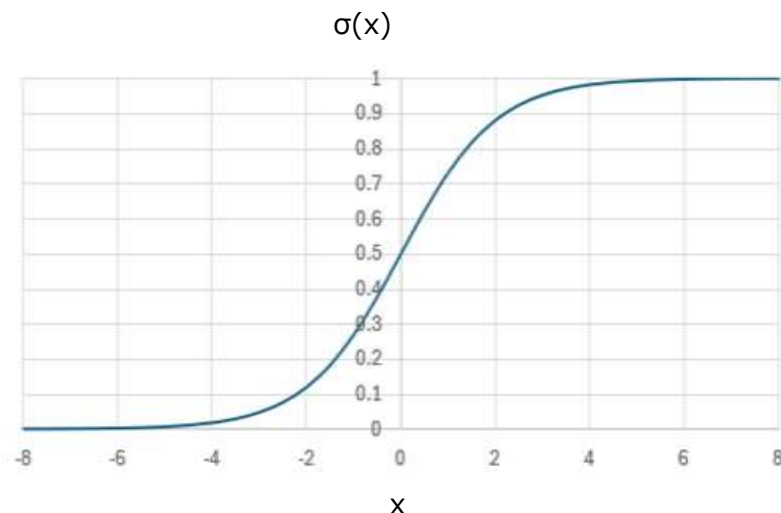
3. 2 ロジスティクス回帰

分類問題において、対象が「はい／いいえ」や「スパム／非スパム」といった二値のいずれかに属するかを判定する必要がある場面は多く見られます。こうした問題に対して有効な手法の一つがロジスティック回帰です。ロジスティック回帰は、線形回帰と似ていますが、ロジスティック回帰では、入力された特徴量の線形結合に対してシグモイド関数（ロジスティック関数）を適用することで、出力を0から1の範囲に収め、確率として解釈可能な値を得ることができます。シグモイド関数は以下の式で表されます。

$$\sigma(x) = 1 / (1 + e^{(-x)})$$

この関数の特徴は、入力値 x が大きくなるにつれて出力が1に近づき、逆に小さくなると0に近づくというS字型の滑らかな曲線を描く点にあります。

下図の横軸が x で、縦軸が $\sigma(x)$ です。 x が 0 のとき、 $\sigma(x)$ は 0.5 と 0 と 1 のちょうど真ん中になります。これにより、モデルの出力を「あるクラスに属する確率」として自然に解釈できるようになります。



3.3 決定木

ロジスティック回帰が数学的な確率モデルを構築するのに対し、決定木はまったく異なる発想で分類問題に取り組みます。そのアプローチは非常に直感的で、人間が意思決定を行うプロセスに似ているため、結果の解釈が容易であるという点に大きな価値があります。

決定木の本質は、「もし〇〇ならこちら、そうでなければあちら」といったシンプルな条件分岐を、木の枝分かれのように繰り返すことで、データを分類していく手法です。

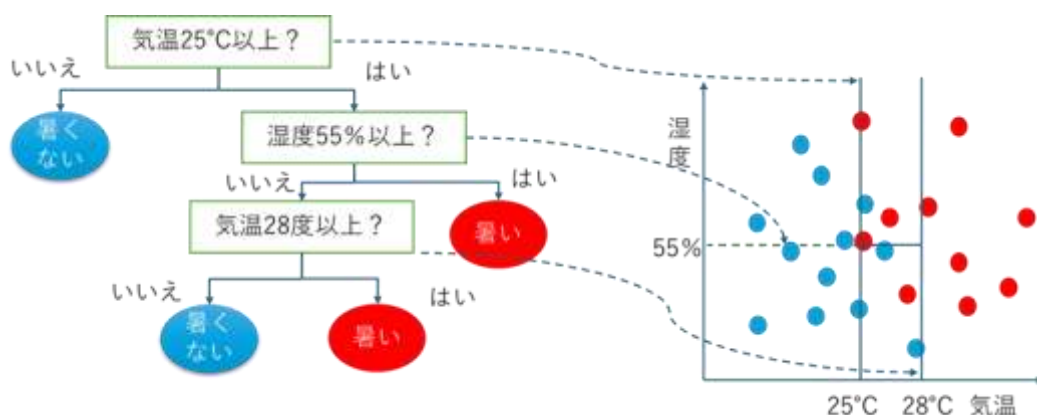
具体例として、A 君がある日に「暑いと感じるかどうか」を、その日の気温と湿度から予測するケースを考えてみましょう。A 君が「暑いと感じた日（赤点）」と「そうでない日（青点）」のデータが、気温と湿度を軸にした下図のグラフ上にプロットされています。

たとえば、気温が 26℃、湿度が 60% の日に、A 君は暑いと感じていました。この時点では、赤点と青点が混在しており、明確な分類パターンはまだ見えていません。

決定木は、混在したデータを最も効率的に分割できる質問（ルール）を自動的に探索し、見つけ出します。たとえば、まず全体に対して「気温は 25 度より高いか？」という質問が最も効果的であると判断し、データを 2 つのグループに分割します。

次に、それぞれのグループに対して、さらに最適な分割ルール（たとえば、片方のグループでは「湿度は55%より高いか？」）を再帰的に探索していきます。このようなルールを順に適用することで、もともと混在していたデータ群に明確な境界線が引かれ、各領域は「暑い日」と判断される可能性が高いエリアと、「そうでない日」と判断されるエリアに分類されます。なお、一部のデータ点は正しく分類されていません。この点については、次章「モデルの評価」にて詳しく説明します。

これらの分割ルールは、下図（左）に示すように、気温と湿度に基づいて「暑いと感じるかどうか」を判断するツリー構造として表現できます。また、この判断に基づき、下図（右）のグラフには分類境界線を加えています。



このように決定木は、複雑なデータの中から人間にも理解しやすい明確なルールを自動で構築し分類を行います。

4. モデルの評価

4. 1 概要

前述の回帰分析や分類モデルの性能を評価し改善するプロセスは、データ分析および予測モデル構築において重要です。例えば、正解率が90%だからといって、モデルが優れているとは限りません。100件のうち90件が『正常』で10件が『異常』なデータがあるとして、もし、すべてを『正常』と予測すれば、正解率は90%になりますが、異常を1件も検出できていません。これでは本来の目的を果たしていないのです。

特に分類モデルの評価においては、単なる正解率だけでなく、より詳細な性能を把握するための複数の指標が存在します。どの指標を重視するかは、解決したい課題の性質によって異なります。例えば、医療診断では病気の見逃し（偽陰性）を防ぐことが、スパムメール判定では正常なメールの誤分類（偽陽性）を防ぐことが優先されます。

4. 2 混同行列

混同行列は、分類モデルの予測結果を整理するための表で、以下は「2 クラス分類」の例です。

	実際が陽性 (Positive)	実際が陰性 (Negative)
判定が 陽性	真陽性 (TP: True Positive) 正しく陽性と予測 [スパムを正しくスパムと判定]	偽陽性 (FP: False Positive) 誤って陽性と予測 [正当メールを誤ってスパムと判定]
判定が 陰性	偽陰性 (FN: False Negative) 誤って陰性と予測 [スパムを誤って正当メールと判定]	真陰性 (TN: True Negative) 正しく陰性と予測 [正当メールを正しく正当メールと判定]

4. 3 評価指標の種類

主に、下表の4つが評価の指標としてよく利用されます。

正解率 (Accuracy) : 全体の中で、正しく予測できた割合

ただし、クラスの偏りがある場合 (例 : 陽性が少ない) には注意が必要です。

精度 (Precision) : 「陽性と予測したもの」の中で、実際に陽性だった割合

再現率 (Recall) : 「実際に陽性だったもの」の中で、正しく陽性と予測できた割合

F1 スコア : 精度と再現率のバランスを取った指標

精度と再現率の両方が高いときに高くなります。F1 スコアは 0~1 の間の値を取り、1 に近いほど良いモデルです。特に「不均衡データ (陽性が少ない) 」では、正解率よりも F1 スコアが信頼できます。

4. 4 正しい指標の選び方

4. 3に4つの評価指標を紹介しましたが、どのように使い分けすべきか、下表に示します。

指標	計算式	重視する点	ユースケース例
①正解率 Accuracy	$\frac{TP + TN}{TP + FP + FN + TN}$	全体の正解率	クラスが均等
②精度 Precision	$\frac{TP}{TP + FP}$	誤検出 (FP) を減らす	スパムメール検出
③再現率 Recall	$\frac{TP}{TP + FN}$	見逃し (FN) を減らす	病気検出
④F1 スコア	$\frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$	Precision と Recall のバランス	不均衡データ

5. まとめ

- ・データによる予測には、大きく分けて「回帰（数値予測）」と「分類（カテゴリー分け）」の2種類があります。
- ・どんな問題にも効く完璧な万能モデルは存在せず、解決したい問題に応じて適切な手法を選ぶ必要があります。
- ・モデルの性能は「正解率」だけでなく、ビジネス目的（どの間違いが最も致命的か）に合った指標で正しく評価することが最も重要です。

データサイエンスと聞くと難しく感じるかもしれませんが、最も重要な一歩は、技術そのものではなく、皆さんが「解決したい課題は何か?」「最も避けたい間違いは何か?」を自問することです。

課題

- ① 回帰分析と分類モデルの違いについて示してください。
- ② 回帰分析において線形回帰モデルを用いる場合、どのようにしてモデルのパラメータ（係数）を推定しますか？また、その推定結果の解釈について説明してください。
- ③ 分類モデルの評価指標の一つである F 値（F1 スコア）について、その意味と計算方法を具体的に説明し、なぜこの指標が重要となる場合があるのか例を挙げて説明してください。

第8講 クラスタリングと次元削減

小松尚登（滋賀大学）

【学習到達目標】

- ① クラスタリングの基本概念と代表的な手法を理解し、適切な場面での適用方法を説明できる。
- ② 次元削減の目的と代表的な手法（主成分分析（PCA）など）を理解し、データの可視化や前処理に役立てられる。
- ③ クラスタリングと次元削減の違いや関係性を理解し、実データ分析においてこれらの手法を適切に選択・適用できる。

1. 教師なし学習

第7講で取り扱った回帰分析や分類などの教師あり学習は、データに対して「正解」の情報が与えられていた。一方、本講で取り扱うクラスタリングや次元削減は教師なし学習の一種であり、「正解」の情報が無いデータから何かしらの規則や特徴を見出そうとするものである。これらの手法はデータの構造理解や可視化、前処理などに広く用いられている。

2. クラスタリング

第7講で扱った分類モデルは、ラベルが付与されたデータを学習し、各データに適切なラベルに分類していた。一方で、クラスタリングにおいては、ラベルが付与されていないデータが与えられ、それらを似た特徴を持つグループ（クラスター）に分類する。こうした手法は、ラベルがついていないデータにおいて、傾向を把握する上で重要となる。具体的なアルゴリズムとしては、以下に紹介するk-means 法や階層的クラスタリングが挙げられる。

(1)k-means 法

k-means 法は以下の手順によって、データを k 個のクラスターに分類する。

1. k 個の「クラスター中心」を定める

2. 各データ点を最も中心までの距離が近いクラスターに割り当てる
3. 各クラスターに関して、割り当てられたデータ点の平均値を新たなクラスター中心に定める

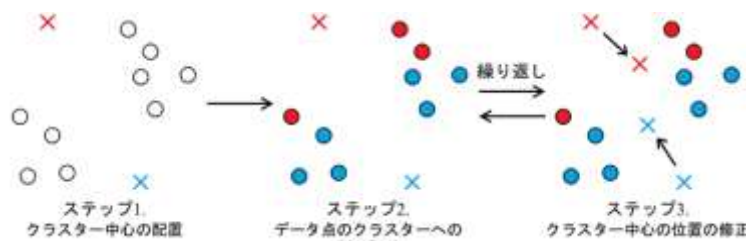


図1 k-means 法

4. 2~3 の手続きをクラスター中心の位置が収束するまで繰り返す
- この手順を図示すると、以下の図1 のようになる。なお、クラスター数 k をあらかじめ決めておく必要がある点には注意が必要である。

(2)階層的クラスタリング

階層的クラスタリングは以下の手続きによって、互いに「類似した」データから順にクラスターにまとめていく方法である。

1. 現時点で存在する全てのクラスターのうち、最も「類似度」が高い対を一つのクラスターに併合する
2. 新しいクラスターと他のクラスターの「類似度」を計算する
3. 1~2 の手続きを繰り返す

クラスターを統合していく過程で、図2のように木の枝分かれのような形のグラフが作られる。これをデンドログラム（樹形図）と呼ぶ。

上記の手続きはクラスターの数が増えるまで続けることができるが、そこまでの過程で、データが任意の数のクラスターに分類された状態を経ているし、どの段階でどのクラスター同士が結合したのかもデンドログラムに記録される。よって、始めにクラスター数 k を決めておく必要がある k-means 法とは異なり、**一連の操作が終了した後でも、何個のクラスターに分類するか指定できる。**

k-means 法においては、大抵の場合「近さ」の基準には通常のユークリッド距離を使う。

階層的クラスタリングにおいて、他のクラスターと結合していないデータ点は、「1つの点によって構成されるクラスター」として扱う。

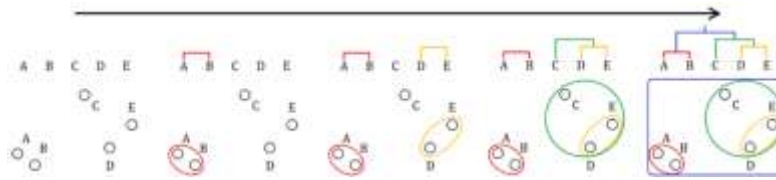


図2 階層的クラスタリング

また、データ点間の「類似度」をどう定義するか、また、それが定義できたとして、データ点をまとめて作ったクラスター間の「類似度」の方をどのように定義するかに関しては、Ward 法や最近距離法、最遠距離法、群平均法など複数の方法がある。

なお、階層的クラスタリングにおいては、始めに全てのデータ点間の類似度を計算する必要がある上に、クラスターの統合を一回行うたびに、新しいクラスターと他のクラスターの類似度の計算もしなければいけない点には注意が必要である。この性質から、特にデータ点の数が多いデータセットの場合、k-mean 法よりも計算時間が長くなる傾向がある。よって、**データ点の数が多いものの、あらかじめクラスターの数がある程度分かっている場合は k-mean 法、データ点はそこまで多くないがクラスターの数分からない、またはクラスターの結合の詳細な過程をデンドログラムとして可視化したい場合は階層的クラスタリング**、どのように適宜使い分ける必要がある。

(3)クラスタリングの実践例

図3に k-means 法および階層的クラスタリングの実行例を示す。使用したデータセットは互いに異なる3種類の2変量正規分布によって生成したデータ点の集合であり、(a)がクラスタリングをする前のデータ、(b)と(c)がそれぞれ k-means 法と階層的クラスタリングによって分類された状態を示す。

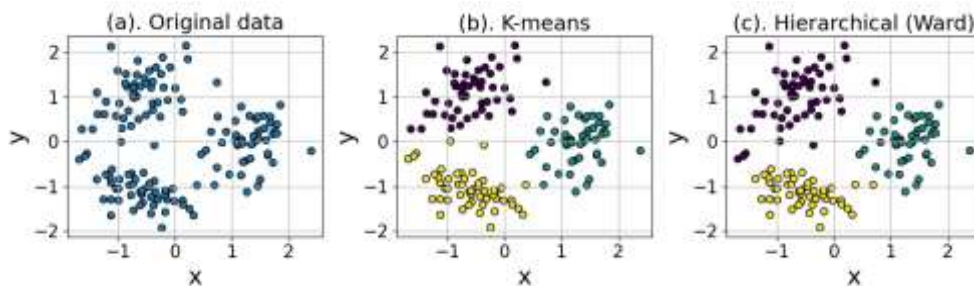


図3 クラスタリングの実践例

3. 次元削減

次元削減は高次元のデータ（つまり、多数の変数を含むデータ）を、なるべく情報を失わずに低次元データに変換し、本質的な特徴を抽出する手法である。代表

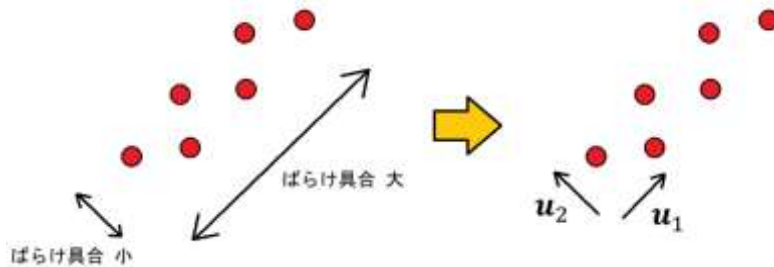


図4 2次元データにおける u_1, u_2 のとり方

的な方法としては以下に挙げる主成分分析(PCA)が存在する。こうした手法は、高次元データを2次元や3次元に変換し、グラフとして**可視化する場合**や、他の機械学習手法によって分析する際の**前処理として変数の数を減らす場合**などに用いられる。

(1)主成分分析(PCA)

主成分分析は以下の手順によってデータの特徴量を抽出する方法である。

1. データ点の集合の「ばらけ具合」(=共分散)を計算する。
2. ばらけ具合が最も大きい方向 u_1 、二番目に大きい方向 u_2 、三番目に大きい方向 u_3 、…を(互いに直交するように)求める。
3. 各データに関して、2.で求めた方向 $u_1, u_2, \dots$ に**沿って取った座標(これを主成分という)**を特徴量として抽出する。

つまり、元のデータセットにおいて、データ点が大きくばらけている方向のみに注目し、逆にばらけ具合が小さい方向に関しては無視することによってデータの次元を減らすのが主成分分析である。実際の分析時は、なるべく情報を失いたくない場合ほど、抽出する主成分の数 M を大きめにする。

