IT資格道場

統計検定データサイエンス発展 学習用テキスト

予備知識ゼロから合格ラインへ —— 読んで理解する1冊目

無料サンプル(第1章まで)

SAMPLE

統計検定 公式「統計検定 データサイエンス発展 出題範囲表」(公式ページ表記202501版)準拠
2026年11月版

ご利用にあたって

- このテキストは「IT資格道場」(www.shikaku-doujou.com)の購入者限定テキストのうち、学習用テキストの第1章までを収録した**無料サンプル**です。どなたでもご覧いただけます。
- 無料サンプルも著作物です。転載・改変しての再配布・二次販売はできません。
- AI モデルの学習データとしてのご利用(学習目的の入力・データセット化を含む)はできません。
- 製品版(学習用テキスト・頻出度別ノート・暗記用ノートの3点)は、ご購入者を識別できる透かし入りのPDFとして提供されます。
- 本PDFはスマートフォン・タブレットでの閲覧に合わせたA5版面・文字サイズで組版しています。

SAMPLE

目次

読み始める前に：この教材の使い方

- 試験の基本情報
- 章の構成
- 学び方

第1章 社会におけるデータ・AI 利活用 [1.1/1.2/1.3/1.4/1.5/1.6]

- 社会で起きている変化 [1.1]
- 社会で活用されているデータ [1.2]
- データ・AI の活用領域 [1.3]
- データ・AI 利活用のための技術 [1.4]
- データ・AI 利活用の現場 [1.5]
- データ・AI 利活用の最新動向 [1.6]

サンプルはここまでです

SAMPLE

3 冊の使い方は、ダウンロード欄の案内を参照してください。

SAMPLE

読み始める前に：この教材の使い方

このテキストは、一般財団法人 統計質保証推進協会が実施する **統計検定 データサイエンス発展** に合格することを目的に作られています。

データサイエンス発展が問うのは、データ・AI の活用を社会の中で捉える知識、データを扱うための数理とプログラミング、確率・統計的推測、種々のデータ解析を、ひとつながりの力として使えるかどうかです。公式の出題範囲表は 13 の大項目で、第1章〜第13章はその順に並ぶ。

前半は用語と区別の章、後半は計算と読み取りの章。数式は平文で書き、数値例は実際に計算して確かめています。式が出てきたら、手を動かして数値例をなぞってください。

試験の基本情報

項目	内容
試験名	**統計検定 データサイエンス発展**
運営	一般財団法人 統計質保証推進協会（日本統計学会 統計検定センター）
出題範囲	公式「統計検定 データサイエンス発展 出題範囲表」（202501版）
受験方法	CBT
問題数	**30問程度**
出題形式	**多肢選択・数値入力**
試験時間	**60分**

SAMPLE

項目	内容
合格水準	**100点満点中 60点以上**
受験料	**一般 6,000円／学割 4,000円（税込）**
前提資格	なし

3級・4級・データサイエンス基礎の範囲表の項目も、データサイエンス発展で出題される。該当する用語は、各章の中で前提から説明している。

章の構成

第1章〜第13章の題は、公式の出題範囲表の大項目名そのままです。

1. 第1章 社会におけるデータ・AI 利活用
2. 第2章 データ・AI 利活用における留意事項
3. 第3章 データリテラシー
4. 第4章 数理基礎
5. 第5章 デジタル情報とコンピュータの仕組み
6. 第6章 アルゴリズム基礎
7. 第7章 データ構造とプログラミング基礎（主に Python）
8. 第8章 データハンドリング
9. 第9章 データ取得とオープンデータ
10. 第10章 確率と確率分布
11. 第11章 統計的推測
12. 第12章 種々のデータ解析
13. 第13章 データ活用実践

SAMPLE

学び方

- 第1章・第2章：用語の定義と、隣り合う用語の違いを押さえる。場面を読んで当てはまる用語を選べる状態が目標
- 第3章〜第9章：データの扱い、数理、コンピュータの仕組み、アルゴリズム、Python のコードの読み方。数値例は最後まで自分で計算する
- 第10章〜第13章：確率と確率分布、統計的推測、種々のデータ解析、データ活用実践。式に数字を入れ、結論の文までたどる
- 手が止まったら、問われているのが用語の区別・計算・結果の読み取りのどれかを先に決める。それだけで使う道具が絞れる

SAMPLE

第1章 社会におけるデータ・AI 利活用 [1.1/1.2/1.3/1.4/1.5/1.6]

第1章は、用語と区別の章。計算はほとんどなく、「この場面で使われているのは何か」「この説明はどの用語のことか」を選ぶ問題が並びます。1つの用語を、定義、隣の用語との違い、取り違えやすい点の順に押さえると、場面の問題で迷わなくなります。

社会で起きている変化 [1.1]

ビッグデータ・IoT・AI・ロボットの意味と互いの関係

ビッグデータは、量、種類、発生の速さの3点で、従来の道具では扱いきれないデータを指します。つまり量だけでは決まりません。[1]英語の頭文字を取って、量（Volume）、種類（Variety）、速さ（Velocity）を3Vと呼びます。正確さ（Veracity）と価値（Value）を足して5Vとする言い方もあります。表計算ソフトで開けない巨大な表、文章や画像が混ざった雑多なデータ、秒単位で流れ込むログがその例です。