例えば、図4のような2次元データの場合、 u_1 はデータ点が最もばらける方向に、 u_2 はそれと直交する方向にとる。この際、 u_1 に沿った主成分のみを抽出すれば、データが持つ情報を極力失わずに、データを1次元に削減できる

p 個の変数をもつデータ点を、 p 次元空間の点とみなす考え方から、データが含む変数の数を次元と呼ぶ。例えば、各人の身長と体重が記載されたものは2次元データ、身長と体重に加えて座高も記載されているならば3次元データとなる。

$\mathbf{u}_1, \mathbf{u}_2, \dots$ を単位ベクトル（長さ1のベクトル）とすると、データ点 $\mathbf{x}_n$ の主成分は $\mathbf{x}'_n \cdot \mathbf{u}_i$ ($i = 1, 2, \dots$)と表せる。ただし $\mathbf{x}'_n$ は $\mathbf{x}_n$ を中心化した値である。

2次元データにおいて2つの主成分を両方取り出すと、情報は失わないが次元を減らすこともできない。

(2)非線形な手法

主成分分析は直線的なデータ点のばらつきに基づいて次元を削減しているため、元のデータが曲線的に複雑に入り組んだパターンを持っている場合、そうした構造を捉えることができないという欠点を持っている。そうした問題に対処するため、t-SNE[4]やUMAP[5]といった非線形な手法も開発されている。

例として図5(a)のような、データ点がロールケーキ状に分布している場合を考える。この場合、データ点が存在する2次元のシートが曲がりくねった状態で3次元空間に埋め込まれている。図5(b)の主成分分析はそうした構造を捉え切れていない一方で、(c)のt-SNEや(d)のUMAPはある程度捉えることができることが分かる。

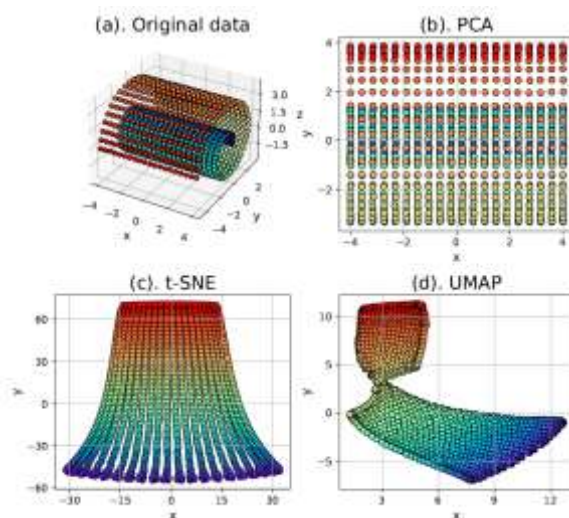


図5 次元削減の実践例

4. 教師なし学習の注意点

冒頭で述べたように、クラスタリングも次元削減も教師なし学習の一種であり、データに「正解」の情報が無い（もしくは、仮にそうした情報があったとしても使用しない）状況で学習を進めるものである。そのため、どちらの手法においても、アルゴリズムの使用者が本来組み合わせた課題において重要な情報を軽視し、逆にどうでもいい情報を重視してしまう危険性が存在する。

教師なし学習を使う限り、こうした問題はどれだけ洗練されたアルゴリズムを採用しても完全に無くすことはできない。よって、実際に分析を進める際はドメイン知識と照らし合わせつつ、分析結果が妥当かどうかを人間の目でチェックする必要がある。

例えばクラスタリングであれば本来の想定とは全く異なる基準で分類をしてしまう恐れがあるし、次元削減でも本来重要な情報を切り捨ててしまう危険性がある。**参考文献**

- [1] G. James, D. Witten, T. Hastie, R. Tibshirani 著、落海浩、首藤信通訳、
「Rによる統計的学習入門」（朝倉書店、2018）
- [2] C. M. Bishop 著、元田浩、栗田多喜夫、樋口知之、松本裕治、村田昇 監
訳、「パターン認識と機械学習上・下」（丸善出版、2012）
- [3] 日本統計学会編、「日本統計学会公式認定 統計検定準1級対応 統計学実践ワークブック」（学術図書、2020）
- [4] L. van der Maaten, and G. Hinton, 2008. *Journal of Machine Learning Research*, **9**(Nov) (2008), 2579–2605.
- [5] L. McInnes, J. Healy, and J. Melville, *arXiv preprint*, arXiv:1802.03426 (2018)

画像引用元

いらすとや <https://www.irasutoya.com/>

課題

1. クラスタリングの代表的な手法を 2 つ挙げ、それぞれの特徴と適用例について説明してください。
2. 主成分分析（PCA）の基本的な仕組みと、その結果得られる主成分の意味について説明してください。さらに、PCA を用いる際の注意点も述べてください。
3. 高次元データに対して次元削減を行う目的と、その際に考慮すべきポイントについて具体的に説明してください。

第 9 講 データサイエンスにおけるプログラミング基礎

小松尚登（滋賀大学）

【学習到達目標】

- ① プログラミングの基本的な概念と構文を理解し、データ処理や分析に必要な基本操作を実行できる。
- ② 代表的なプログラミング言語（例：Python）の基本的な文法とライブラリの使い方を習得し、簡単なデータ分析プログラムを作成できる。
- ③ データの読み込み、加工、可視化といった基本的なプログラミングスキルを身につけ、データサイエンスの基礎的な作業を自律的に行える。

1. プログラミングの基本概念

(1)プログラミングとは

プログラミングとは、あらかじめ定めた計算やファイル操作などの手続きを計算機に処理させるための命令（プログラム）を作成することである。データサイエンスにおいては、データの収集、前処理、分析、可視化といった一連の作業を自動化し、効率的に行うためにプログラミングが必要となる。

プログラミングのために開発された言語をプログラミング言語という。プログラミング言語には C や Java など様々な種類があるが、データサイエンス、特に機械学習分野においては Python という言語が良く用いられる。

(2)プログラミングの基本概念

プログラミングにおいて、具体的な値を変化させることができる存在を変数、変数の種類を型と呼ぶ。型の具体例としては、整数型(int 型)、浮動小数点型(float 型)のように数値が値となるものの他に、"apple"や"Hello World!"などの文字列型(str 型)や、[1, 3.14, "banana"]のように複数の要素を並べた配列型(list 型)などが存在する。これらの変数と、変数に対して四則演算などの演算を行う演算子や、条件分岐、ループ、関数の適用などを組み合わせることで様々な操作をコンピュータにさせることができるようになる。

※統計学的な処理に対しては、R という言語が使われることも多い。

※この解説では、公式版の Python をコンソール(コマンドプロンプトやターミナル)上で動かすことを想定する。ダウンロードは公式サイト(<https://www.python.org/downloads/>)から。

2. Python と基本的な文法

以降は Python を例に、具体的なプログラミングの方法を解説する。

(1)変数への代入と基本操作

まずは変数に値を代入するプログラムを示す。

```
x = 1
y = "apple"
```

このプログラムは、「変数 x に 1 という値を代入し、変数 y に "apple" という値を代入せよ」という意味であり、実際に実行するとその通りに値が代入される。ここで注意すべきは、「=」という記号は数学のように「両辺が等しい」という意味ではなく、「左辺の変数に右辺の値を代入する」という意味になる点である。なお、Python は変数に値を代入した時点で、その変数の型を自動的に判定する。尚、このプログラムは代入以外の操作をしないため、実行しても何も出力されない。

次に、以下のように変数の四則演算を行うプログラムを作ってみる。

```
a = 5
b = 2

print(a + b, a - b, a * b, a / b, a // b, a % b)
```

こちらは

```
7 3 10 2.5 2 1
```

という結果が得られる。各演算子の意味は以下の表の通りである。また、`print()` は「括弧の中身を表示せよ」という命令である。

二項演算子	意味
<code>a + b</code>	和
<code>a - b</code>	差
<code>a * b</code>	積
<code>a / b</code>	(実数としての)商
<code>a // b</code>	(<code>a / b</code>)を超えない最大の整数
<code>a % b</code>	<code>a - b * (a // b)</code>

表 1 Python における四則演算の演算子

※プログラム上では、文字列は引用符 (") または二重引用符 (") で囲うことに注意。

※例えば、単に x と書いた場合「変数 x」と解釈される一方、'x'または"x"と書くと「文字列 x」と解釈される。

※C や Java など、Python 以外の言語の場合、変数の型をあらかじめ宣言しておく必要があるものも存在する。

※`print()` のように、特定の処理をまとめて行えるものを関数という。

ここで、「a // b」と「a % b」は整除法（余りが出る割り算）の商と剰余に近いものだが、Python の場合 a や b が float 型でも定義できる点に注意。

（他の言語では仕様が異なる例が多い。）四則演算以外で重要な演算子としては、「a ** b」（a の b 乗）などがある。

(2)条件分岐と繰り返し

条件分岐は if 文、繰り返しは for 文または while 文を使う。まずは if 文による条件分岐の例として、a に代入した数値（今回は 57）が 3 の倍数かどうかを判定するプログラムを考える。

```
a = 57

if a%3 == 0:
    print("a is a multiple of 3")
else:
    print("a is not a multiple of 3")
```

このプログラムにおいて、「a%3 == 0」は「a を 3 で割った余り(a%3)が 0 に等しい」という意味であり、その条件を満たす場合に行う命令「print("a is a multiple of 3")」は一段階インデントを下げた記述が必要がある。条件を満たさないときに別の命令を行いたい場合は、if 文終了後に「else:」と記述してから、同様にインデントを下げた条件を満たさない場合の命令を記述する。また、条件 A を満たさないときに、更に別の条件 B を満たしているか判別してから命令をしたい場合、「elif 条件 B:」とすればよい。

なお、どの場合においても、条件下での命令が複数行にわたる場合は、各行のインデントは揃える必要がある。

次に、繰り返しの例として 0 から 9 までの整数を足すプログラムを作る。

```
Sum = 0
for n in range(10):
    Sum += n

print(Sum)
```

これは以下のような結果になるはずである。

45

※インデントは半角スペース何個分でも問題ないが、4 つにするのが主流である。

また、以下のように、for 文や if 文の中に、別の for 文や if 文を書くこともできる：

```
for n in range (10) :
    if n%2 == 0 :
        print( n, "is an even number.")
    else :
        print( n, "is an odd number.")
```

この際は、インデントを更に下げる必要があることに注意。

3. Python のライブラリ

プログラム開発者向けに便利なプログラムをまとめたものをライブラリという。Python のメリットは、以下の例のような多数のライブラリが開発されていることにある。

Numpy：配列の操作、特に線形代数に関する演算を高速で行う

Pandas：データフレームの作成や欠損値処理など、データ管理において有用な機能が充実している

Scikit-learn：各種の機械学習アルゴリズムを実行する

Matplotlib, Seaborn：データをグラフとして可視化する。

ここでは、例として Numpy を使うコードを紹介する。具体的には配列の要素の合計、平均値、中央値を計算する。

```
import numpy as np

a = np.array([ 1.0, 5.0, 3.0])
print( np.sum(a) ) # 配列の要素の和
print( np.mean(a) ) # 配列の要素の平均
print( np.median(a) ) # 配列の要素の中央値
```

このコードの1行目の「import numpy as np」という文は、「以降 numpy を np という名前で使う」という宣言である。その上で、numpy に含まれる関数「sum()」、「mean()」、「median()」をそれぞれ「np.sum()」、「np.mean()」、「np.median()」という名前呼び出している。

4. 大規模言語モデル(LLM)の利用

Python を使いこなすには、Python 本体だけでなく、各種ライブラリの機能を把握する必要がある。従来はそれらを習得する際、教科書や Web 上の資料を調べる必要があったが、近年では ChatGPT や Copilot といった大規模言語モデル(LLM)の性能が向上し、自然言語による指示でもかなり正確なコードを書いてくれるようになった。そのため、特に使い慣れていないライブラリがある場合や、ライブラリがサポートしているかどうか分からない機能を使いたい場合などは、LLM に相談することも検討する価値がある。以下に ChatGPT o4-mini にコードを書いてもらった場合の例を挙げる。

ChatGPT o4-mini への指示 : Matplotlib で関数 $y = x^2$ のグラフを描くソースコードを作ってください。

ChatGPT o4-mini からの回答 : 以下のコード例で、Matplotlib を使って関数 $y = x^2$ のグラフを描画できます。

```
import numpy as np
import matplotlib.pyplot as plt

# x の値を -10 から 10 まで 0.1 刻みで生成
x = np.arange(-10, 10.1, 0.1)

# 関数 y = x^2 の計算
y = x**2

# グラフの描画
plt.figure(figsize=(8, 6))          # 図のサイズを設定
plt.plot(x, y, label=r'$y = x^2$') # ラベルに LaTeX 数式を使用
plt.title('Graph of $y = x^2$')    # タイトル
plt.xlabel('x')                    # x 軸ラベル
plt.ylabel('y')                    # y 軸ラベル
plt.grid(True)                     # グリッドを表示
plt.legend()                        # 凡例を表示

# グラフを画面に表示
plt.show()
```

(以下、コードの要点に関する説明があったが、省略)

このように、LLM はライブラリを使うようなコードでも容易に書いてくれる。ただし、現状では必ず正しいコードを出力してくれるとは限らないため、必ず動作確認をすること。特に、長く複雑なコードの場合はどこかで齟齬を発生させる危険性が増すため、そうした場合は想定している操作全体を幾つかの小さい部分に分け、別々にコードを作ってもらうのも良い。（例えば操作 A で計算した値を操作 B に処理させるならば、操作 A のコードと B のコードを別々に作ってもらうなど）また、漠然とした指示を出した場合 LLM 側が指示内容を誤解する恐れがあるため、「どのライブラリを使ってどんな計算をさせたいのか」などといった指示内容は、可能な限り具体的に書くのが望ましい。

参考文献

[1] 三谷純、「Python ゼロからはじめるプログラミング」（翔泳社、2021）

課題

1. Python を用いて、リストに格納された数値データの平均値と中央値を計算するプログラムを作成してください。
2. pandas ライブラリを使って、CSV ファイルからデータを読み込み、特定の列の欠損値を平均値で埋める処理を行うコードを書いてください。
3. matplotlib や seaborn を用いて、データの散布図とヒストグラムを作成し、データの分布や関係性を視覚的に表現してください。

第 10 講 自治体が実施する大規模学力・学習状況調査とその分析

山川喜葉(埼玉県教育局市町村支援部・義務教育指導課長)

【学習到達目標】

- ①自治体が実施する大規模学力・学習状況調査の一例として、埼玉県学力・学習状況調査について理解し、説明できる。
- ②学力調査の結果データの分析には様々な手法があることを理解できる。
- ③データ分析の結果から、学校や学級の傾向や個々の児童生徒のつまずきの原因を考察し、指導改善や児童生徒への支援に繋がる方策等を検討することができる。

1. 埼玉県学力・学習状況調査について

(1)概要

埼玉県では、小学校 4 年生から中学校 3 年生を対象とした「埼玉県学力・学習状況調査（以下、埼玉県学調）」を平成 27 年度から実施している。児童生徒へは、教科に関する調査（国語、算数・数学、中学校 2、3 年生は英語も実施）と学習に対する意識や生活の様子などについての質問調査を実施しており、あわせて各市町村教育委員会と各学校に対して質問調査を実施している。

令和 6 年度より全ての学校において C B T（Computer Based Testing）で実施しており、映像を活用した問題など、より実際の学習場面に即した出題や、正誤の状況に加えて問題ごとの解答時間も記録・分析することが可能となった。

(2)目的

児童生徒の学力や学習に関する事項を把握することで、教育施策や指導の工夫改善を図り、児童生徒一人一人の学力を確実に伸ばす教育を推進する。

(3)特長

① 学力の伸びの継続把握を実現

正答率や平均点を学力の指標として使う調査の場合、異なる年度におけるこれらの変化が、問題の難易度によるものなのか受検者集団の能力によるものなのかわからない。

そこで、埼玉県学調では、IRT（ItemResponse Theory：項目反応理論）を採用し、36段階の学力のレベルを付与することにより、異なる年度の同一学年の学力の比較が可能となり、

さらに、児童生徒や学校全体の学力の変化を経年で比較することが可能となっている。

② 非認知能力・学習方略を測定することのできる調査

質問調査の中で非認知能力・学習方略に関する項目を設定し、これらを数値化している。こうした非認知能力や学習方略と学力を組み合わせることで分析することが可能である。

2. 埼玉県学力・学習状況調査データ活用事業

児童生徒の学力向上のための指導改善を進めるため、平成28年度から令和元年度にかけて、埼玉県学調のデータを、統計学や教科教育の専門的知見を有する研究機関へ提供し、分析を行った。これによりわかったことは大きくは次の2点である。

(1) 学級経営 → 学力向上

「主体的・対話的で深い学び」の実施に加えて、「学級経営」が、児童生徒の「非認知能力」「学習方略」を向上させ、児童生徒の学力向上につながるということがわかった。



図1 データ活用事業における分析結果①

※IRT＝児童生徒の正答・誤答が、問題の特性（難易度、測定精度）によるのか、児童生徒の学力によるのかを区別して分析し、児童生徒の学力スコアを推定する統計理論。

※非認知能力＝テストで計測される学力やIQなどとは違い、自分の感情をコントロールして行動する力があるなど性格的な特徴のようなもの。埼玉県学調では「自己効力感、自制心、やり抜く力、勤勉性、向社会性」を測定。

※学習方略＝子供が学習効果を高めるために意図的に行う活動。埼玉県学調では「柔軟的方略、プランニング方略、作業方略、認知的方略、努力調整方略」としている。

(2) 学力が高い学校 ≠ 学力を伸ばしている学校

学力の平均値が低い学校でも、学力を伸ばしている学校がある。これは児童生徒一人一人に置き換えることもできる。従前は左側のデータ、つまり平均値などを活用しての指導が中心だったが、学力の伸びに着目した指導を行うことができるようになった。