IoT（Internet of Things、モノのインターネット）とは、センサーや機器がネットワークにつながり、稼働や環境のデータを自動で送り続ける仕組みのこと。工場の機械、車、家電、農地の温湿度計が送るデータは人が入力しなくても溜まります。IoT はビッグデータの主な発生源です。

AI は、知的な判断を行う処理です。画像から物体を見分ける、文章の意味を取る、将来の需要を予測するといった判断を、データから学んだ規則で実行するものにほかなりません。ロボットは物理的に動く機械です。AI はソフトウェアの判断で、ロボットは体を持つ機械なので、両者は別物です。AI がカメラで部

SAMPLE

品を認識し、ロボットのアームが取り上げる場合は、認識、判断、動作が1つにつながって作業の自動化になります。

4つの関係は流れで覚えること。IoT が現実世界のデータを集め、ビッグデータとして蓄え、AI が分析して判断し、その結果をロボットや画面が現実に返します。この循環ができて初めて、データが価値を生みます。

次の数字で規模をつかみます。センサー1,000台が1秒に1回ずつ測定値を送ると、1日で1,000×86,400＝86,400,000件のデータ量。1件を100バイトとすると約8.64ギガバイトです。小さな装置が少数でも、増える速さは人手の入力とは桁が違います。

間違えやすい点が3つあります。第1に、IoT は「インターネットにつながったパソコン」を指す語ではありません。第2に、ビッグデータは量だけの話ではなく、種類と速さも含みます。第3に、AI とロボットを同じ意味で使うと、ソフトウェアと機械の区別を落とします。問題文の空欄に「通信機能を搭載したモノから取得されるデータ」とあれば、入るのは IoT です。通信規格では、5G が超高速、超低遅延、超大容量、超大量接続を特徴とします[2]。

データ量の増加・計算機の処理性能の向上・AI の非連続的進化

AI が急に実用になった背景には3つの条件が同時にそろったことがあります。データ量の増加、計算機の処理性能の向上、手法の進歩です。

データ量は、センサー、スマートフォン、Web の利用ログで急増しました。保存と処理の費用が下がったことも利活用を後押ししました。計算機では、GPU のように計算を多数同時に進める装置が、深層学習が必要とする大量の行列計算を速くしました。学習が現実的な時間で終わるようになったのです。

非連続的進化とは、性能が少しずつ上がるのではなく、ある時点で跳ぶように上がることです。代表例は2012年の画像認識コンテスト ILSVRC です。深層学

SAMPLE

習を使った AlexNet が、上位5つの予測に正解が入らなかった割合（top-5 誤り率）で15.3%を記録し、2位の26.2%を大きく下回りました[3]。それ以前は、人が設計した特徴量を使う手法が少しずつ誤り率を下げていた時期でした。この転機以後、画像認識、音声認識、翻訳で深層学習が主流になります。

3つの条件は、どれか1つでは足りません。手法だけあってもデータが無ければ学習できません。データがあっても計算機が遅ければ学習が終わりません。「深層学習の考え方は昔からあったのに2010年代に花開いた」という問題の答えは、データと計算力がそろったからです。

AI、機械学習、深層学習は入れ子の関係です。AI が最も広い言葉で、その中に、データから規則を学ぶ機械学習があり、さらにその中に、多層のニューラルネットワークを使う深層学習があります。包含の向きは「AI ⊃ 機械学習 ⊃ 深層学習」で、逆にはなりません。機械学習は学び方で3つに分かれます。

学び方	与えるもの	代表的な使い道
教師あり学習	入力と正解ラベルの組	予測（数値を当てる回帰、種類を当てる分類）
教師なし学習	正解の無いデータ	グルーピング（クラスタリング）、パターン発見
強化学習	行動の結果としての報酬	囲碁、ロボット制御

問題で「正解ラベルの有無」が示されたら、この表で学び方が決まります。ラベル付きの過去データから来月の売上を当てるなら教師あり、ラベルの無い購買履歴から似た顧客をまとめるなら教師なしです。ルールを人が書く方式は機械学習の外にあるので、ルールベースと呼んで区別する。

SAMPLE

第4次産業革命・Society 5.0・データ駆動型社会

第4次産業革命は、IoT、ビッグデータ、AI、ロボットによる産業の変革です。区分は次のとおりです。

- 第1次: 蒸気機関による機械化
- 第2次: 内燃機関や電力による大量生産
- 第3次: コンピュータによる生産の自動化
- 第4次: デジタル技術と IoT、AI による変革

総務省の情報通信白書は、この区分を世界経済フォーラムの整理を踏まえたものとして説明しています。ドイツの「インダストリー4.0」も第4次産業革命を指す名称です[4]。

Society 5.0 は、内閣府が示した日本の未来社会像です。狩猟社会（Society 1.0）、農耕社会（2.0）、工業社会（3.0）、情報社会（4.0）に続く社会で、「サイバー空間とフィジカル空間を高度に融合させたシステムにより、経済発展と社会的課題の解決を両立する人間中心の社会」と定義されています[5]。数字は産業革命の次数と社会の段階で別の系列です。第4次産業革命と情報社会（Society 4.0）は、どちらも「4」ですが同じものではありません。

2つの系列を並べると取り違えを防げます。

数字	産業革命（技術の変革）	Society（社会の段階）
1	蒸気機関による機械化	狩猟社会
2	電力による大量生産	農耕社会
3	コンピュータによる自動化	工業社会
4	IoT・AI による変革	情報社会

SAMPLE

数字	産業革命（技術の変革）	Society（社会の段階）
5	（産業革命の数え方に無い）	サイバーとフィジカルを融合する社会

産業革命は第4次までで、Society は5.0までです。表の同じ行は対応していません。たとえば「Society 2.0」は農耕社会であり、電力の時代ではありません。工業社会（3.0）が、蒸気機関や電力の産業革命の時期にあたります。

用語を当てる問題では、決まった手がかりの語が出ます。「サイバー空間とフィジカル空間の高度な融合」「人間中心」は Society 5.0 です。「蒸気機関による機械化」は第1次産業革命、「電力による大量生産」は第2次、「コンピュータによる自動化」は第3次。「IoT・ビッグデータ・AI による産業の変革」は第4次産業革命で、ドイツではインダストリー4.0と呼ばれます。「データを集めて分析し、結果を現実に返す循環で動く社会」はデータ駆動型社会です。

場面の問題では、まず取り違えの種類を見分けます。「Society 4.0 は第4次産業革命のことだ」という選択肢は誤り。前者は社会の段階（情報社会）、後者は産業の技術変革で、同じ数字でも別の系列です。「Society 5.0 は産業のための概念だ」も誤りで、経済発展と社会的課題の解決の両立、人間中心が定義に入っています。「データ駆動型社会ではデータを人が集めて人が分析する」も誤りです。集めるのは IoT などの仕組み、分析は AI などの仕組みが担い、人は目的と価値判断を持つ。

データ駆動型社会は、現実のデータを集め、サイバー空間で分析し、結果を現実に返して価値を生む循環で動く社会です。情報社会（Society 4.0）は、人が情報を探して分析していました。Society 5.0 は、IoT で集めたデータを AI が分析し、その結果が自動で現実に反映される点が違います。

SAMPLE

間違えやすい点です。「サイバー空間のデータを人が分析する」だけならSociety 4.0 の姿です。現実から集め、サイバー空間で処理し、現実へ返す双方向の循環こそ Society 5.0 の特徴。経済発展だけでなく社会的課題の解決を両立する点も定義に含まれます。

複数技術を組み合わせた AI サービスと、人間の知的活動と AI の関係

AI サービスの多くは、複数の技術の組み合わせです。スマートスピーカーは、音声認識で声を文字にし、言語理解で意図を取り、検索や知識で答えを作り、音声合成で話し返します。自動運転は、カメラ画像の認識、位置の把握、経路の判断、車体の制御の組合せ。問題で「このサービスに使われている技術はどれか」と問われたら、入力から出力までの流れを順にたどって、各段で必要な技術を拾います。

人間の知的活動と AI の関係は、役割分担で捉えます。AI が得意なのは、大量データからのパターン抽出、疲れずに高速で繰り返す処理、多数の候補の比較です。人が持つのは、目的の設定、何を良しとするかの価値判断、結果への責任であり、AI には任せきれません。AI は人の判断を支援し、一部の作業を代替します。判断の責任ごと代替するわけではありません。

間違えやすい点は、AI が「理解している」ように見えても、与えられた目的の中で最適な出力を作っているにすぎないことです。目的そのものを問い直すのは人の仕事です。

社会で活用されているデータ [1.2]

データの出どころ（調査・実験・人の行動ログ・機械の稼働ログ）の見分け

データは出どころで4つに分けられます。[6]

SAMPLE

- 調査データ: アンケートや聞き取りで集めたデータ。標本の取り方と質問文の設計で質が決まります
- 実験データ: 条件を人が操作して集めたデータ。処理（薬の投与や広告の表示）の割り当てを無作為にすると、原因と結果の比較ができます
- 人の行動ログデータ: Web の閲覧、購買、位置情報など、人の行動が記録として自動で残ったデータ
- 機械の稼働ログデータ: 機械、設備、センサーが出す動作の記録（温度、回転数、エラー）。故障の予兆を見つける予知保全に使う

見分けの手がかりは、誰が何のために集めたかです。人に尋ねて集めたなら調査、条件を操作したなら実験、人が普通に行動した結果が溜まったなら行動ログ、機械が自動で出したなら稼働ログです。EC サイトの購買履歴は行動ログです。同じサイトで表示デザインを A と B に無作為に分けて売上を比べた場合は、実験データ。

間違えやすい点です。アンケート（調査データ）は、回答者が覚えている範囲や答えやすさで偏ります。行動ログは実際の行動を反映しますが、行動の理由は分かりません。実験データは、割り当てを無作為にしたときだけ因果の比較に使えます。条件を操作せずに溜まったデータの比較では、原因と結果を結論できません。

4つの出どころを場面で見分ける例です。「店舗のアプリでの閲覧と購入の記録」は人の行動ログ、「工場の旋盤が1秒ごとに出す回転数と温度」は機械の稼働ログ、「顧客満足度を郵送で尋ねた回答」は調査、「新しい画面を無作為に半数の利用者へだけ表示して購入率を比べた結果」は実験です。ログはどちらも自動で溜まるので、誰が出したか（人か機械か）で分けます。調査と実験は、研究者が自分で設計して集める点が共通ですが、尋ねるのが調査、条件を操作するのが実験。