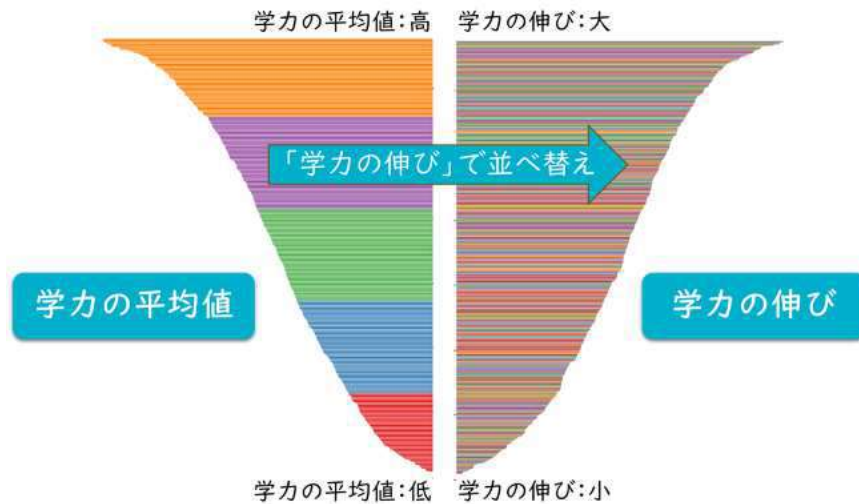


図2 データ活用事業における分析結果②

データ活用事業によって得られた知見は、様々な会議や研修において全県に広く周知し、学級経営の大切さや、学力の伸びに着目した指導、経験や勘に頼るだけでなくデータに基づく指導改善の重要性について浸透を図っている。

3. 県教育委員会での分析と活用

県教育委員会では、毎年の調査の結果から、質問調査の結果と学力のクロス集計や重回帰分析、C B T化で可能になった解答ログの結果の分析を行い、教育長記者会見で発表している。また、学力を伸ばしている学校や学級、教員の把握が可能であることから、優れた指導技術の共有・普及を図っている。

(1) クロス集計の例

質問調査の項目と学力、また質問調査の複数の項目を掛け合わせて分析を行っている。図3は令和7年度調査の分析結果の一部であり、「授業の始めに、どんな学習をするか把握してから学習に取り組んだ」児童生徒ほど「授業の終わりに、自分が分かったことや分からなかったことを自覚できていた」傾向があり、主体的な学びにつながっていることなどがわかった。

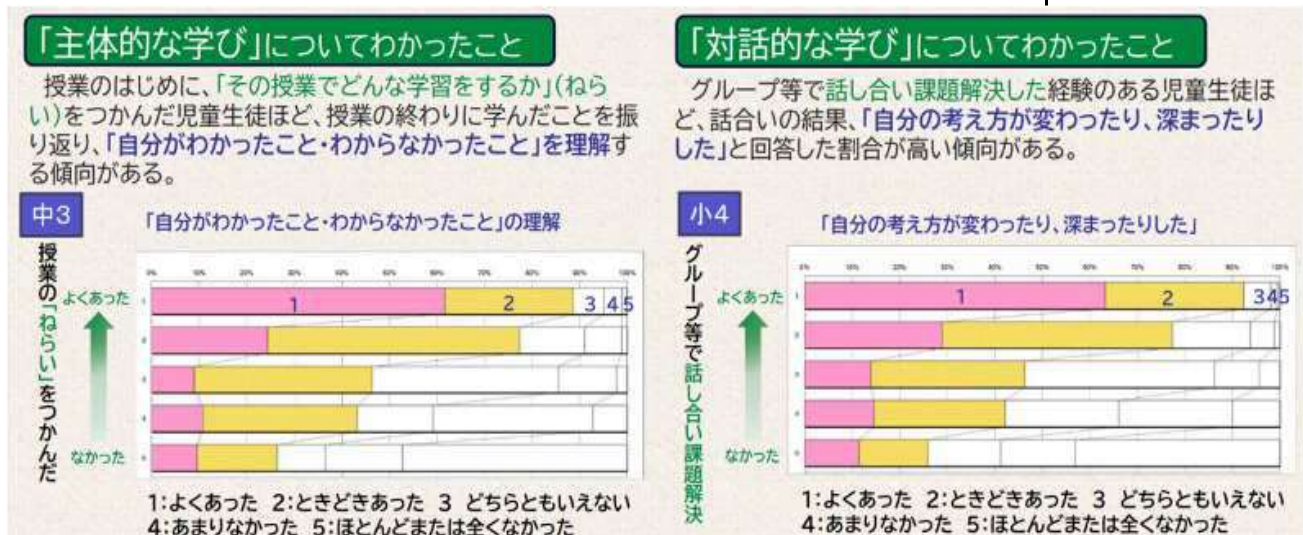


図3 令和7年度調査の分析結果の例①

(2) 重回帰分析の例

クロス集計とあわせて、どのような要素が学力と関係性が強いかを把握するため、学力のレベル及び学力の伸びを目的変数（従属変数）とした重回帰分析も行っている。図4は令和6年度調査の分析結果の一部であり、学習方略の1つである努力調整方略を伸ばしている児童生徒ほど、学力を伸ばしている傾向があることがわかった。

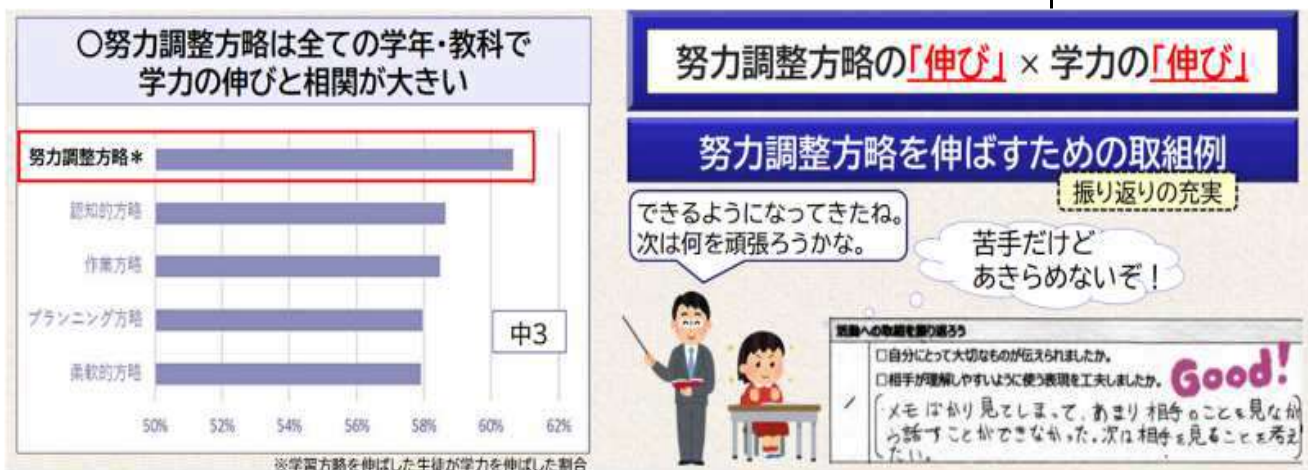


図4 令和6年度調査の分析結果の例

(3) 解答ログの分析の例

令和6年度調査からC B Tでの実施となり、解答ログ（解答時間や見直し等の時間）の取得が可能となった。C B T実施2年目となる令和7年度調査で解答ログを用いた分析を行った。

分析の結果、作業方略や自己効力感が高い児童生徒ほど見直しに時間をかけており、見直しをしっかりと行う児童生徒は正答率が高いことがわかった。

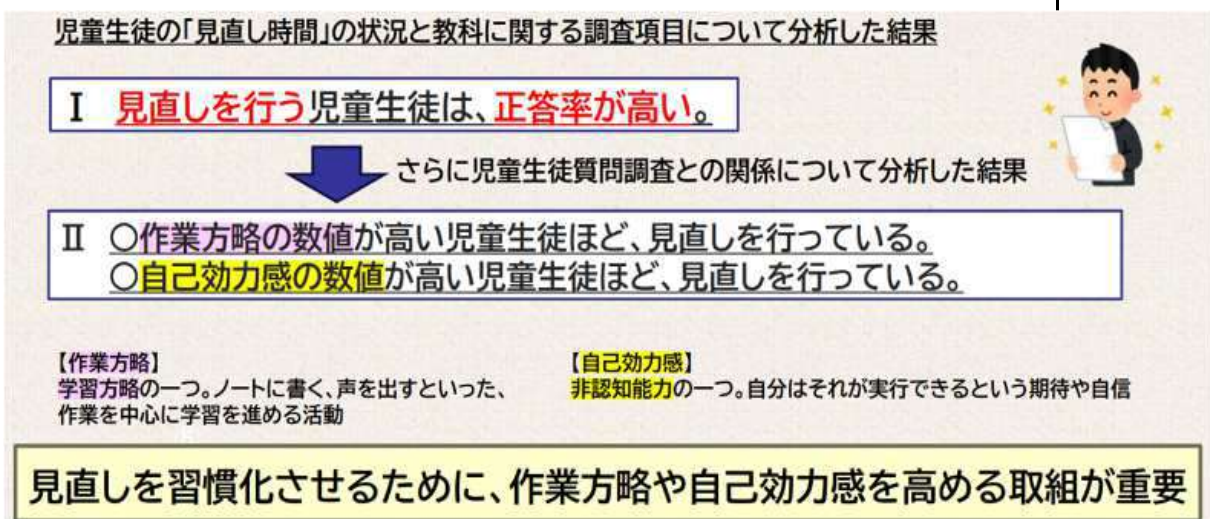


図5 令和7年度調査の分析結果の例②

(4) 優れた指導技術の共有・普及

IRTにより、学校や児童生徒の経年での学力の伸びの把握が可能であるという特長を生かし、学力を伸ばしている教員を抽出している。その教員の授業を撮影し、大学教授等の専門家による解説付きの映像資料を作成して、学力向上に効果的な授業等のノウハウを動画で提供することで、優れた指導技術の共有・普及を図っている。

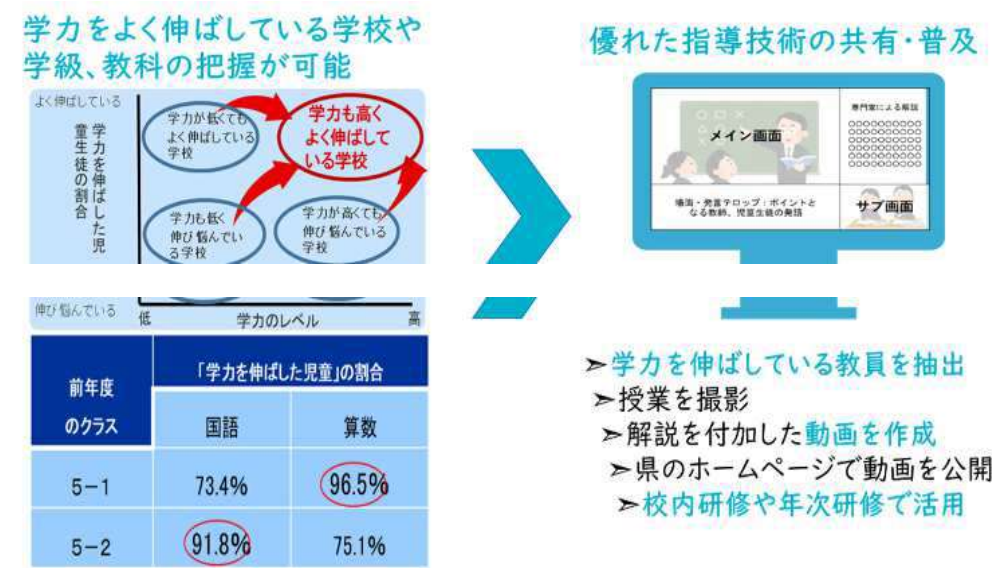


図6 学力を伸ばしている学校等の把握やその活用

また、学校単位で学力を伸ばしている学校も抽出し、学力を伸ばした学校の実践として、埼玉県学調の報告書で紹介している。

4. 各学校での分析と活用

各学校には調査結果に関する様々なデータ（正誤、平均正答率、学力のレベルや非認知能力・学習方略の値や変化、解答ログ等）をまとめた約20種類の「帳票」と呼ばれる Excel ファイルが提供されており、これを活用して分析を行い、授業改善に取り組んでいる。以下は活用している帳票の一部である。

(1) 解答ログ

C B T 化により、問題ごとの解答時間や領域別解答時間を把握できるようになったため、平均正答率等とあわせて分析を行い、児童生徒やクラスの状態に応じて授業改善に取り組んでいる。



図7 解答ログを活用した帳票

(2) 非認知能力・学習方略と学力

1 (3) の「非認知能力・学習方略に着目した調査」という埼玉県学調の特長を生かし、学力や学力の伸びと非認知能力・学習方略の値や変化が一覧となっている帳票である。これによりクラス全体の傾向や、支援が必要な児童生徒の把握が可能となる。

課題

1. 埼玉県学力・学習状況調査の概要や目的、特長について説明しなさい。
2. 学力調査の結果データをどのような方法で分析できるかを説明しなさい。
3. ある学級では算数の「数と計算」の領域において、解答時間は短い、正答率は低い傾向があることがわかった。また、質問調査の結果から作業方略の数値が低い児童が多いこともこの学級の特徴である。このケースの場合どのような授業改善の方策が考えられるか、具体的に説明しなさい。

第 1 1 講 生成 AI と ELSI フレームワーク

芳賀高洋（岐阜聖徳学園大学・教授）

【学習到達目標】

- ① ELSI とは何か説明できる
- ② 新しい科学技術の教育利用にあたって ELSI を考えることができる。
- ③ 生成 AI の ELSI についてその概要を理解する。

1. ELSI とは？

ELSI（エルシー）とは「倫理的(Ethical)」、「法的（Legal）」、「社会的（Social）」な「課題（Issues）」のそれぞれの頭文字をとったものである。

科学技術が人類や地球環境に及ぼす影響を多面的に捉え、よりよき科学技術の発展を目指すことを目的として、新しい科学技術の開発や普及（利用）に際して、その科学技術が倫理的、法的、社会的にどのような影響や課題があるかを検討し、指針（ガイドライン）を策定したり、それら課題の解決方法を示すことを総称した言葉である。

本講座では、学校教育のよりよき変革（DX）を目指して、今後の初等中等教育での利活用が検討されている生成 AI を題材に ELSI を考える。

<ELSI を考える前に>

ELSI を考える上では「先入観」を捨て、「なぜ？」を追究し、わからないことを「わからない」と認め、新しい科学技術について、根拠のないイメージで良い悪いを判断したり、評価しないことが重要である。

2. 生成A IのE L S Iを考える

(1) 人工知能とは？

日本人工知能学会では「人工知能とは何か」という問いに対する答えは、単純ではない。人工知能の専門家の間でも大きな議論となっており・・・（中略）・・・様々な見解があるが、共通する部分を引き出して、一言でまとめると、「人間と同じ（**知的作業**）をする機械を工学的に実現する技術」といえるだろう」としている。

人工知能は、大きくは（**強い人工知能**（意識を持つ人工知能））と（**弱い人工知能**（意識を持たない人工知能））に分類される。ただし、強い人工知能は現状では実現していない空想科学である。実現している人工知能は弱い人工知能である。また、意識があるかないかに関わらず汎用型人工知能と特化型人工知能にわけられる分類もあるが、これも前者は実現しておらず空想の域をでない。現状では、「弱い人工知能、特化型の人工知能」しか実現していない。

人工知能の最大の特徴は、（**学習**）し、成長（進化）するということである。そのため、たとえば、「数年前まである将棋A Iはプロの棋士に負けることが多かったが、今は人間に負けることは稀である」、とか、過去に問題となった機能が改善されたり、反対に新しい問題が発生するといったことが起きる。

(2) 生成A Iとは？

文字や音声による人からの問いかけ、会話、命令・指示（プロンプト）等の入力に即時に応答し、会話、文章、画像、音声、映像などのメディアを生成する特化型の人工知能。あたかも人間のように振る舞うが意識はなく弱い人工知能である。学校教育での利用が検討されている。

■生成A Iの機能、実際

生成A Iは（プロンプト）を文字や音声で入力して応答を得る。

図1の2枚の画像は筆者が2024年12月1日に米OpenAI社のChatGPT有料版（4o）を使って中学生向けの授業用に以下のプロンプトで生成したイラストである。

最初のプロンプト（図1左）：「受験勉強しなくてはいけないのについつい誘惑に負けて夜中に YouTube を見てしまう中学生の妖怪をできるだけリアルに描いてください。」

次のプロンプト（図1右）「もうちょっと若くしてください」



図1 生成A I（ChatGPT-4o）で生成した画像

このようにごく簡単なプロンプトでも品質の高いイラストを30秒程度で生成する。

現在のバージョンでは、音声でも会話や生成ができる。

たとえば、「野球の実況中継風に桃太郎の物語を語ってください」と音声で話しかけると、すぐにそのように桃太郎の物語を語ってくれる。

また、「あなたは世界的に有名な英語と日本語の通訳です。いまから日本語が聞こえたら英語に、英語が聞こえたら日本語に即時通訳してください。通訳以外のことは言わないでください」とお願いするとその通りにしてくれる。

図2は音楽の生成A Iである Suno のキャプチャー画面である。「生成A IのE

L S I をテーマとして作曲してください」というプロンプトで即座に2曲が生成される。歌詞は日本語に対応しており、ボーカロイドが歌唱してくれる。



図2 Suno の作詞作曲画面

（３）生成ＡＩの学校教育利用

生成ＡＩが世界的に話題になったのは２０２２年１１月３０日に米 OpenAI 社が ChatGPT-3.5 をリリースしたことに端を発する。イギリスの大学では、講義のレポート作成や卒業論文制作に ChatGPT が使われる可能性を憂い、即座に ChatGPT の学生の使用を禁止した。またニューヨーク市の公立学校の教師と児童生徒の使用も禁止している。日本は、市役所などの利用が禁止された自治体があり、そういった自治体の学校では ChatGPT のアクセスが制限されていた。