SAMPLE

1 次データと 2 次データ・データのメタ化（メタデータ）

1 次データは、自分の目的のために自分で集めたデータです。2 次データは、他者が別の目的で集めた既存のデータで、公的統計や他社の公開データがそうです。

項目	1 次データ	2 次データ
集めた人	自分	他者
目的の合い方	自分の目的に合わせて設計できる	目的とずれることがある
費用と時間	かかる	安く速い
注意点	設計の質で結果が決まる	定義、調査時期、集計単位の違い

自分が行ったアンケートは1 次データで、その結果を別の研究者が使えば、その人にとって2 次データです。同じデータでも使う人次第で区分が変わる点に注意。

場面で見分ける例です。ある市が、自分たちで住民に配ったアンケートの結果を使って施策を考えます。これは市にとって1 次データです。同じ市が、国が公表している人口の統計を使うなら、2 次データ。後者は費用も時間も小さく済みますが、国の統計の調査時点、地域の区切り、用語の定義が、市の知りたい内容とずれていないかを確かめます。たとえば国の統計が5年に1回の調査で、市が知りたいのは今年の状況なら、時点がずれています。ずれたまま使うと、結論が現実とずれます。「集めた人」と「使う人」が同じかどうかで決まる。

2 次データを使うときの確認事項は、定義、期間、対象範囲、単位、集計の仕方です。たとえば「失業率」でも、調べた機関によって定義が異なれば比べられません。

SAMPLE

メタデータ（データのメタ化）は、データについてのデータです。作成日時、作成者、単位、列の定義、撮影位置、収集の方法などがその例です。写真ファイルなら撮影日時や位置情報がメタデータにあたります。メタデータがあれば、データの検索も再利用も正しい解釈も可能。単位の記載が無い表は、数字の意味が分かりません。

間違えやすい点は、メタデータが元のデータの「中身」とは別の情報であることです。写真の被写体は中身で、撮影日時は中身についての情報です。データを共有するときは、中身と一緒にメタデータも渡します。

構造化データと非構造化データ（テキスト・画像/動画・音声/音楽）

構造化データとは、行と列の表に収まり、列ごとに型が決まったデータのこと。リレーショナルデータベースの表や CSV ファイルがそうです。売上表の「日付、商品、数量、金額」は構造化データです。

非構造化データは、決まった表の形を持たないデータです。テキスト、画像、動画、音声、音楽が代表。そのままでは行と列に並べられないので、分析するには数値に変える処理が要ります。文章は単語に区切って数え、画像は画素の値の並びで表し、音声は波形を数値の列にします。JSON や XML のように、構造は持つが表ほど固定されないデータを半構造化データとして別に分ける流儀もあります。

見分ける手がかりは、集計の道具。平均や合計を表計算で出せるなら構造化、そのままでは出せずに言語処理や画像処理の前処理が要るなら非構造化です。

データ作成（ビッグデータとアノテーション・教師ラベル）

アノテーションは、データに正解のラベルや注釈を付ける作業です。画像の物体を枠で囲んで名前を付ける、文章に感情（肯定か否定か）を付ける、音声を

SAMPLE

文字に書き起こすといった作業が該当。付けた正解を教師ラベル、ラベル付きのデータを教師データと呼びます。教師あり学習は、この教師データから規則を学びます。

アノテーションは人手で行うことが多く、課題が2つあります。費用と、人によるばらつき。1万枚の画像に、1枚30秒で枠とラベルを付けると、10,000×30＝300,000秒で約83.3時間かかります。作業者が複数なら、同じ画像でも判断がずれることがあるので、付け方の基準を文書にして、複数人の結果を突き合わせます。

ビッグデータの量だけでは足りません。ラベルが無いデータは、教師あり学習には使えません。教師ラベルが間違っていれば、モデルも間違いを学びます。ラベルの質が、モデルの質の上限を決めます。

間違えやすい点です。アノテーションは「データを集めること」を意味しません。ラベルが不要な教師なし学習（グルーピングなど）にはアノテーションが要りません。

データ・AI の活用領域 [1.3]

活用領域の広がり（生産・消費・文化活動）と業務の流れ（研究開発〜サービス）

データ・AI の活用は、生産、消費、文化活動の3つの領域に広がっています。生産は製造や農業で、消費は購買とレコメンド（おすすめ表示）で、文化活動は作曲や画像の生成のような創作で使われます。

企業の業務の流れに沿うと、活用の場面は次の順に並びます。

業務	活用の例
研究開発	新しい材料や薬の候補の探索

SAMPLE

業務	活用の例
調達	需要予測にもとづく発注量の決定
製造	外観検査、機械の故障の予兆検知（予知保全）
物流	配送経路の最適化、倉庫内作業の自動化
販売・マーケティング	顧客の分析、広告の出し分け、価格の調整
サービス	チャットボットによる問い合わせ対応

場面から業務を当てるには、その活動が業務の流れのどこにあるかを見る。「需要を予測して仕入れ量を決める」は調達、「製品の傷を画像で検出する」は製造、「トラックの配送順を最短にする」は物流です。複数の段にまたがる例もあります。需要予測は調達にも製造計画にも使えます。問題が「最も適切」を問うなら、文中の目的語（発注、検査、配送、顧客）に注目して決めればよい。

活用の目的（仮説検証・知識発見・原因究明・計画策定・判断支援・活動代替・新規生成）の見分け

活用の目的は7つに分けられます。[6]この7つを取り違える問題が、この節の難所です。

目的	中身	例
仮説検証	先に立てた仮説をデータで確かめる	「雨の日は来店が減る」を過去の来店数で確かめる
知識発見	データから未知のパターンや規則を見つける	購買履歴から、意外な商品の組合せを見つける
原因究明	問題の原因を突き止める	不良品が増えた原因を製造条件のデータから探す

SAMPLE

目的	中身	例
計画策定	予測をもとに計画を作る	需要予測から生産計画を作る
判断支援	人の判断の材料を出す	与信審査のスコアを審査担当者に示す
活動代替	人の作業を AI が置き換える	自動運転、定型事務の自動化（RPA）
新規生成	新しいコンテンツを作り出す	文章、画像、音楽の生成

見分けの軸は2つです。1つ目は、仮説が先かデータが先か。仮説を先に立てて確かめるなら仮説検証で、データから先に何かを見つけるなら知識発見（データマイニング）です。2つ目は、最終判断を誰がするかです。人が最終判断をして、AI は材料を出すだけなら判断支援です。人の作業そのものを AI が担うなら、活動代替になる。

場面で判断する例。クレジットカードの与信審査で、AI が申込者ごとに返済見込みのスコアを出し、審査担当者がそれを見て可否を決めています。この場合、スコアは人の判断の材料なので、目的は判断支援です。もし AI が審査の可否まで決め、人は関与しないなら、活動代替であって判断支援ではありません。

別の例です。ある工場で、先月から不良率が上がっています。担当者は「温度が高いと不良が増えるのではないか」と考え、温度と不良率のデータを調べる。仮説が先にあるので、これは仮説検証です。仮説を立てず、全ての測定項目から不良と関係の強い項目を機械に探させたなら、知識発見になります。「不良の原因は何か」を突き止めることが目的なら、原因究明です。同じデータ解析でも、目的の言い方で名前が変わる。

SAMPLE

目的を決める手がかりの語を表にします。

文中の手がかり	目的
「〜という仮説を確かめる」	仮説検証
「これまで知られていなかった傾向を見つける」	知識発見
「なぜ〜が起きたのかを調べる」	原因究明
「予測をもとに〜の計画を作る」	計画策定
「担当者が判断する際の材料として示す」	判断支援
「人に代わって自動で〜する」	活動代替
「新しい文章・画像を作る」	新規生成

2つの目的が並ぶ場面では、文の主たる動詞で決めます。「需要を予測して、その結果から生産計画を作る」は、予測は手段で、成果物が計画なので計画策定です。「異常を検知して、担当者に知らせる」は、人が確認して判断する前提なので判断支援。「異常を検知して、機械を自動で止める」は、人の操作を AI が置き換えているので活動代替です。

間違えやすい点は2つです。原因究明と知識発見は、どちらもデータから何かを探しますが、原因究明は「ある問題の原因」と目的がはっきりしています。計画策定と判断支援は、どちらも将来に向けた使い方ですが、計画策定は計画という成果物を作り、判断支援は人の判断に材料を渡す点が違う。

生成系の活用（対話・コンテンツ生成・翻訳・要約・執筆支援・コーディング支援）

生成 AI は、人の指示に応じて、文章、画像、音声、コードなどを新しく作り出す AI。[6]活用の型は次のとおりです。

SAMPLE

- 対話: チャットボット、質問への回答、相談の受け答え
- コンテンツ生成: 文章、画像、音楽、動画の作成
- 翻訳・要約: 文章を別の言語へ、または短くまとめる
- 執筆支援: 文章の下書き、言い換え、校正
- コーディング支援: コードの補完、説明、変換、テストの生成

生成系は、新規生成の目的に使えるだけでなく、判断支援（質問への回答の材料）にも使えます。使い方に注意が必要な点があります。出力はもっともらしいが誤っていることがあります。事実でない内容を事実のように出力する現象は「ハルシネーション」と呼ばれる。結果は人が確かめて使います。また、入力した情報が外部に渡ることがあるので、個人情報や社外秘を入れるときは利用条件を確認します。

間違えやすい点です。翻訳と要約は、既にある文章を別の形に変える作業で、新規生成の中では言い換えの生成にあたる。一方で、検索は既存の情報を探す作業で、生成ではありません。生成 AI の出力が元データの丸写しとは限らず、学習したデータの特徴から新しく作られる点が、検索との違いです。

データ・AI 利活用のための技術 [1.4]

データ解析の型（予測・グルーピング・パターン発見・最適化・シミュレーション・データ同化）

データ解析は6つの型に分けられます。[6]

- 予測: 過去のデータから、将来や未知の値を推定する。売上予測、需要予測、故障予測。回帰、時系列、分類のモデルを使う
- グルーピング: 似たもの同士をまとめる。顧客のセグメント分け。正解ラベルが無い教師なし学習（クラスタリング）

SAMPLE

- パターン発見: データの中に繰り返し現れる規則を見つける。併買の関係（アソシエーション分析）、異常検知
- 最適化: 制約のもとで目的を最大か最小にする解を求める。配送経路、在庫量、人員配置
- シミュレーション: モデルで現象を模擬計算する
- データ同化: 観測データでシミュレーションの状態や母数を逐次補正する。気象予報が代表

予測は「正解がある」型で、グルーピングは「正解を使わない」型です。予測では、答えの分かっている過去のデータで学習し、答えの分からない未来を当てます。グルーピングでは、答えの無いデータを似ているもの同士に分けるだけで、分けた集団の名前は人が付けます。