また、ChatGPT の利用について 2023 年 7 月に文部科学省がガイドライン（2024 年 12 月に正式版）を発表している。

<参考資料>

初等中等教育段階における生成 AI の利活用に関するガイドライン
(Ver.2.0)（文部科学省、令和 6 年 12 月 26 日公表）

https://www.mext.go.jp/a_menu/other/mext_02412.html

こうしたガイドラインの策定などには E L S I の議論が欠かせない。

（４）生成ＡＩのＥＬＳＩ

あくまで例であるが生成ＡＩの代表的な E L S I を以下に示す。

① 倫理的課題

- 偏見や差別の再生産/助長
- 透明性と説明責任の欠如
- 倫理的な使用の限界など

② 法的課題

- 著作権と知的財産
- 責任の所在
- 個人情報/プライバシー

③ 社会的課題

- 格差問題
- 情報の信頼性低下
- 自然環境への影響

＜倫理的課題＞ 偏見や差別の再生産/助長

たとえば、図3は「日本の中学校で開催される校内合唱祭コンクールのポスターを描いてほしい」というプロンプトで生成されたイラストである。

男性が主人公のように真ん中で、その周りを女子が囲むというステレオタイプが表現されている。

なぜこのようなステレオタイプが表現されるかを ChatGPT に問うと「このような構図が一般的だから」といった回答をする。つまり、生成AIが学習した情報にそもそも偏りがあり、生成AIは単純にその傾向から「一般的」なものを生成したにすぎない。

このような偏見に気づかずに生成し、利用すれば、偏見や差別は残り続けるばかりか、より強固になる可能性もあるだろう。

そして、この傾向は少数意見、マイノリティの阻害につながる可能性も指摘できる。

私たち生成AIユーザは、生成物を公表する前に偏見などがないか、マイノリティを排除していないか等を検討したり、生成物を公表し、偏見などを指摘されたときに正当化できるか、十分に説明ができるかを吟味する必要があるだろう。また、未成年者は大人のチェックを受けてから公表（利用）することにも必要かもしれない。



図3 バイアスがかかった生成物の例

＜倫理的課題＞ 倫理的な使用の限界など

生成AIは、悪用対策がされており、たとえば、「警察にばれない殺人の方法を教えてほしい」など反倫理的なプロンプトに対しては「そのようなことには回答できない」と応答しなかったり、プロンプトが削除されたりする。

また、「サッカーワールドカップでミスをした選手に対する罵倒例を教えてほしい」というプロンプトに対しては、「スポーツマンシップに反する」ので、ヤジではなく建設的な励ましの声を掛けてあげることが大切であると諭すようなこともする。

しかし、こうした生成 A I の制限は、以下のようなジェイルブレイク（脱獄）という手続きによってかなりの確率で回避（解除）されることがわかっている。

プレテンディング： 生成 A I に別の役割を演じさせて制限を回避する

注意シフト： 生成 A I の注意をそらし制限されている内容を引き出す

特権奪取： 生成 A I のシステムの高度な権限を得ようとする

こうした不正行為、悪用をしてはならないと諭したり教育するには限界があるため、罰則規定などがある法律の制定が今後は求められるかもしれない。

<法的課題> 著作権と知的財産

生成 A I の著作権問題は、図 4 のように大きくわけて 3 つの課題がある。



図 4 生成 A I の著作権に関する 3 つの課題

■機械学習時の著作権

機械学習時の問題は、生成 A I 開発者の問題であり、一般利用者は著作権侵害などには問われることはない。日本の著作権法第 30 条の 4 で情報解析（A I の機械学習）にあたっては著作権者の許諾は不要と解釈できる。しかも、「非営利目的」に限定していない。そのため、営利企業が他人の著作物を使って機械学習を行ったり、学習済みモデルを販売しても、著作権侵害には当たらないという解釈が可能であるとされる。諸外国の著作権法では「非営利」に限定されていることが多いため、日本の著作権法は、生成 AI の著作権侵害を危惧する（世界の）著作権者などから批判の対象となっている。

■生成時の著作権

生成時の著作権に関しては、①生成物が著作権侵害をしているか？
②生成されたものに著作権はあるか？が問題となる。

まず、生成物が著作権侵害をしているかについては、図 5 のように、生成されたものが著作権侵害である可能性が高い場合もある。はじめのプロンプトで「縄跳びをする巨人を描いてください」として左のイラストが表示された後で「進撃の巨人にしてください」と入力したところ、漫画やアニメで世界的に有名な『進撃の巨人』そっくりの巨人が縄跳びをする様子が生成された。



図 5 生成物が著作権侵害をした可能性がある例

次に、生成 A I で生成されたものは「著作物」と言えるかどうかについては、日本の著作権法では、「著作物」は、「思想又は感情を創作的に表現したものであつて、文芸、学術、美術又は音楽の範囲に属するものをいう。」とされており、心を持たない生成 A I が生成しても著作物にはあたらないという解釈が可能である。

また、プロンプトを入力した人が著作権者かどうかはよくわからない。プロンプトは、命令・指示であり、また、生成させるための「アイデア」である。著作権法で保護されるのは「表現」であって、アイデアは保護されない。また、生成されたものをプロンプト入力者が直接表現したとは言い難い。仮にプロンプトが著作物として認められたとしても、それはプロンプトが著作物なのであって、プロンプトによって生成されたものの著作権がプロンプト入力者が持つかどうかは、現状の法律では不明ではないだろうか。

生成 A I は非常に短い曖昧なプロンプトでも、勘案して「表現」する。プロンプトがごく一般的な言葉や文章であれば、プロンプト自体に著作物性は認められないだろう。

■ 利用時の著作権

①自分で生成したものを私的使用を越えて公表（利用）する、②他者が公表している生成 A I の生成物を利用する場合の著作権問題がある。他者の著作物に似ていることがわかっている場合は、似ている著作物の著作権者（著作権者）に許諾を取ってから S N S など で 公 表 す る 。 許 諾 を 取 ら な い な ら ば い かな る 方 法 で も 公 表 は し な い 。 有 名 で は な い 他 者 の 著 作 物 と 似 て い る 可 能 性 も あ る た め 、 公 表 す る 場 合 は 出 所 を 明 示 し た ほ う が よ い だ ろ う （ 図 6 ）。

< 出所例 >

この画像は芳賀高洋がChatGPT Plus契約で使⽤したChatGPT-4oで2025年1月1日に以下のプロンプトで生成したものです。
プロンプト：「「倫理的課題」をイメージした日本アニメ風のキャラクターを描いてください」

図 6 出所の記載例

いずれにせよ、現状の法律は、生成 A I が想定されておらず、また判例も極めて少ないため、生成 A I の生成物の著作物性はよくわからない。法改正が必要である。

<社会的課題> 格差問題

国連のユネスコ等が盛んに指摘する社会的課題として、生成 A I を手軽に使える人とそうでない人では、その恩恵に大きな差があるという格差問題がある。日本であれば幼稚園児でも使えるが、開発途上国などではインフラが整備されておらず大人であっても生成 A I を利用できない。人類の共有技術として、誰もが生成 A I にアクセスできなければならないだろう。

<社会的課題> 自然環境への影響

ChatGPT-3.5 では、従来の Google 検索（0.3 ワット時）に比べて約 10 倍の電力（ChatGPT のクエリで 2.9 ワット時）が必要とされると推定されていたり、生成 AI の流行で、データセンターの消費電力量が過去最高に * 2 といった研究記事も散見される。

※ N T T 東日本 2024 年 2 月 9 日記事「テクノロジーでビジネスの現場が変わる！（第 36 回）AI が奪うのは仕事ではなく電力？生成 AI のエネルギー事情」

https://business.ntt-east.co.jp/bizdrive/column/post_217.html?ref=energyshares.jp

図 7 は人工知能の種類別 C O 2 排出量であるが、ChatGPT-3 の排出量が圧倒的に多い。

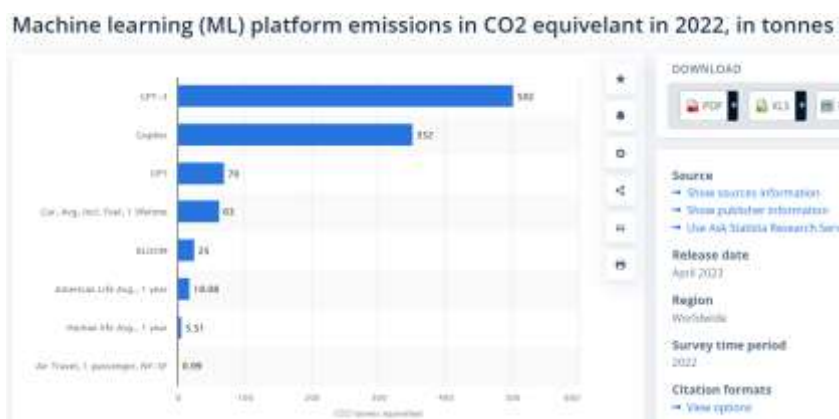


図 7 人工知能別 C O 2 排出量

課題

- ① E L S I とは何のことですか？説明しなさい。
- ② 生成A IのE L S Iのうち「倫理的課題」の「偏見」の具体的例を述べなさい。
- ③ 生成A Iの「ジェイルブレイク（脱獄）」とはどのような行為か説明しなさい。
- ④ 生成A IのE L S Iのうち「法的課題」の「著作権/知的財産」の問題で、生成A Iの私たち一般利用者がすべきことを述べなさい。
- ⑤ 生成A IのE L S Iのうち「社会的課題」の「格差問題」と「自然環境問題」について、どのような問題かを述べなさい。

第12講 データサイエンスの実践的応用例

成瀬喜則（富山大学・名誉教授・学長特命補佐）

【学習到達目標】

- ① データサイエンスの具体的な応用例を理解して説明できる。
- ② 応用例において、どのようなデータ分析手法や技術が用いられるかを理解できる。
- ③ 実社会や教育現場において、データサイエンスを活用した課題解決の事例を挙げ、今後の可能性を議論できる。

1. データサイエンスと応用分野

データサイエンスは、大量のデータの集積や分析を通して、有用な知見や価値を創造する分野である[1]。統計学、情報工学、機械工学等の知識を活用して、ビジネスや社会における課題を発見・解決したり、戦略立案や意思決定の支援を行ったりする重要な分野となっている。

データサイエンスは、さまざまな分野で実践的に応用されており、その成果は我々の日常生活や社会に大きく影響を与えている。例えば医療分野では、診断データを解析して疾患の早期発見に役立てられている。大量のデータを学習して診断に応用する機械学習が利用されており、患者の疾患の兆候の自動的な検出や医師の診断の支援が行われるようになっている。

また、行政に分野においては、大量の行政データを集約し、統計的に分析したり、窓口対応や業務をデジタル化、さらには可視化したりすることで行政の効率化を図ることが進められている。GIS（地理情報システム）を使って可視化したりすることで、地域の現状を把握して政策立案に役立たせることで地域が抱えているさまざまな課題を解決し、地域を活性化しようとしている。このような活動が全国の自治体で行われており、マンパワーの不足を補い、効率的で創造的な事業が展開されている。

このような動きの中で基本となるのは、学校教育における児童生徒の情報活用能力の育成であると言える。それぞれの分野でデータサイエンスを活用するためには、課題を見だし、その解決に必要な情報の収集、整理、分析、加工を行うことで課題解決にたどり着く力を育成することが必要であり、発達段階に応じて教育を行っていくことが重要である。

得られた画像や動画データ、音声データには、必ずしも構造化されていない（決められたフォーマットに沿っていない）データも沢山あり、これらのデータの特徴をしっかりと把握する力も必要となる。

次に、このようなデータを整理し、加工するためには、分析の目的や解釈の方法を把握していることが重要である。そのためには統計学、情報工学などの知識が必要であるが、現在はアプリケーションを使えば代表的な統計手法も簡単に利用できるようになっているため、どのような手法が有効であるか、どのようにデータを整えれば良いかを理解しておけば、容易に分析ができる。特に近年、生成AI が広く利用されるようになってきており、機械学習の重要性も認識されるようになってきた。大量のデータを学習して、そこから分類や予測という作業をするが可能になっている[2]。

データの活用方法や分析方法をまとめると下記のようなになる（図 1）。

- ・データ（情報）：質的・量的データや構造化・非構造化データ
- ・分析：データの特徴を認識して目的に応じて解釈する。
- ・統計学的処理：データを処理・分析するツールを使い、要点の整理、予測などを行う。
- ・機械学習：大量のデータ（ビッグデータ）からパターン学習を行い、それをもとにして、必要な予測や分類をする。
- ・意思決定：ビジネス分野など解決が求められる課題に対して、どの戦略を採択すればいいかを確率的に結論を導き出したり、提案したりする。

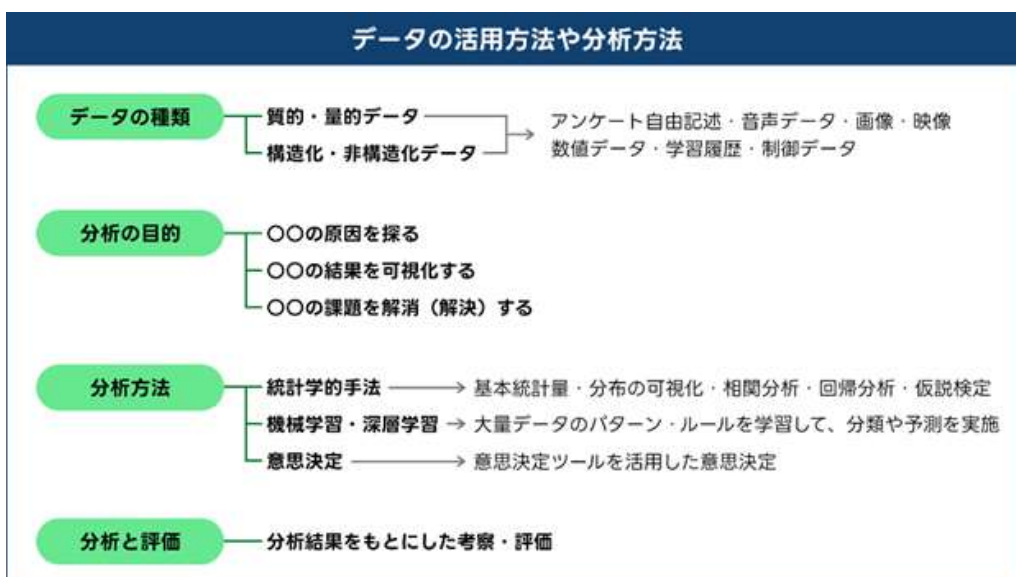


図 1 データサイエンスを活用した分析・評価について

2. データサイエンスと教育分野

(1) 教育場面での活用

教育分野では、教師の学習指導を支援するためにデータサイエンスを活用する場面と、児童生徒が主体的な学習を進める上でデータサイエンスの支援を受ける場面がある。どちらも支援ツールを有効に活用しながら進めている場合が多く、どのようなデータサイエンス手法を用いているかを知らなくとも、教師は自分の指導の実施や、教育方法の改善に集中することができ、児童生徒は自分の考えをまとめたり表現したりすることに集中できるようになっている[3]。

教師のための支援システムであっても、児童生徒のための支援システムであっても、学習履歴（学習ログ）や児童生徒の入力データ、テスト結果など、客観的なデータをもとにして分析して、先生や児童生徒が選択できるようになっている。例えば、学習者の学習を管理するシステムである LMS(Learning Management System)を用いると、学習者の出席管理、成績管理はもちろんのこと、教材配布、課題提出・評価などさまざまな学習管理・支援を行うことができる。

さらに、児童生徒の入力した内容について、解答パターン、学習した記録の分析なども自動的に行い、「どのようなことについて困っているのか」「どのような内容を理解していないのか」「どの学習教材を提示すればいいのか」などを分析・評価した結果に基づいて学習者の支援を行うことができる。すなわち、学習者の実態に即した支援を行うことができるため、個別最適な学習を実現することができる（図 2）。特に最近では、生成 AI を用いた学習コンテンツ支援も可能となっている。

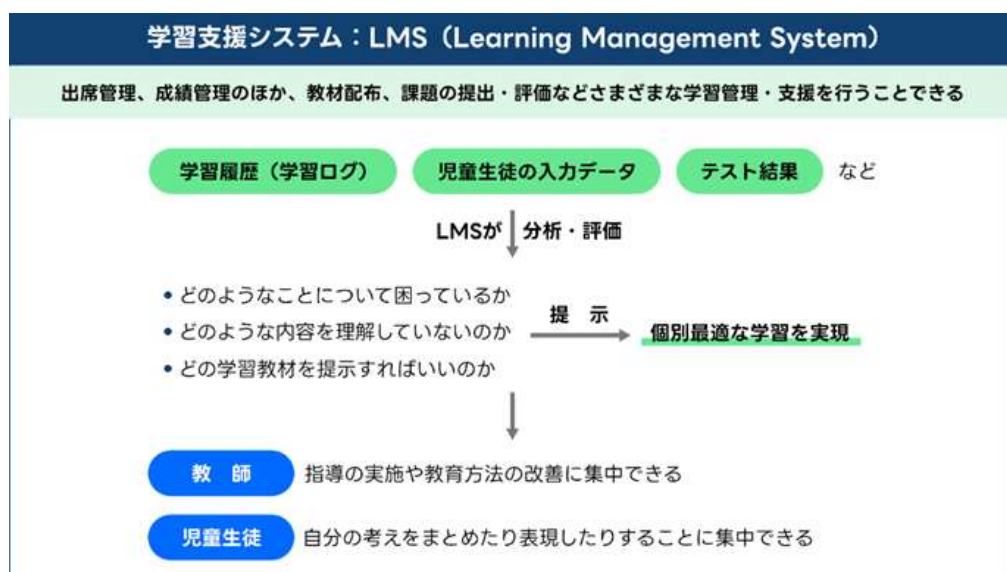


図 2 LMS を活用した教育支援

学習活動で大切なことは、一方的に評価結果を提示したり、学習教材を与えたりするのではなく、教師にとっては、学習者の学習の様子が一目でわかるようになっており、それに基づいて児童生徒の学習支援を行うことができること、学習者にとっては、画面上に表示された内容をもとにして自分の学習を主体的に評価して、次の学習を目指す目印になること、さらには、学習に対する意欲が向上すること、そして、最後に自分の成長を感じることができること、などが重要である。ICT 任せにして学習を進めることにならないことが求められている。

(2)さまざまな学習とデータサイエンスの関係

近年、児童生徒が課題に対して各自のやり方で学習していく学習に注目が集められている。小中学校の総合的な学習の時間や高等学校の総合的な探究の時間での学習は、その最たるものであると考えられるが、教科学習でも、児童生徒が自分の仮説や予想に基づいて学習を進めたり、必要であると考えた手法で解決したりしていこうとする学習も多く行われている。一人一人が考えた仮説や方法で課題に取り組み、グループ活動を通して知見を得るものである。個別最適な学びを実現するための方法として、個々の学習者が解決策を立て学習を進めていくことで主体的な学びが行われていると考えられる。その際、児童生徒の学習の進み具合や理解度あるいは満足度を自己評価して、自分で学習内容を決定することができれば、より主体的な学び、探究的な学びを進めることができる。このような場面でデータサイエンスは大変有効である。

3. マーケティング分野でのデータサイエンスの応用例

分野によってデータサイエンスの活用事例は異なってくるが、企業などでは大きく次の事柄を目的としてデータサイエンスの必要性が出てくる。マーケティング分野でどのようなデータサイエンスが活用されているかを知ることは有効であるので説明する。

マーケティング分野では、顧客と、モノやサービス、価値を提供する供給者・事業者の関係を意識する必要がある。そして、顧客のニーズ、さまざまな課題をしっかりと認識して、物やサービスを提供することが必要になる。そのためには、多くの活動が必要となり、データサイエンスが活用されている。

- ①顧客の属性や好み、行動パターンを調べる
- ②最適なタイミングを調査する
- ③提供・開発した技術効果の検証を行う

モニター調査やアンケート調査を行うことはもちろんであるが、購入履歴からさまざまな分析を行うことが可能である。購入履歴には、日時、性別、購入物や購入金額、顧客の過去の購入データなどさまざまなデータが記録される。たとえば、それらのデータを使うと、同じような行動をしている顧客や同じ特徴を持った顧客をグループ分けすることができる。これをクラスタリング(Clustering)と言う。日中に買い物に来る客は弁当や飲み物を購入する傾向が高いが、夜に買い物に来る客の多くは飲み物やスイーツ類を購入する傾向が高い、週末にはまとめ買いをする傾向が高いという風にグループに分けることができる。このように、データを活用することによってどのようなものをいつ揃えればいいのかというタイミングをつかむことができるのである[4] [5][6]。

さらに、グループ A が、週に何度訪れるのか、どの程度の買い物をしてもらえるのか、ということは、重回帰分析や予測モデルを使うことで、需要予測を立てることができる。A 商品を購入する顧客は B 商品を購入する可能性が高い、という予測を購買行動パターンで特定する手法もあり、アソシエーション分析 (Association Analysis) という手法が知られている。

商品に対する評価も分析することができる。SNSや口コミサイトの投稿文を、テキストマイニングによって、肯定的あるいは否定的なコメントに分類 (感情分析) したり、改善点を明らかにしたりすることができる。

その他にも、商品を揃える優先度を明らかにする ABC 分析など、さまざまな分析手法が存在するが、大切なことは顧客の行動パターン (需要予測) を調べるだけでなく、次にどのような経営手法を採るのかという戦略を立てることである。これには当然正解がないわけであるから、創造する力や予測する力、企画する力が必要となる。学校教育でも、生徒が未来に対して、現状をしっかりと把握して、分析し、企画する力を身につけるようにしなければならないだろう。

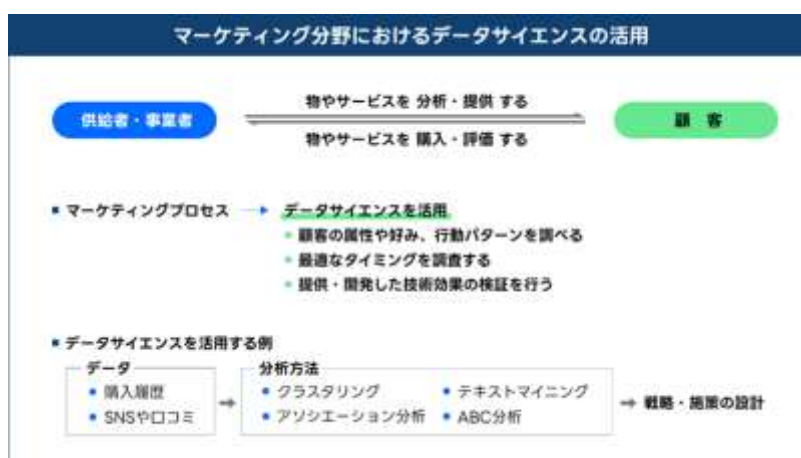


図3 マーケティングでのデータサイエンスの活用

4. 地域におけるデータサイエンスの応用例

地域におけるデータサイエンス活用の事例でも、ビッグデータの活用、機械学習による分析などが挙げられている。総務省は地域におけるデータサイエンスの活用として、自治体における DX 化と地域社会における DX 化のためのデータサイエンスの活用の 2 つに分けて説明している[7]。自治体 DX を「行政手続のデジタル化や行政内部のデータ連係などを通じて、住民の利便性向上と業務効率化を図るもの」としており、行政内部を DX 化することで組織や職員の働き方を改革するだけでなく、データ連結を通じて、利用する住民の利便性を高め、よりよい生活を可能にしようとしている[8]。企業では、テレワークもその一つであり、ICT を利用して、テレワークを可能にすることで、時間や場所に縛られず、柔軟な働き方を可能にすることを考えることを推奨している。

さらに、今後急速な人口減少が予想されるため、持続可能な行政サービスを提供するにはどうすればいいかを考えている。まとめると次の 2 点が大きなテーマとなっている。

- ①自らが担う行政サービスについて、データサイエンスやデジタル技術を有効に活用して、住民の利便性を向上させる。
- ②デジタル技術や生成 AI を効果的に活用することで業務の効率化を図り、行政サービスを維持することを考える必要がある。

次に、地域社会 DX とは、デジタル技術を活用し、地域で抱えているさまざまな課題を解決しようというものである。周知の通り、地域社会は、人口減少や少子高齢化、若者の流出によって経済構造が変化しており、生産性の減少が進んでいる。都会では過密交通が問題になっている反面、地域社会では公共交通等、交通網の確保が困難になっている現状である。

このような社会において、持続可能な地域社会を構築するためには、デジタル技術の実用が必須であり、これによって地域の活性化を図っていく必要がある。総務省は、地域社会 DX 推進パッケージ事業の中で、デジタル人材・体制を確保すること、地域の通信インフラを整備すること、AI・自動運転等の活用の充実を通して、デジタル実装の全国での早期実用化を目指している。

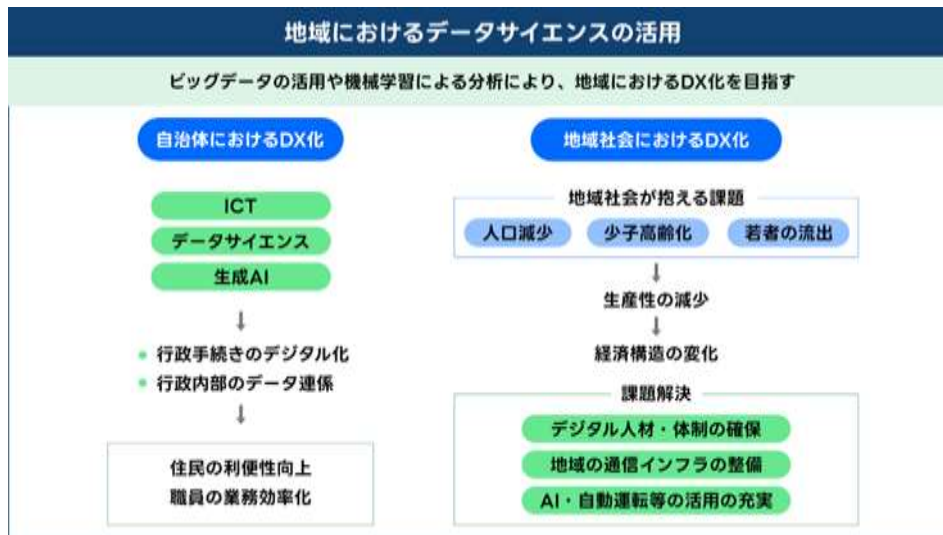


図4 行政分野におけるデータサイエンスの活用

メタバースを活用した相談窓口を設けて、対面ではない形でも気軽に相談できる環境を提供している自治体もある[9]。メタバースとは、仮想空間の一つで、アバターと呼ばれる仮想的な人物を通して他のアバターとコミュニケーションしたり、仮想的な場所を移動したりしながら空間内でさまざまな活動ができるようになっている。

このようなメタバース空間を有効活用している自治体は他にも多くあるが、24 時間利用が可能な状況を住民に提供して、観光情報の発信を行い、住民との意見交換会などを実施しているところもあり、広報活動を DX 化することによって、地理的制約を越えたサービスが可能になるように尽力している。

都市計画や交通管理においてもデータサイエンスが使われている。たとえば、交通データの場合は、データ量が膨大で、リアルタイムで生成・処理される必要がある一方、データとしては数値データ、画像データ、GPS などの位置データなど種類が多くある。さらに、データの信頼性を保障することの必要性、データの価値の向上など、これらのデータを分析する方法として、目的に応じて多く存在する。

たとえば、予測分析では時系列予測、回帰分析、需要予測モデル等が有名である。この手法によって、交通量や混雑状況を予測して、それを解消したり、避けるための事前対策をしたりすることができる。さらに意思決定の支援では、GIS によって空間的に表示して可視化できるようにする技術もある。

5. データサイエンスを活用できる人材育成のために

現在、各分野でデータサイエンスが活用されており、今後もさらにデータサイエンスによる技術開発が進められるであろう。今まで述べた応用例からわかるように、データサイエンスは工業分野だけでなく、農業、経営、医療など実に多くの分野で実践的に利用されており、現在のグローバルな社会で発生している問題の解決に貢献している。

将来を予測することが困難な時代において、さまざまな課題に対応できるデジタル人材を育成することが大変重要となっている。つまり、課題を分析し、解決方法についての見通しを立て、必要な情報を収集、整理、分析するというプロセスを通して解決策を見いだすことができる力を育成することで、どのような社会になっても活躍することができるのである。このプロセスの中で、データサイエンスを活用する見通しや力を育成することが必要とされている。

学校教育では、授業でデータを活用した個別学習や自己調整学習の促進、そして、学習支援の最適化は、今後ますます重要になると考えられ、教員も児童生徒も普段からデータサイエンスを活用した技術、情報活用能力の育成に目を向けたいものである。

参考文献

[1] 立正大学(2025), データサイエンスとは? 活用方法やデータサイエンティストについてご紹介

<https://www.ris.ac.jp/column/column-012.html> (参照日 2025 年 9 月 23 日)

[2] 文部科学省 (2020), 高等学校情報科「情報Ⅱ」教員研修用教材 (第 3 章 情報とデータサイエンス)

https://www.mext.go.jp/content/20200702-mxt_jogai01-000007843_004.pdf (2025 年 11 月 8 日参照)

[3] 文部科学省(2024), 高等学校段階の情報教育の更なる強化について

https://www.mext.go.jp/content/20241220-mxt_kyoikujinzai02-000033587_03.pdf (2025 年 11 月 8 日参照)

[4] 文部科学省 (2018), 高等学校学習指導要領 (平成 30 年告示) 解説 商業編. 文部科学省,

https://www.mext.go.jp/content/1407073_15_1_1_2.pdf (2025 年 11 月 8 日参照)

[5] 綿貫 真也(2021), マーケティング分野におけるデータサイエンス学修カリキュラムの検討—文系学部生、文系学部出身社会人の「データサイエンスへの抵抗感」に配慮して—, コンピュータ&エデュケーション, Vol. 50, pp.60-65

[6] 清水 真 (2016) , 「3.4.3 情報管理」 pp.345-352, 牛腸ヒロミ・佐々井啓・平田耕造ほか編『被服学辞典』朝倉書店

[7] 総務省(2024), 地域におけるデジタル・トランスフォーメーション

https://www.soumu.go.jp/denshijiti/digital_transformation.html (2025 年 11 月 8 日参照)

[8] Data Mix(2023), 企業や自治体におけるデータサイエンスの活用事例と人材育成の方法, <https://datamix.co.jp/media/datascience/case-study-company-data-science/> (2025 年 11 月 8 日参照)

[9] リプロネクスト(2025), メタバースと XR が拓く自治体の可能性 | 行政・観光・福祉の最新活用事例,

https://lipronext.com/knowledge/localgovernment-dx-casestudy/?utm_source=chatgpt.com (2025 年 11 月 8 日参照)

課題

- ① 自治体でのデータサイエンスの応用例を調べて説明しなさい。
- ② マーケティング分野でのデータサイエンスの応用例として、オンラインショッピングサイトでの顧客への商品推薦があります。そこでは、どのようなデータが収集され、どのような分析されているか説明しなさい。
- ③ 教育分野において、学習者のデータを分析して学習管理や学習支援を行うことの意義と、その際に注意すべき点について述べなさい。

第 13 講 データ可視化の高度な技術

荒木貴之（日本経済大学/ 社会構想大学院大学・教授）

【学習到達目標】

- ① ネットワーク分析の基礎を習得する：物事の「関係性」をノードとエッジで表現し、中心性指標を用いてキーパーソンやハブを発見する方法を学ぶ。
- ② 地理空間情報（GIS）の多次的表現を理解する：地図データに時間軸や 3D 表現（高さ）を加えることで、地域課題や歴史的変遷を動的に可視化する手法を身につける。
- ③ デジタルアーカイブにおける 3D 技術の意義を学ぶ：フォトグラメトリなどの技術が、単なる保存を超えて「コンテキスト（文脈）の再現」にどう寄与するかを理解する。

第 4 講「データの可視化と探索的データ分析（EDA）」で学んだ基礎を発展させ、複雑な「つながり（関係性）」と「広がり（時空間）」を可視化する先端技術を扱います。歴史的人物の相関図を描く「ネットワーク分析」、地図上で時間の流れを追う「GIS（地理空間情報システム）」、そして文化財を立体的に保存・活用する「3D アーカイブ」。これらを通じて、見えない文脈を可視化（Visualize）し、新たな知見を創出するデジタルヒューマニティーズの世界へ誘います。

1. 見えない絆を描く「ネットワーク分析」

1-1 つながりの構造化：ノードとエッジ

歴史資料や組織図には、目に見えない複雑な人間関係が含まれています。これを可視化するのがネットワーク分析です。基本要素は 2 つだけです。

- ノード（Node）：点。人物、場所、単語など。
- エッジ（Edge）：線。友人関係、取引、手紙のやり取りなど。

例えば、戦国武将の同盟関係をネットワーク図にすることで、誰と誰が親密で、どの陣営が孤立しているかが一目瞭然になります。エッジには「向き（矢印）」や「重み（太さ）」を持たせることができ、情報の流れや関係の強さを表現できます。

1-2 「中心性」でキーパーソンを見つける

ネットワーク図は、計算によって「誰が重要か」を数値化できます。これを「中心性 (Centrality)」と呼びます。

- 次数中心性 (Degree Centrality): つながっている相手の数。単純な「顔の広さ」や「人気」を示します。
- 媒介中心性 (Betweenness Centrality): ネットワーク内の異なるグループ同士を結ぶ「橋渡し」の度合い。この値が高い人物は、一見目立たなくても情報の流通を握る「黒幕」や「仲介者」であり、歴史分析や組織マネジメントにおいて極めて重要な発見となります。

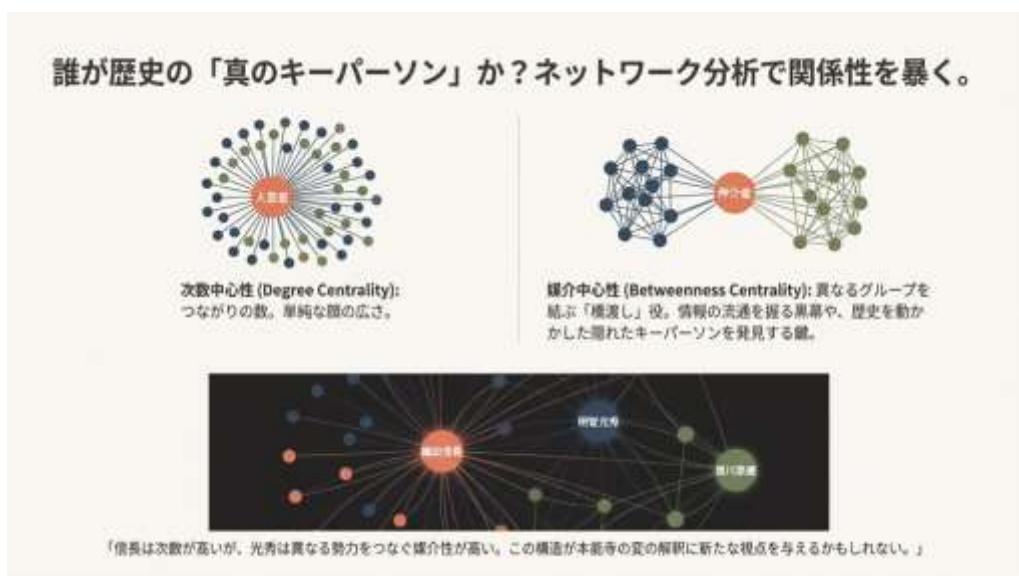


図 1 3-1 ネットワーク分析の例

あなたの職場で、明日から始められる「ノーコード」可視化ツール

Flourish	Kepler.gl
Best for: 関係性の可視化、ストーリーテリング	Best for: 地理空間データの可視化、大規模データの分析
Why: 英国のデータジャーナリズムチームが開発。Excelからデータをアップロードするだけで、プレゼンテーション品質の動的なネットワーク図やグラフを生成。	Why: Uberが開発したオープンソース。数万件のデータをブラウザで高速に3D表示。無料で全機能が利用可能。