最適化と予測は区別して覚える。予測は「どうなるか」を当てる型で、最適化は「どうすれば最良か」を求める型です。明日の来客数を当てるのは予測で、その来客数に対して人員をどう配置すれば費用が最小になるかを求めるのは最適化です。予測の結果を最適化の入力に使うことは多くあります。

シミュレーションとデータ同化の関係は、天気予報で考えると分かる。シミュレーションは、大気の物理法則をもとにした計算で、先の天気を計算します。ただし初期の状態が少しずれると、先で大きくずれます。観測データ（気温や気圧の実測値）を取り込んで計算の状態を修正し続ける手法がデータ同化です。シミュレーションだけでは観測とずれていき、観測だけでは未来が計算できません。両方を組み合わせるのがデータ同化です。

異常検知は、パターン発見の一種です。異常は数が少ないので、評価には適合率と再現率を使います。たとえば1,000件の中に異常が10件ある場合。システムが20件を異常と判定し、うち8件が本当の異常だったなら、適合率は8÷20＝0.4、再現率は8÷10＝0.8です。適合率は「異常と判定したもののうち本当

SAMPLE

に異常だった割合」、再現率は「本当の異常のうち見つけられた割合」です。誤報が多いと適合率が下がり、見逃しが多いと再現率が下がります。

異常検知を正解率だけで評価すると、判断を誤る。1,000件のうち異常が10件のデータで、すべてを正常と判定するシステムを作ります。正しく当たるのは正常の990件なので、正解率は990÷1,000＝0.99です。数字は高く見えますが、異常は1件も見つけておらず、再現率は0です。異常のように数の少ないものを探す問題は、適合率と再現率を見る。

場面から型を選ぶ例。ある小売店が、レシートの記録から「ビールを買う客は同時にチーズも買いやすい」という組合せを見つける。ラベルも仮説も無く、規則を見つけたのでパターン発見です。同じ店が、顧客を購買の傾向で5つの集団に分けたならグルーピングです。来月の売上を推定したなら予測です。

データ可視化の型（複合グラフ・2軸・多次元・関係性・地図上・挙動や軌跡・リアルタイム）の選び方

可視化は、何を見せたいかで型を選ぶ。[6]

見せたいこと	型	例
種類の違う量を同じ時間軸で	複合グラフ	売上（棒）と利益率（折れ線）を重ねる
単位の違う2つの量を1枚で	2軸グラフ	気温（左軸）と売上（右軸）
3つ以上の変数の関係	多次元の可視化	散布図行列、平行座標、ヒートマップ
つながりの構造	関係性の可視化	ネットワーク図（利用者どうしのつながり）

SAMPLE

見せたいこと	型	例
場所による違い	地図上の可視化	地域別の人口を色で塗り分ける（コロプレス図）、密度を示すヒートマップ
動きの経路	挙動・軌跡の可視化	人や車の移動経路を線で描く
今の状態の更新	リアルタイム可視化	稼働状況を自動で更新するダッシュボード

複合グラフは、棒と折れ線のように種類の違うグラフを重ねたものです。2 軸グラフは、左右に別の目盛りを持つグラフで、単位の違う量を1つの図で比べられます。2軸グラフは、左右の目盛りの取り方で相関が強く見えたり弱く見えたりします。目盛りの範囲を変えると、印象は大きく変わる。図が示す関係の強さは、図の見た目だけでは分かりません。

多次元の可視化は、3つ以上の変数を一度に見る手段です。散布図行列は変数の組ごとに散布図を並べ、平行座標は変数ごとに平行な軸を立てて、1件のデータを折れ線でつなぎます。ヒートマップは数値の大きさを色の濃さで表します。変数が多いと、紙や画面では次元が足りないので、見せ方の工夫が要る。

関係性の可視化は、点（人や商品）とそれを結ぶ線（つながり）で表すネットワーク図です。SNS の友人関係や取引の関係を見るのに向いています。地図上の可視化は、位置の情報があるデータを地図に載せます。人口を地域ごとの色で表すのが階級区分図（コロプレス図）。挙動・軌跡の可視化は、人や車の位置の時間変化を線で描きます。リアルタイム可視化は、新しいデータが届くたびに表示を更新し、状態をその場で見られるようにします。

SAMPLE

選び方の問題は、見せたい内容の語から型を決めます。「つながり」はネットワーク図、「場所による違い」は地図、「移動の経路」は軌跡、「今の状況」はリアルタイム。「単位の違う2つを1枚に」は2 軸グラフ、「3つ以上の変数」は多次元の可視化です。

非構造化データ処理（言語処理・画像/動画処理・音声/音楽処理）

非構造化データは、種類ごとに処理が違います。

- 言語処理: 文章を単語に区切り（形態素解析）、意味を捉え、翻訳、要約、感情の判定、質問への回答に使う
- 画像・動画処理: 画像の分類、物体の検出（どこに何があるか）、顔の認識、動画では動きの追跡
- 音声・音楽処理: 音声認識（声を文字に）、音声合成（文字を声に）、音楽の分類や生成

言語処理の入り口は、文章を数値に変えること。単語に区切って出現回数を数える、単語を数値のベクトルに置き換えるといった方法があります。画像は画素の値の並びで、音声は波形の数値で表されるので、数値に変える段階は既にあります。

見分け方は、入力の種類。入力が文章なら言語処理、写真や映像なら画像・動画処理、声や音なら音声・音楽処理です。出力が文字になる音声認識は、入力が音声なので音声処理に入ります。画像に写った文字を読み取る光学式文字認識（OCR）は、入力が画像なので画像処理です。

特化型 AI と汎用 AI・今の AI で出来ることと出来ないこと・AI とビッグデータ

特化型 AI とは、特定の課題だけを解く AI のこと。画像の分類、囲碁、翻訳がそうです。汎用 AI（AGI）は、人間のように幅広い課題を自分でこなす AI です。

SAMPLE

現在実用になっている AI は、すべて特化型です。囲碁で人に勝つ AI が、同じ仕組みのまま料理を作ったり会話したりできるわけではありません。

大規模言語モデルは、1つのモデルで多くの文章課題をこなすので汎用に見えますが、研究上の定義の汎用 AI（AGI）とは言えません。誤りが多く、得意な課題が文章に偏っているためです。

今の AI の得意と苦手は次のとおりです。

- 得意: 大量のデータからのパターンの認識と予測、疲れずに同じ処理を高速で繰り返すこと
- 苦手: 学習データに無い状況への対応、因果の理解（相関と因果の区別）、常識の判断、責任を持つ判断

学習データに偏りがあると、AI はその偏りを引き継ぎます。過去の採用データで学習した AI が、過去の偏った判断を再現するのがその例です。AI が間違える理由は、アルゴリズムの欠陥だけでなく、データの偏りにもあります。

AI とビッグデータの関係は相互の関係。深層学習には大量のデータが要り、ビッグデータは AI の分析で価値が出ます。データが多いほど精度が上がるとは限りません。データの質、偏り、ラベルの正しさが、精度を左右します。

場面で判断する例。ある会社が「画像から不良品を見つける AI」を作りました。精度が高いので、同じ AI に製造計画の立案も任せようとしています。この AI は不良の画像分類に特化したもので、計画の立案は学習していません。特化型 AI は、学習した課題の外では使えません。

「今の AI でできること」を問う選択肢は、次のように見分けます。「大量の画像から傷のある部品を高速に見つける」は得意な作業なので正しい記述です。「学習データに一度も現れなかった状況でも、人と同じ常識で対応できる」は、苦手なので誤りです。「データから見つけた相関を、そのまま原因として断定で

SAMPLE

きる」も誤りで、因果の理解は今の AI の苦手分野。「学習データの偏りは、AIの出力に影響しない」は誤りで、偏りは出力に引き継がれます。

認識技術・ルールベース・自動化技術

認識技術は、画像、音声、文字を識別する技術です。光学式文字認識（OCR）、音声認識、顔認識がそうです。自動化技術とは、定型的な作業を機械に任せる技術のこと。RPA（ロボティック・プロセス・オートメーション）は、パソコン上の定型作業をソフトウェアのロボットが代わりに行う仕組み。工場のロボットは、物理的な作業を自動化します。

ルールベースは、人が書いた「もし〜なら〜する」の規則で判断する方式です。機械学習は、データから規則を学ぶ方式です。

項目	ルールベース	機械学習
規則の出どころ	人が書く	データから学ぶ
説明しやすさ	判断の理由が追いやすい	理由が分かりにくいことがある
向く場面	規則が明確で数が少ない	規則が複雑、または書き出せない
弱点	規則の網羅が難しい	データが要る、偏りを引き継ぐ

迷惑メールの判定を例にする場合。「件名に特定の語を含むなら迷惑メール」と人が規則を書けばルールベースで、規則に漏れがあると見逃します。大量の迷惑メールと通常メールから特徴を学ばせると機械学習で、規則を人が書かなくても新しい手口に対応できることがあります。ただし、なぜその判定になったかは説明しにくくなります。

SAMPLE

間違えやすい点は、RPA が機械学習とは限らないこと。RPA は決められた手順を繰り返すのが基本で、学習はしません。AI と組み合わせると、画像の文字を読んで入力するといった、判断を含む作業まで自動化できます。

マルチモーダル・プロンプトエンジニアリング

マルチモーダルは、画像、音声、テキストなど複数の種類のデータを同時に扱う AI のことです。1つの種類だけを扱うのは単一モーダルです。写真を見せて「何が写っていますか」と文章で聞き、文章で答えが返ってくるのは、画像と文章の2つを扱うので、マルチモーダルにあたる。

プロンプトは、生成 AI に与える指示の文です。プロンプトエンジニアリングは、指示の書き方を工夫して、望む出力を得る技術です。主な工夫は次のとおりです。

- 役割の指定: 「あなたは経理の専門家です」と立場を示す
- 例示（few-shot）: 入力と望む出力の例をいくつか見せる。例を示さない指示はゼロショットと呼ぶ
- 手順を踏ませる指示: 段階を追って考えさせる（段階的な推論を促す）
- 出力形式の指定: 表、箇条書き、文字数、言語を指定する
- 条件や背景の明記: 目的、読み手、制約を書く

指示があいまいだと出力もあいまいになる点に注意。「要約して」だけでなく、「小学生向けに、3文で要約して」と書くと、望む出力に近づきます。例を見せる方法は、モデルを再学習しなくても出力の型を揃えられる点が特徴です。

間違えやすい点です。プロンプトエンジニアリングは、モデルの中身を学習し直す作業ではなく入力の工夫。モデルを追加の学習で調整するのは、転移学習の一種のファインチューニングで、別の手段です。

SAMPLE

データ・AI 利活用の現場 [1.5]