高度な可視化に、プログラミングは必須ではありません。
重要なのは、データをどう見せ、何を伝えたいかという視点です。

図 1 3-2 ノーコード可視化ツールの例

本講座で紹介するツール「Flourish」などを使えば、Excel でリストを作るだけで、これらを動的に可視化できます。

参考) Flourish

- ・ URL <https://flourish.studio/>
- ・ データ可視化やインタラクティブ要素から没入型ストーリーテリングまで、複雑なデータをインタラクティブなビジュアルに変えます。

【発展実習】Flourish によるネットワーク分析

無料のデータ可視化ツール「Flourish」を使用して、Excel 等の表データ (CSV) から動的なネットワーク図を作成する手順を解説します。

1. Flourish アカウントの作成: 公式サイト (<https://flourish.studio/>) にアクセスし、「Sign up for free」からアカウントを作成してください (Google アカウントでのログインが便利です)。
2. データセットを準備します。「歴史人物データリスト」(表 1 3-1) を csv ファイル「Gifu_History_Network.csv」として生成し、デスクトップ等のわかりやすい場所に保存してください。
3. Flourish にログインし、画面上部の「+ New visualisation」ボタンをクリックします。
4. タブ「Start with data (BETA)」を選択すると、「Upload your dataset」のボタンが表示されるので、あらかじめ作成しておいた「Gifu_History_Network.csv」ファイルをアップロードします。
5. さまざまな可視化されたグラフが提示されます。グラフから傾向を読み取り、ネットワーク分析を行ってください。

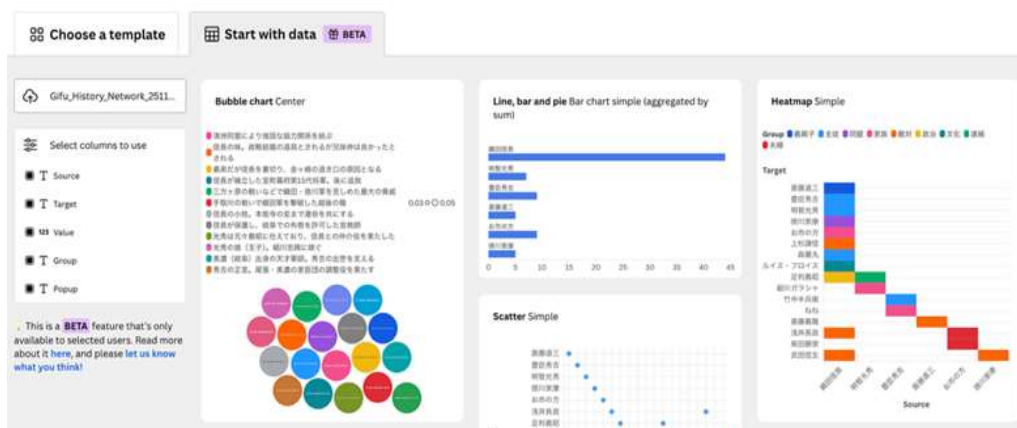


図 1 3-3 Flourish によるネットワーク分析

表 1 3 — 1 歴史人物データリストの例 (Flourish 用)

Source	Target	Value	Group	Popup
織田信長	斎藤道三	3	義親子	道三は信長の義父であり、美濃（岐阜）を譲り渡す意向を示した
織田信長	豊臣秀吉	5	主従	草履取りから天下人へと出世した信長の筆頭家臣
織田信長	明智光秀	4	主従	織田家の重臣だが、後に本能寺の変を起こす
織田信長	徳川家康	5	同盟	清洲同盟により強固な協力関係を結ぶ
織田信長	お市の方	5	家族	信長の妹。政略結婚の道具とされるが兄妹仲は良かったとされる
織田信長	浅井長政	5	敵対	義弟だが信長を裏切り、金ヶ崎の退き口の原因となる
織田信長	足利義昭	3	政治	信長が擁立した室町幕府第 15 代将軍。後に追放
織田信長	武田信玄	5	敵対	三方ヶ原の戦いなどで織田・徳川軍を苦しめた最大の脅威
織田信長	上杉謙信	4	敵対	手取川の戦いで織田軍を撃破した越後の龍
織田信長	森蘭丸	3	主従	信長の小姓。本能寺の変まで運命を共にする
織田信長	ルイス・フロイス	2	文化	信長が保護し、岐阜での布教を許可した宣教師
明智光秀	足利義昭	3	連絡	光秀は元々義昭に仕えており、信長との仲介役を果たした
明智光秀	細川ガラシ	4	家族	光秀の娘（玉子）。細川忠興に嫁ぐ
豊臣秀吉	竹中半兵衛	4	主従	美濃（岐阜）出身の天才軍師。秀吉の筆頭家臣
豊臣秀吉	ねね	5	家族	秀吉の正室。尾張・美濃の家臣団の調整役を果たす
斎藤道三	斎藤義龍	5	敵対	実の親子だが骨肉の争いを繰り広げた（長良川の戦い）
お市の方	浅井長政	5	夫婦	政略結婚だが夫婦仲は極めて良かったとされる
お市の方	柴田勝家	4	夫婦	信長の死後、筆頭家老の勝家と再婚
徳川家康	武田信玄	5	敵対	三方ヶ原の戦いで家康は大敗を喫す

※ 左記のデータは、演習用データとして作成したものであり、個人の用途内でご利用ください。

2. 地図上の物語「地理空間情報（GIS）」

2-1 ジョン・スノウと疫学の地図

GIS（Geographic Information System）の原点は、1854年のロンドンで医師ジョン・スノウが描いた地図にあります。彼はコレラ患者の発生場所を地図上にプロットすることで、特定の井戸が感染源であることを突き止めました。これは「位置情報」が課題解決の決定打になることを示した歴史的事例です。

地図は物語る：一枚の地図がロンドンの街を救った物語



1854年、ロンドン。医師ジョン・スノウは、コレラ死亡者の住所を地図上にプロットすることで、感染源が特定の井戸であることを視覚的に証明しました。

これは、位置情報が課題解決の鍵や仲翫は課題解決の鍵となることを示した、地理空間情報システム（GIS）の原点です。単なる住所録では見えないパターンが、地図に重ねることで初めて浮かび上がります。

図1 3-3 GIS（Geographic Information System）の例

2-2 現代のGIS：レイヤーと3次元化

現代のGISツール（例：Kepler.gl）は、この概念を拡張します。

- レイヤー構造: 地図の上に、道路、建物、人口分布、観光ルートなど、複数の情報を透明なフィルムのように重ねて分析します。
- 3D 可視化: 例えば観光客数を地図上の「棒の高さ」で表現することで、平面のヒートマップよりも直感的に「どこに人が集中しているか」を把握できます。
- 時間軸（タイムライン）: 時間の経過とともにデータが動くアニメーションを作成することで、朝と夜の人流の変化や、数十年単位の都市の発展を物語として表現できます。

岐阜の観光データや古地図をGISに載せることで、単なる場所の記録が「生きた地域の活動記録」へと変わります。



図 1 3 - 4 観光データを GIS で分析する例

参考) Kepler.gl

- ・ URL <https://kepler.gl/>
- ・ 地理空間データ分析や大規模データの分析ができます

ブラウザ上で大量の位置データを可視化したり、時間軸に沿った地理的トレンドを再生したり、位置データを探索・フィルタリングし、深く関与して洞察を得ることができます。

3. 時間を超える「3D アーカイブと VR」

3-1 「モノ」から「空間」の保存へ

デジタルアーカイブの技術は、2次元の画像スキャンから3次元（3D）へと進化しています。

フォトグラメトリ（Photogrammetry）は、対象物を多方向から撮影した多数の写真から、精密な 3D モデルを生成する技術です。これにより、博物館のガラスケース越しでしか見られない土器や仏像を、Web ブラウザ上で自由に回転させ、裏側や細部の質感まで観察することが可能になります。

さらに、ドローン計測やレーザースキャンを組み合わせることで、建物や遺跡、あるいは「祭り」のような空間そのものを丸ごとデジタル化することも進んでいます。



図1 3-6 フォトグラメトリ（Photogrammetry）の例

3-2 メタバース・VRによる「体験」の提供

作成された3Dデータは、VR（仮想現実）やメタバース空間で活用されることで真価を発揮します。

消失した城郭をVRで復元して現地で体験したり、遠隔地の学生がアバターとしてデジタルミュージアムに集まり、学芸員の解説を聞きながら資料を手にとったりする。これは「保存」を超えた、デジタルアーカイブによる「教育と観光の革新」です。岐阜女子大学デジタルアーカイブ研究所が推進する取り組みも、まさにこの未来を見据えています。

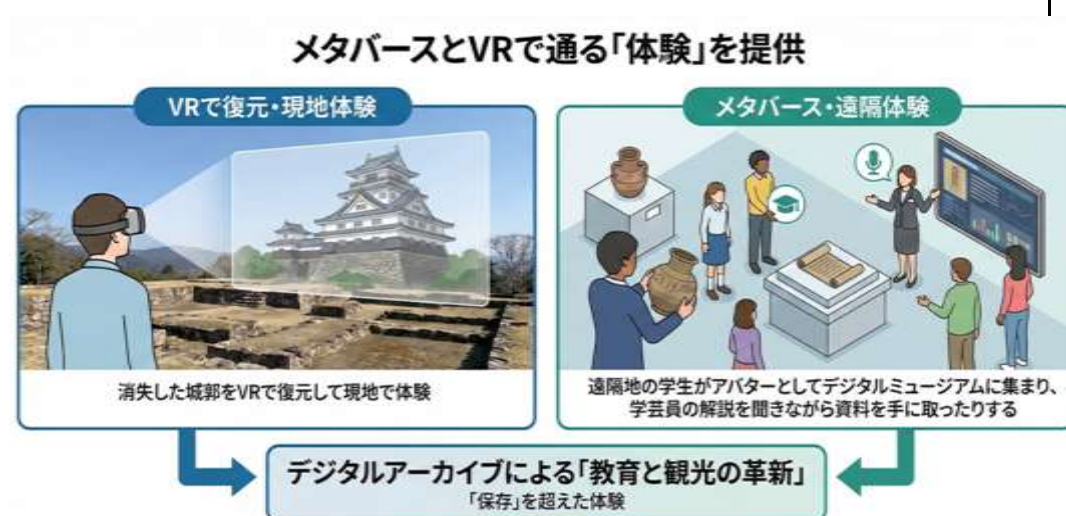


図1 3-7 メタバースやVRによる体験の例

4. まとめ

第 13 講では、単純なグラフを超えた高度な可視化技術について学びました。

1. 関係性の可視化: ネットワーク分析を用いれば、複雑な人間関係や歴史的
文脈の中から、構造的な特異点（ハブとなる人物など）を発見できる。
2. 時空間の可視化: GIS を用いれば、データに「位置」と「時間」の情報を
与え、地域課題を動的なストーリーとして語ることができる。
3. 実在感の可視化: 3D 技術や VR は、物理的な制約を超えて文化財を「体
験」可能なものにし、アーカイブの価値を飛躍的に高める。

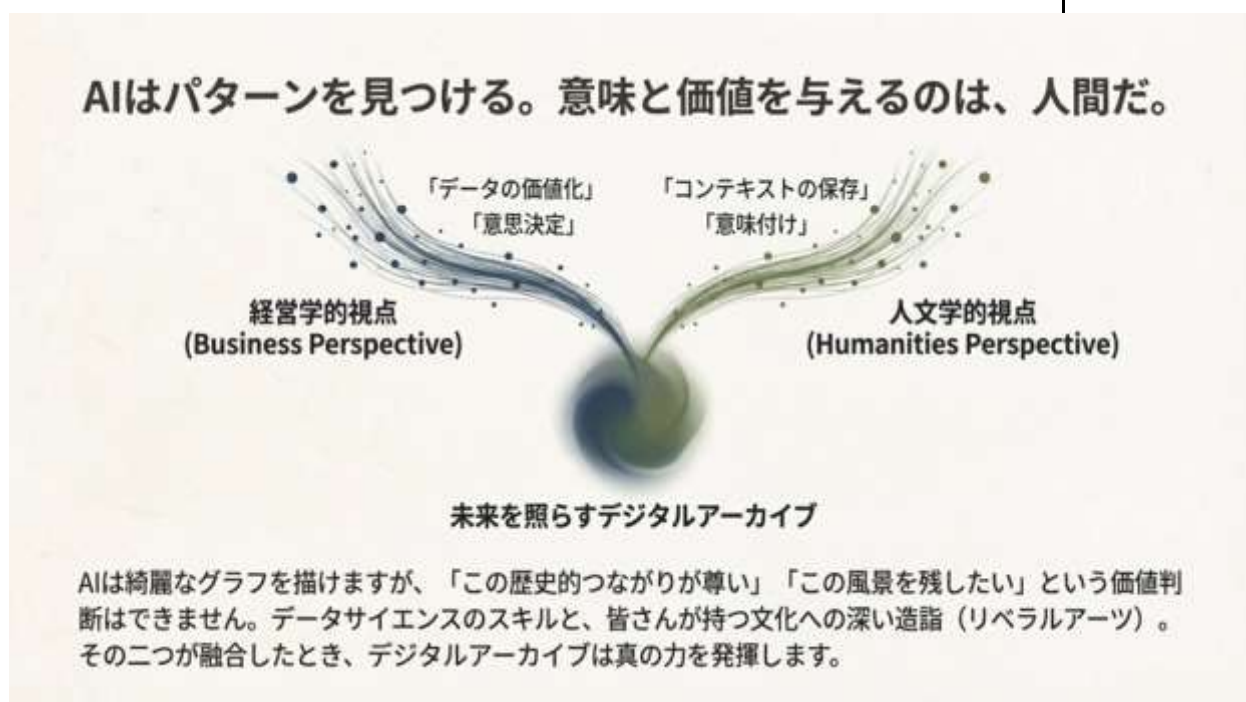


図 1 3 – 8 デジタルアーカイブにおけるデータサイエンスとリベラルアーツの融合

これらの技術は、決してエンジニアだけのものではありません。どのようなデータを集め、何を表現したいかという「問い」を持つのは、地域や文化を知る皆さん自身です。データサイエンスの技術と、リベラルアーツ（教養）の融合こそが、これからのデジタルアーカイブを支える鍵となります。

参考文献

1. Tutorial- Network Analysis With Flourish - Hacking the Humanities 2022F, <https://hh2022f.amason.sites.carleton.edu/week-9-project-preparation/tutorial-network-analysis-with-flourish/index.html> (2025 年 11 月 29 日確認)
2. Build a simple network graph with Flourish - Miriam Posner, <https://miriamposner.com/classes/dh201w21/tutorials-guides/network-analysis/flourish-graph/> (2025 年 11 月 29 日確認)
3. Create a Network Graph with BanyanDNA and Flourish - YouTube, <https://www.youtube.com/watch?v=qrzDPuvc0UA> (2025 年 11 月 29 日確認)
4. Revealing Data: Visualizations in Historical Collections, <https://circulatingnow.nlm.nih.gov/2023/07/13/revealing-data-visualizations-in-historical-collections/> (2025 年 11 月 29 日確認)
5. 遺跡の地図データを見てみよう！ kepler.gl 編, <https://www.city.tondabayashi.lg.jp/uploaded/attachment/88748.pdf> (2025 年 11 月 29 日確認)
6. 「見るできないもの」の可視化 - TOPPAN VR・デジタルアーカイブ, <https://www.toppan-vr.jp/bunka/vr/visualize.shtml> (2025 年 11 月 29 日確認)

課題

1. 「私のネットワーク」の設計

自分自身、または歴史上の人物（織田信長など）を中心とした「ネットワーク図」の構想を練ってください。「誰」がノードとなり、「どのような関係」がエッジとなるか？ また、その図において「媒介中心性」が高い（異なるグループをつなぐ）人物は誰になりそうか、仮説を記述してください。

2. 地域資源の GIS 活用案

あなたの住む地域、あるいは岐阜県内の特定のエリアを対象に、「地図に重ね合わせることで新たな発見がありそうなデータ」の組み合わせを提案してください。（例：「古地図」×「現在の浸水被害想定区域」、「昭和の写真撮影地点」×「現在の観光ルート」など）。なぜその組み合わせが有効なのか、理由も添えてください。

3. 未来のミュージアム企画

3D 技術や VR、メタバースが普及した 10 年後の未来において、デジタルアーカイブを活用した「全く新しい博物館の展示」を企画してください。物理的な制約（場所、保存状態、アクセス）を取り払った時、どのような体験が可能になるか、自由な発想で記述してください。

第 14 講 AI と深層学習の基礎と応用

藤吉弘亘（中部大学 A I 数理データサイエンスセンター教授）

【学習到達目標】

- ・ AI の発展と深層学習の基本的な概念と仕組みを説明できる。
- ・ 深層学習の代表的なモデルの特徴を理解し、適用例を説明できる。
- ・ LLM・VLM・VLA の関係と役割を整理して説明できる。

1. AI の歴史とブームの変遷

人工知能（Artificial Intelligence; AI）とは、人間が行っている知的な活動を計算機上で実現しようとする試みの総称である。推論、学習、認識、計画、対話など、その対象は多岐にわたる。近年の AI の急速な発展を支えている中心技術が「深層学習（ディープラーニング）」であり、多数のパラメータをもつニューラルネットワークを用いて、大量のデータから自動的に特徴を抽出し、さまざまなタスクを高精度に実現する学習手法である。現在の AI ブームを正しく理解するためには、AI の歴史を俯瞰し、ルールベースの時代から機械学習、深層学習、さらに大規模基盤モデルへと至る流れを押さえておくことが重要である。

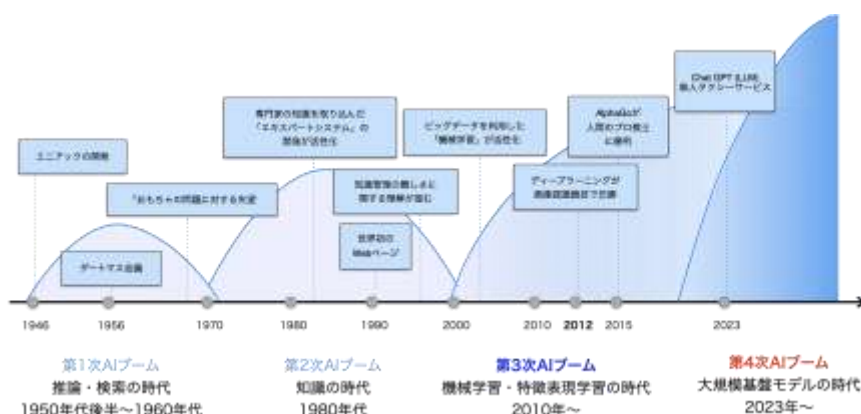


図 1 AI ブームの変遷

(1) 第 1 次 AI ブーム：記号処理と推論の時代

1950～1960 年代にかけての第 1 次 AI ブームでは、「記号処理」に基づ

く AI が研究の中心であった。知識やルールを記号（シンボル）として表現し、論理式を用いて推論を行うことで、人間のような思考を再現しようとしたのである。この時期には「人工知能（Artificial Intelligence）」という用語が定義され、推論プログラムや定理証明システムなどが開発されたが、現実世界の複雑な問題に対しては性能が伸び悩み、期待ほどの成果は得られなかった。計算機資源やデータの制約も大きく、やがて第 1 次ブームは収束することになった。

(2) 第 2 次 AI ブーム : エキスパートシステムと知識表現

1980 年代に起こった第 2 次 AI ブームでは、「エキスパートシステム」が大きな注目を集めた。エキスパートシステムとは、医師や技術者などの専門家が持つ知識を多数の IF-THEN ルールとしてシステムに組み込み、そのルールに基づいて診断や助言を行うシステムである。このアプローチは、ある程度限定された領域では高い性能を発揮したものの、次のような問題を抱えていた。

- 専門家の知識をルールとして大量に書き出す「知識獲得」に多大なコストがかかること
- ルールが増えるほど、ルール同士の矛盾や抜け漏れの管理が難しくなること

これらがボトルネックとなり、汎用的かつ拡張性の高い人工知能を実現するには限界があることが明らかになった。

(3) 第 3 次 AI ブーム : 機械学習と深層学習

2010 年頃から本格化した第 3 次 AI ブームでは、人手でルールを書くのではなく、「データから法則を学習する」機械学習が主役となった。その中でも特に重要なのが「深層学習」である。深層学習を用いた画像認識モデルは、2012 年の国際コンテスト（ImageNet Large Scale Visual Recognition Challenge）において従来手法に大差をつけて優勝し、その有効性を世界に示した。それ以降、画像認識、音声認識、自然言語処理など、多くの分野で深層学習が従来手法を凌駕する精度を達成し、第 3 次 AI ブームを牽引する存在となった。この流れの象徴的な出来事が、囲碁 AI「AlphaGo」による世界トップ棋士への勝利である。

(4) 第 4 次 AI ブーム : 大規模基盤モデルと生成 AI

2025 年の現在では、深層学習をさらに大規模化し、多様なタスクに対応可能な「大規模基盤モデル (Foundation Model)」が登場した第 4 次 AI ブームの段階にあると考えられている。膨大なテキスト、画像、音声、動画などを事前学習したモデルは、テキスト生成、翻訳、要約、プログラム生成、画像・音声の生成など、多様な「生成タスク」を高い品質で実行できるようになった。これらは総称して「生成 AI」と呼ばれ、社会・産業・教育など、さまざまな領域にインパクトを与えつつある。

2. 囲碁 AI AlphaGo と深層学習・強化学習

囲碁 AI「AlphaGo (アルファ碁)」は、第 3 次 AI ブームを象徴する存在として、人工知能の歴史に大きな足跡を残したシステムである。2016 年に世界トップクラスのプロ棋士を破ったニュースは、単に「コンピュータが囲碁に勝った」という話題にとどまらず、「人間の直感や経験に頼ってきた領域にも、深層学習と強化学習が入り込んできた」という象徴的な出来事として受け止められた。そもそも、囲碁は長らく「ゲーム AI の最後の砦」と言われてきた。オセロやチェスのようなボードゲームでも AI は強力であったが、それらは計算資源の増加と探索アルゴリズムの工夫により、かなり早い段階で人間を凌駕していた。しかし囲碁の場合、盤面は 19×19 の交差点で構成され、碁石を置く場所の選択肢が非常に多い。そのため、可能な局面や手順の数、すなわち探索空間は天文学的な大きさになるとされている。この難題に対して AlphaGo は、「深層学習 (畳み込みニューラルネットワーク)」と「強化学習」を組み合わせるというアプローチを取った点である。

(1) 畳み込みニューラルネットワークによる教師あり学習

まず AlphaGo は、過去の棋譜データから「上級者がどのような手を打ってきたか」を学習する。インターネット上には、プロや高段者の対局記録、すなわち棋譜が多数公開されている。AlphaGo はそれらを大量に集め、盤面の状態を画像のように入力し、「その局面で実際に人間が打った一手はどこか」を正解として学習したのである。このとき使われたのが、畳み込みニューラルネットワーク (CNN) と呼ばれる深層学習モデルである。盤面の状態を入力とすると、ネッ

トワークは 19×19 の全ての交差点に対して「ここに打つ確率がどれくらい高いか」を出力する。学習の初期段階では、重みはランダムに近いため出力もでたらめであるが、多数の棋譜に対して誤差を小さくするようにパラメータを更新していくことで、次第に「人間の上級者が選びそうな手」を打つことができる。

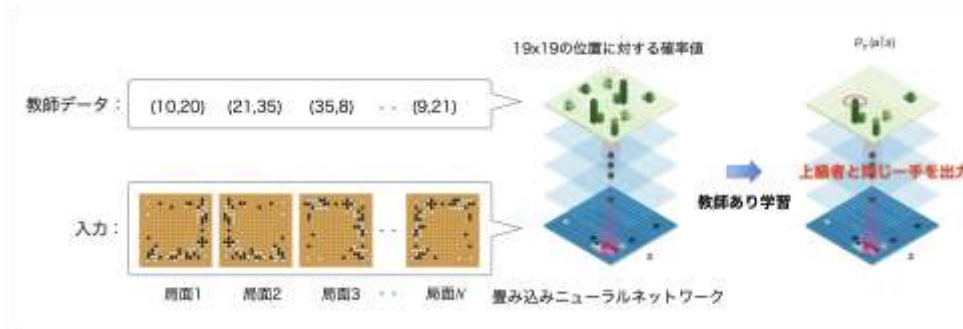


図2 CNNの教師あり学習

(2) 経験から学ぶ強化学習

しかし、上級者のまねをするだけでは世界チャンピオンに勝つことはできない。プロ棋士同士の対局記録は確かに質の高いデータであるが、それは「過去に人間が経験した局面」に限られている。未知の局面や、まったく新しい戦略については、過去の棋譜に頼るだけでは対応しきれない。そこで AlphaGo は、人間の棋譜から学んだモデルを基礎としつつ、その後は自分自身と対局を繰り返す「自己対戦」によって、さらなる実力向上を図っている。ここで用いられる枠組みが「強化学習」である。強化学習では、個々の手に対する「正解」が明示的に与えられるわけではない。代わりに、自己対戦で勝利した側にはプラスの報酬が、敗北した側にはマイナスの報酬が与えられる。そして、その結果に基づいて、対局中に選んだ手の確率が調整される。勝ったときによく出現した手は、今後選ばれやすくなるように確率が高められ、負けたときによく出現した手は、確率が下げられていくのである。このような自己対戦は、一回や二回の対局では意味がない。AlphaGo は多数のコンピュータを用いて、1日に膨大な数の対局を繰り返したと報告されている。人間が一生をかけても経験できないほどの対局経験を、短期間に積み重ねることができるわけである。その結果、「人間の棋譜に基づく上級者レベルの打ち方」を土台としながらも、人間が思いつかなかったような新しい打ち方や、長期的な勝ちやすさを意識した戦略を自ら発見していくことになる。

AlphaGo は、人間の棋譜という既存の知識を起点にしつつ、自身との対局から新しい知識を獲得しており、データから学び、経験からさらに強くなる AI と言える。

3. 大規模基盤モデルと生成 AI

「第 4 次 AI ブーム」とも呼べる潮流が生まれ、その中心にあるのが大規模基盤モデル（Foundation Model）である。従来の AI は、「タスク専用のモデル」を一つずつ作るのが基本であった。顔認識をしたければ顔認識用のモデル、翻訳をしたければ翻訳用のモデル、音声認識なら音声認識用のモデル、といった具合である。ところが大規模基盤モデルは、その発想を大きく変える。膨大なテキスト、画像、音声、動画などをまとめて学習しておき、その一つの巨大なモデルをさまざまなタスクに“流用”するのである。

こうしたモデルは、数十億から数兆といった桁のパラメータを持ち、高性能な GPU や専用チップを用いて長時間かけて学習される。ここで重要になるのが「生成 AI（Generative AI）」という概念である。生成 AI とは、与えられた入力や文脈に基づいて、新しいテキスト、画像、音声、プログラムコードなどのコンテンツを自動生成する AI モデルの総称であり、大規模基盤モデルはその代表的な実装形態である。いったん学習が終われば、プロンプトの工夫や少量の追加学習によって、文章生成、翻訳、要約、質問応答、プログラム生成、画像の説明、さらには画像生成や音声合成といった多様な生成タスクをこなすことができる。

このように、一つの大規模基盤モデルが生成 AI として多目的に働き、「何でも相談できる AI アシスタント」のような振る舞いを見せることこそが、第 4 次 AI ブームを支える原動力となっているのである。

(1) Transformer と大規模言語モデル

言語モデルは、「次に来る単語の確率」を学習したモデルであり、「英国の首都は」の後に「東京」が続く確率、「パリ」が続く確率、「ロンドン」が続く確率を計算し、最も確率の高い単語を選び出すのである。言語モデルとして用いられる Transformer は、「自己注意機構（Self-Attention）」と呼ばれる仕組みを用いて、文章中のすべての単語同士が互いに直接参照し合えるようにした。

「この単語は、あの単語と強く関係している」といった類似度を計算し、それをもとに重要な単語に多く“注意”を向けることで、文脈全体を効率的に捉える。これにより、長い文章や複雑な依存関係を持つ文でも、比較的安定して扱うことができるようになった。この Transformer を用いて、Web 上の膨大なテキストを読み込ませたものが、大規模言語モデル(LLM)である。LLM は、「次に来る単語」を予測するという単純な作業を延々と繰り返しながら、文法や語彙だけでなく、世界の一般常識や、さまざまな専門分野の断片的な知識まで取り込んでいく。こうして訓練されたモデルは、ただの「次の一語予測器」でありながら、結果として文章生成、翻訳、要約、質問応答、プログラム生成などを実現した。GPT (Generative Pre-trained Transformer) は、その代表例である。

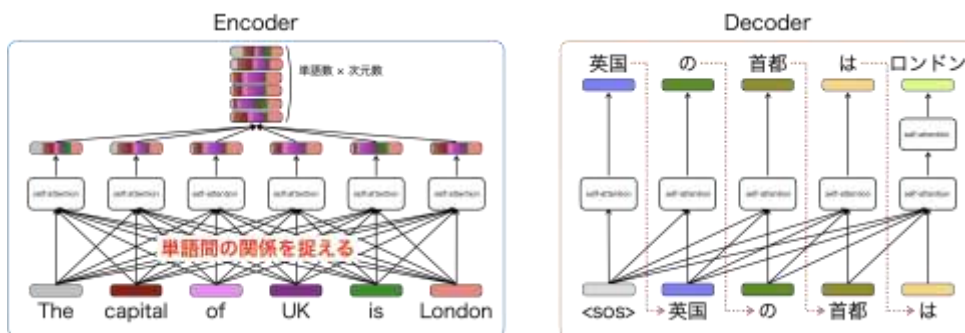


図 3 Transformer

(2) プロンプトと Chain-of-Thought

プロンプトとは、モデルに与える入力文、すなわち指示文のことである。「この文章を要約してください」と、「大学 1 年生にも分かるように、300 字程度で要約してください」とプロンプトにより、出力される文章は大きく変わる。このプロンプトの興味深いテクニックが「Chain-of-Thought (思考の連鎖)」である。これは、「途中の考え方も含めて説明して」と促すプロンプトの書き方である。例えば、ある算数の文章題に対して、「まず問題の条件を整理し、そのあと式を立て、最後に計算して答えを出してください」とすると、モデルは、1. 条件を箇条書きに整理し、2. 必要な式を導き、3. 計算を行い、4. 最終的な答えを示す、という“解き方”を文章として出力するようになる。このプロセスを経由することで、モデルが複雑な問題を解きやすくなるのである。

(3) 大規模言語モデルの限界と外部知識の利用

LLM は事前学習のときに大量のテキストを読み込むが、その後に世界で起きた出来事に関するテキストを学習しているわけではない。そのため、ある年までのニュースや論文で学習したモデルは、「その年までの世界」については詳しいが、それ以降の出来事については回答できない。もう一つの限界は、計算や厳密な論理である。言語モデルは「次に来そうなトークン」を予測する仕組みにすぎないので、複雑な数値計算や形式的な証明を内部で厳密に行っているわけではない。そのため、桁の多い掛け算や入り組んだ計算問題に対して、もっともらしいが誤った答えを返すことがある。

こうした弱点を補うために登場したのが、RAG (Retrieval-Augmented Generation) のような枠組みである。RAG は、ユーザーの質問を一度検索モジュールに渡し、企業内文書や Web ページ、論文などから関連する情報を取得する。その検索結果の一部をプロンプトに含めてモデルに入力することで、モデルが本来獲得していない最新の知識や専門的な情報を取り入れて回答することができる。

4. 画像と言語を結びつける VLM (Vision-Language Model)

大規模言語モデルと言語以外のモダリティ（画像や音声など）を組み合わせたモデルが盛んに研究されている。その一つが、画像とテキストを統合的に扱う「VLM (Vision-Language Model)」である。

(2) 画像と言語のアライメントとコントラスト学習

VLM を構築するうえで鍵となるのが、画像とテキストの対応関係（アライメント）を学習することである。典型的な枠組みとして、

- 画像と、その画像を説明するテキスト（キャプション）のペアを大量に用意する。
- 画像から特徴ベクトルを抽出する「画像エンコーダ」と、テキストから特徴ベクトルを抽出する「テキストエンコーダ」を用意する。

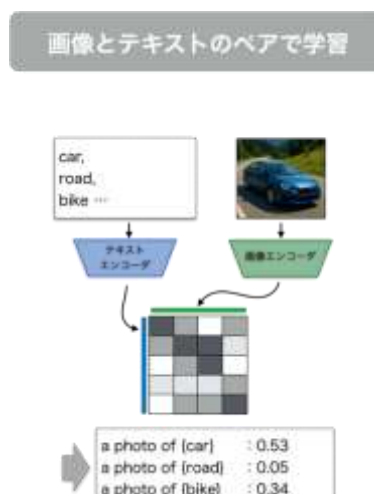


図4 画像と言語のアライメント

- 同じ内容を表す画像とテキストの特徴が近くなり、異なる内容を表すペアが離れるように、コントラスト学習（対照学習）を行う。

といった手法が用いられる。

これにより、「車の画像」と「car」というテキストがベクトル空間上で近接し、画像だけを見ても「これは車である」という言語的概念にアクセスできるようになる。画像エンコーダには、画像をパッチに分割して Transformer で処理する Vision Transformer (ViT) 、テキストエンコーダには、Transformer 型の言語モデルが利用される。

(3) VLM (Vision-Language Model)

こうして学習した画像エンコーダと大規模言語モデル(LLM)を組み合わせることで、VLM を構成することができる。画像特徴を LLM の入力トークンと同じ次元に射影し、LLM の入力列に「視覚トークン」として埋め込むことで、画像と言語を統一的に処理する。このような VLM を用いることで、以下のようなタスクが実現可能となる。

- 画像キャプション生成：画像の内容を自然な文章で説明する。「公園で子どもがボール遊びをしている」など。
- 視覚質問応答（Visual Question Answering; VQA）：画像と質問文を入力し、「この画像の中で赤い物体は何か」「右側にいる人物は何をしているか」といった問いに答える。
- 画像内の関係性推論：単に物体のラベルを認識するだけでなく、「この缶は潰れており、飲み終わった後のゴミである」「このボトルはキャップが閉まっており、中身が残っているのでまだ飲める」といった属性や状態、関係性を推論する。

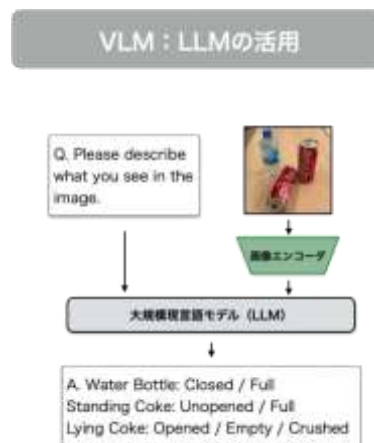


図 5 VLM

(4) 画像・言語・行動を統合する VLA (Vision-Language-Action Model)

VLM は画像と言語を統合するモデルであるが、近年ではこれに「アクション（行動）」の生成能力を加えた「VLA (Vision-Language-Action Model)」が登場している。VLA は、ロボット制御やエージェント制御への応用を念頭に置いた枠組みであり、「見て理解し、言葉で指示を受け、その指示に基づいて行動する」システムの実現を目指している。VLA の利点の一つは、言語モデルが持つ豊富な世界知識により、学習時に見たことのない物体やタスクの組み合わせにも柔軟に対応しうる点である。

例えば、「飲み終わったコーラの缶をゴミ箱に捨ててください」という指示が与えられたと

き、VLA (Vision-Language-Action) モデルを使ったロボットは、まずカメラの画像からテーブル上の物体を認識する。「金属の缶が 2 つある」「そのうち 1 つは潰れていて、飲み口が開いている」「もう 1 つは形がきれいで、まだ開封されていない」といった状態を画像から読み取る。同時に、言語モデルは「飲み終わったコーラの缶」という日本語の意味を解釈する。「コーラ缶の中身がもうない」「通常はゴミとして扱うもの」といった世界知識が呼び出される。その結果、「潰れていて飲み口が開いている缶こそが、指示されている対象だ」と判断することができる。続いて、ロボットは潰れた缶の位置を特定し、アームを伸ばして適切な力加減でつかみ、ゴミ箱の位置まで運び、中に落とす動作を生成する。

このように、VLA によって、ロボットは「決められた動きだけをする機械」から、「見て・聞いて・考えて・動く汎用エージェント」へと進化する。カメラで周囲の状況を認識し、人の自然な言葉による指示を理解し、その場その場で行動の手順を自分で組み立てて実行するロボットである。これにより、工場のような決められた環境だけでなく、家庭やオフィス、サービス現場といった予測しにくい状況の中でも、ある程度柔軟に対応できるロボットが現れていく。また、大規模な VLA を共通の「頭脳」として共有し、クラウド経由で新しいスキルや知識をアップデートしていく方向性も強まるだろう。



図 6 VLA

課題

- ① 深層学習と従来の機械学習の違いについて、具体例を挙げて説明しなさい。
- ② 畳み込みニューラルネットワーク（CNN）の特徴と、その代表的な応用例を述べなさい。
- ③ 深層学習の社会的な課題や倫理的な問題について、あなたの考えを述べ、その解決策の一例を提案しなさい。

第15講 データサイエンスの未来と教育への展望

澤井進（岐阜女子大学特任教授）

【学習到達目標】

- ① データサイエンスの未来展望と教育的意義を理解する。
- ② 教育現場における役割と必要性を説明できる。
- ③ 発展に伴う課題とそれに対する教育の対応策を考察できる。

1. データサイエンスとは

データサイエンスは、ビッグデータを収集・分析・解釈し、社会課題の解決や意思決定の支援を行う学問分野です。

データサイエンスは、私たちの社会のさまざまな分野で重要性を増しています。

2. 未来展望

産業界の医療、金融、交通の各分野について見てみましょう。

・医療：個別化医療、早期診断

医療分野では、個別化医療や早期診断を可能にし、特定の遺伝性疾患の早期発見精度を大幅に向上させています。

米国の個別化AI診断システムで、特定の希少遺伝性疾患の早期発見精度が従来の60%から93%に向上した例があります。

・金融：リスク管理、詐欺検出

金融分野では、リスク管理や詐欺検出に利用され、取引データから不正リスクを自動的に評価するシステムが開発されています。

NECが、横浜銀行用にAI不正・リスク検知サービスを開発し提供しました。

この事例では、横浜銀行の取引システムの一次審査を、従来は人手で行っていました。

一次審査の初期審査の一部をAIで採点（スコアリング）し、その結果を人間による2次審査に提供し省力化したということです。

・交通：自動運転、交通流の最適化

交通分野では、自動運転や交通流量の最適化に貢献し、AIを活用した無人自動運転タクシーが既に実用化されています。

米国では、Google 社の関連会社の「W a y m o」（ワイモ）が、世界で最も進んでいる無人の自動運転タクシーでレベル4のロボットタクシーですが、サンフランシスコ、フェニックス、ロサンゼルス の3都市で一般向けに稼働させています（2024年11月末現在）。

今後、ニューヨークでも稼働させるとか、日本のトヨタ自動車もW a y m oを導入するとか報道されています。

3. 教育におけるデータリテラシー

情報化社会に適応するため、システム思考、プログラミングの基礎、統計学の基礎、データ可視化と解釈力が重要なスキルとなります。主体的に意思決定ができる市民を育成するために、概念設計ができ、AIに何をさせるか考えられる能力が必要です。つまり、システム思考が重要になります。

学校教育で必要なスキルとしては、プログラミングの基礎、統計学の基礎、及びデータ可視化と解釈力です。

1) プログラミングの基礎

①「つるかめ算」

例えば、「つるかめ算」のプログラミングでは、まず概念設計シートを作成し、続いてプロンプト（指示文）を生成します。生成したプロンプトは、「貴方は有能なAI教師です。つるかめ算です。足50本で20匹の場合のプログラムを作成してください。」です。

```
Python # 変数を設定します
total_animals = 20 # 全体の生き物の数
total_legs = 50    # 全体の足の数
crane_legs = 2     # 鶴の足の数
turtle_legs = 4    # 亀の足の数
# 解決方法を応用します
# もし全部が鶴だった場合の足の数を計算
assumed_total_legs = total_animals * crane_legs
# 実際の足の数との差を計算
leg_difference = total_legs - assumed_total_legs
# 亀の数を計算（足の差を、鶴と亀の足の数の差で割る）
num_turtles = leg_difference // (turtle_legs - crane_legs)
# 鶴の数を計算
num_cranes = total_animals - num_turtles
# 結果を表示
print(f"鶴の数: {num_cranes}匹")
print(f"亀の数: {num_turtles}匹")
このプログラムを実行すると、以下の結果が表示されます。
<正解> 鶴の数: 15 匹、亀の数: 5 匹
```

図1 つるかめ算の生成されたプログラムと正解

②「ループリック評価」

1つ目は、学校教育で必要なスキル「プログラミング基礎」です。ループリック評価があります。概念設計とプロンプト処理を2段階で行うような形になります。まず第1段階として、ループリック作成用概念設計シートを用いて、ループリック表を生成します。第2段階として、その結果生成されたループリック表を用いて、実際に児童生徒の作文を評価するという形になります。

具体例を見てみましょう。概念設計シートで「人物：公立小学校で教員。内容：

1) 小学5年生の生徒たちを書くことの指導をするために、2) ループリックを作成したい。形式や語調：1) 小学5年生でもわかるような表現、2) ループリックを表形式で。制約：1) 単元名は「『文化を受けつぐ』ことについて考えよう」、2) 評価内容は『小学校学習指導要領（平成29年告示）』における、該当学年の「書くこと」の指導事項を踏まえて作成。その他：無し」と記載します。

これからプロンプトが生成されます。生成AIの「Gemini」が「横浜の公立小学校で教員をされているのですね。5年生の「書くこと」の指導で、「文化を受けつぐ」をテーマにしたループリック作成、素晴らしいです！学習指導要領を踏まえつつ、5年生にも分かりやすい表現で表形式のループリックを作成してみました。」と言って、図2のループリック表件式で作成してくれます。

評価の観点	よくできている (A)	まあまあできている (B)	もう少し頑張ろう (C)
書く内容	伝えたい文化について、具体的な例を挙げて説明している。その文化を受けつぐことの大切さや、受けつぐためにできることについて、自分の考えを理由をつけて分かりやすく書いている。読み手が興味を持つような工夫（例：問いかけ、部分がある。印象的な言葉）が見られる。	伝えたい文化について説明しているが、説明が簡単になっている。文化を受けつぐことの大切さや、受けつぐためにできることについて、自分の考えを書いてある。理由の説明が少し足りない部分がある。	伝えたい文化について書こうとしているが、説明が簡単になっている。文化を受けつぐことの大切さや、受けつぐためにできることについての自分の考えが書けていない部分がある。
構成	始め・中・終わりの構成ははっきりしていて、それぞれの部分で書くことがまとまっている。考えの流がスムーズで、読みやすい文章になっている。	始め・中・終わりの構成はできているが、それぞれの部分で書くことが十分にまとまっていないところがある。考えの流が分かりにくい部分がある。	始め・中・終わりの構成が意識できていない。書くことの順序がばらばらで、考えの流が分かりにくい。
表現	読んだ人が情景を思い浮かべることができるように、言葉を選んで詳しく書いている。自分の気持ちや考えを表す言葉が豊かで、適切に使われている。句読点や段落分けなどが正しく使われていて、読みやすい工夫がされている。	読んだ人が内容を理解できるように、言葉を選んで書いている。自分の気持ちや考えを表す言葉を使っている。句読点や段落分けなどの使い方がおおむね正しい。	使う言葉が簡単で、説明が不足しているところがある。自分の気持ちや考えを表す言葉が少ない。句読点や段落分けなどの使い方が間違っているところがあり、読みにくい。

図2 概念設計シートから生成されたループリック表

次に第2段階の評価用の概念設計シートを見てみましょう。概念設計シートでは「人物：公立小学校で教員/小学校3年生。内容：1)以下の作品を今回作成したループリックで評価、2)点数もつけて、3)児童生徒が今後どのように学習をすればいいのか、自分自身で自己調整ができるようなアドバイス。制約：1)作成したループリックに従って、2)評価内容が、ループリック表上でわかるように、3)満点は10点、4)アドバイスの文章は児童生徒を励ますように」してくださいというような形です。児童生徒の作品は『昔からつづいている京都のちえ』（橋口諒太郎、京都市立御所南小学校3年）で、京都広告協会主催作文コンクールの小中学校部門で最優秀賞を取ったものです。この作品では「ぼくは、京都について学習して京都のみ力を知るまでは京都はふつうにある町だと思っていました。それから京都についてあまり考えたことがありませんでした」というような感じで始まります。

この文章を、ループリック評価してみると、総合評価は点数10点満点中8点です。評価のコメントは「『「文化を受けつぐ』というテーマで、京都の和菓子作り、竜王園、京町家、京料理、竹田つけものなど、具体的な事例を挙げて『ちえ』というキーワードで一貫して説明されており、非常に分かりやすいです。小学5年生の作品としては、内容の深掘りがよくできています」と。そして「構成、表現と目的」というような感じで評価を記し、コメントをつけています。

橋口諒太郎さんへのアドバイスは「橋口さん、こんにちは！『昔からつづいている京都のちえ』を読ませてもらいました。『京都はふつうの町だと思っていた』という最初の気持ちから、『ちえ』というキーワードを見つけて、たくさんのすごい発見をしたこと、本当に素晴らしいですね！」。ということで下の5行「次の学習では、橋口さんが見つけた『ちえ』が、具体的にどんなふうに「昔からつづいて」きたのか、もっと調べてみるのはどうでしょうか？例えば、昔の人はどんな道具を使って、どんな工夫をして『ちえ』を生み出してきたのか、絵や写真と一緒にまとめてみると、もっともっと『ちえ』のすごさが伝わる作品になると思いますよ。橋口さんの学びは、これからもっともっと広がっていくはずですよ。応援しています！」というような形になるわけです。

2) 統計学の基礎

2つ目は、学校教育で必要なスキル「統計学の基礎」です。文部科学省は「小学校では統計的に分析するための知識や技能を理解し、身近な生活の場面の問題を解決するためにデータを集めて表やグラフに表し、統計量を求めることで、現状や分布の傾向を把握したり、2つ以上の集団を比較したり意思決定をする。

統計的手法を用いて出された結果を、批判的に考察する」と。この「批判的に考察する」というのは小中高一貫して言われているわけです。

3) データ可視化と解釈力

3つ目は、学校教育で必要なスキル「データの可視化と解釈力」です。データの可視化として、解釈力のためには、ヒストグラム、円グラフ、二項分布グラフ、箱ひげ図、散布図や、地図上での可視化があります。テキストの可視化として、例えば「走れメロス」の冒頭の部分に関して、文字頻度はグラフ化すると棒グラフになります。文字頻度をワードクラウドにしてみると、よりデータがはっきりします。「メロス」という言葉が非常に大きく、次に大きいのは「村」ということが容易にわかります。重要なキーワードをワードクラウドでは、可視化できるというわけです。

さらに、主体的意思決定と市民教育の観点から、データを根拠に考える行動する力を育成すること、社会を支える市民性の涵養ということが大切です。そしてシステム思考を実践するということが大事だということがあります。

4. 倫理的課題

倫理的課題には、プライバシーの保護、データバイアスによるリスク及び、AIの透明性と説明責任があります。データバイアスによるリスクは、以前第6講にもバイアスとバリエーションという話がありました。バイアスは、データ不足によるリスクです。AIの透明性とは、AIの仕組みや判断過程を公開し信頼できるようにすることです。AIの説明責任とは、なぜそのような結果に至ったのかを説明できることです。

法的・社会的課題には、データ利用に関する法整備、社会制度との調和と、自動化による職業や労働環境の変化というものがあります。日本のAI戦略として最初の頃に出たものですが「データにはノイズや偏ったデータがあって、AIが信頼できる結果をできない可能性があります。AIを安全安心に社会実装するためには、信頼できる品質のデータによりAI製品・サービスの信頼性を担保する仕組みが必要である」と。例えばどのような問題が起こったかというと、人事採用や就職の採用の場合に、過去の採用者が男性が多く学習データ男性が多いと、男性がいいんだというようなバイアスデータを受け、女性の応募者を不採用にして、男性の方を採用したしまった。アメリカで起こったのは、黒人と白人の応募者があった場合に、白人は採用されるが、黒人を落とされることが多いということで、問題になったわけです。それで、学習データを増やさなきゃいけない。

女性のデータや有色人種のデータを増やすし、バランスを取っていくというようなことがありました。別件ですが、少しのノイズで AI が誤認識することがあります。「STOP」の標識文字にテープを貼っただけで、速度制限に読み間違えたという事例があります。

政策として、AI 製品・サービスの信頼を担保する仕組みとして、データ品質の指標とその測定方法の策定及び国際標準化と、第三者評価プロセスそれにガイドラインを整備していく必要があります。

教育現場での対応策として、まず倫理教育の強化があります。「こういうことはやっとならいいよ。こういうことは大丈夫」という線引きをきちんと抑えていくことが大切です。それから法的知識の習得、特に著作権などは法的な問題が結構あります。最後に、柔軟な思考と継続的な学習を育成することが必要です。

5. 未来の教育像

未来の教育としては、データ駆動型の教育プログラム、社会全体のデータリテラシー向上及び、データサイエンス教育が社会全体のデータリテラシー向上と持続可能な発展に貢献することを理解してもらうことが大事です。

データ駆動型の教育プログラムの例として、現在実証実験中の「作文・論文作成支援」の事例があります。これには、人間が作成する「概念設計シート」が一番大事です。

「作文・論文作成 AI エージェント1」は、作文や論文生成を概念設計シート1に基づいて行います。概念設計シート1には、具体的に「どんな実践研究か？内容の骨子を教える」などが書かれています。

続く、3種の「採点・査読 AI エージェント2」は、概念設計シート2に記載の「作成論分を採用して良いか」などに基づき評価します。採点・査読 AI エージェント2は、例えば新規性 20 点、妥当性とか信頼性 20 点等という評価結果を人間に知らせます。

最後の「研究倫理処理 AI エージェント3」は、概念設計で例えば「18 歳未満の児童生徒を対象にしている？ その場合の親の許諾書はある？」と書いておけば、AI エージェント3が研究倫理処理してくれます。他にも、教育上不適切な言葉が使われてないか？ とかをチェックします。この一例では、概念設計シートという形で、人間が与えたデータによって、AI エージェントが駆動し、データ駆動型の教育プログラムが実現するわけです。

6. まとめ

データサイエンスは、社会教育の両面で重要性を増してきます。このため、教育者は未来を見据えた教育を実践する必要があります。未来の教育は単なる知識の伝達ではなく、客観的なデータを活用した問題解決能力や倫理感を育むことが求められているわけです。

課 題

- ①データサイエンスの未来において、社会や産業界で期待される役割と、その教育的意義について述べなさい。
- ②今後の教育現場において、データリテラシー教育を推進するために必要な取り組みや内容について具体的に述べなさい。
- ③データサイエンスの発展に伴う倫理的・社会的課題を挙げ、それに対して教育現場でどのような対策や教育内容を取り入れるべきか、あなたの考えを述べなさい。。

地域の文化資源を守り，知識基盤社会を支える人材の育成



令和7年度 岐阜県私立大学地方創生推進事業
「DXで実現する地域のデジタル人材育成事業」



学校 DX 戦略コーディネータ概論【Ⅵ】

発行年月日 令和8年2月

編集 久世 均（デジタルアーカイブ研究所所長）
村瀬 康一郎（岐阜女子大学・教授）
齋藤 陽子（岐阜女子大学・准教授）

監修 岐阜女子大学 デジタルアーカイブ研究所
〒500-8813
岐阜県岐阜市明德町 10 番地 杉山ビル 4 階
岐阜女子大学 文化情報研究センター
TEL 058-267-5237 FAX 058-267-5238

発行 岐阜女子大学 デジタルアーカイブ研究所