データサイエンスのサイクル（課題抽出と定式化〜課題解決に向けた提案）の段階と順序

データサイエンスの仕事は、6つの段階を順に進める循環です。[6]

1. 課題抽出と定式化
2. データの取得・管理・加工
3. 探索的データ解析
4. データ解析と推論
5. 結果の共有・伝達
6. 課題解決に向けた提案

課題抽出と定式化は、現場の困りごとを、データで答えられる問いに言い換える段階です。「売上を伸ばしたい」を、「来店客の何が購入額を左右するかを調べ、購入額を予測する」のように、目的の変数と使うデータ、評価の物差しがはっきりした形にします。この段階を飛ばして解析を始めると、立派な結果が出ても、現場の課題に答えていないことがあります。

データの取得・管理・加工は、必要なデータを集め、保管し、使える形に整える段階。欠測値（抜けている値）の扱い、異常な値の確認、形式の統一、複数のデータの結合を行います。

探索的データ解析（EDA）は、本格的な解析の前に、要約統計量（平均や分散）やグラフでデータを眺める段階です。分布の形、外れ値、欠測、変数どうしの関係を見て、仮説を立て、前処理の方針を決めます。ここは仮説を確かめる段階ではなく、データから手がかりを探る段階。

SAMPLE

データ解析と推論は、モデルを当てはめたり、検定や推定を行ったりして、結論を導く段階です。結果の共有・伝達は、図や言葉で、解析を知らない人にも分かる形で伝える段階です。課題解決に向けた提案は、結論を行動につなげ、現場に何をすべきかを示す段階です。

順序の問題は、前後関係の理由を押さえれば解ける問題。評価の物差しは、解析の前に決めます。解析してから「良さそう」な指標を選ぶと、結果に合わせた都合のよい評価になるためです。探索的データ解析は、データを集めて加工した後、モデルを作る前です。共有・伝達は、解析が終わってから。

場面で判断する例。あるチームが解析を終えて、結果の資料を作っています。作る途中で、使った変数に重大な欠測が偏っていると気づく。この場合、資料作りの段階から、データの取得・管理・加工の段階へ戻って、欠測の扱いを見直し、解析をやり直します。サイクルは一方向の直線ではなく、必要に応じて前の段階へ戻る循環です。戻る先は、問題が生じた段階です。

間違えやすい点。探索的データ解析は「結論を出す段階」ではありません。探索で見つけた関係は、仮説の候補にとどまる。同じデータで見つけて同じデータで確かめると、たまたまの関係を本物と誤ることがあります。確かめには別のデータを使うか、検定の設計を整えます。また、データの取得・管理・加工は、解析の下ごしらえなので手間が少ないように見えますが、実務では時間の多くを占めます。

作業を段階に振り分ける練習。「売上の減少を、データで答えられる形の問いにする」は課題抽出と定式化です。「複数の店舗の売上表を結合し、欠測を埋める」はデータの取得・管理・加工です。「売上の分布と外れ値をグラフで確認する」は探索的データ解析です。「回帰モデルを当てはめて、広告費の効果を推定する」はデータ解析と推論。「経営会議向けに、結果をグラフ中心の資料にまとめる」は結果の共有・伝達です。「この結果から、来月は広告費を増やす

SAMPLE

よう勧める」は課題解決に向けた提案です。作業の中身から段階を当てられれば、順序の問題も解けます。

流通・製造・金融・サービス・インフラ・公共・ヘルスケア等の利活用の事例

分野ごとの代表的な事例は、課題とデータと解析の型の組で覚える形。[6]

分野	事例	使うデータ	型
流通	需要予測と発注、POS データの分析	販売履歴、天気、催事	予測
製造	外観検査、予知保全	製品の画像、機械の稼働ログ	画像認識、異常検知
金融	不正取引の検知、与信スコア	取引履歴、属性	異常検知、予測
サービス	レコメンド、チャットボット	行動ログ、問い合わせ文	推薦、言語処理
インフラ	設備の劣化予測、電力の需給予測	センサーデータ、使用量	予測
公共	人流データの防災への活用	携帯電話の位置情報	可視化、予測
ヘルスケア	画像診断の支援、罹患の予測	医用画像、検査値	画像認識、予測

製造の予知保全は、機械の稼働ログ（振動や温度）から故障の予兆をつかみ、故障の前に保守する手法。壊れてから直す事後保全、決まった周期で点検する定期保全と比べて、無駄な点検と突然の停止の両方を減らせます。

金融の不正取引検知は、普段と違う取引を見つける異常検知です。不正は数が少ないので、見逃しと誤報のバランスが課題になります。ヘルスケアの画像

SAMPLE

診断の支援は、AI が病変の候補を示し、診断は医師が下す判断支援の例。

分野と事例を結ぶ問題では、分野の特徴と使うデータの関係から選びます。振動のデータなら機械の保全、位置情報なら人流、購買履歴ならおすすめの表示です。

データ・AI 利活用の最新動向 [1.6]

新しいビジネスモデル（シェアリングエコノミー・商品のレコメンデーション）

シェアリングエコノミーは、個人や企業が持つ遊休資産（車、部屋、スキル）を、プラットフォームを通じて貸し借りする経済です。配車サービス、民泊、フリーマーケットがその例。需要と供給のマッチング、価格の決定、利用者の評価にデータと AI が使われます。持たない人が使い、持つ人が収益を得るので、資産の稼働率が上がります。

レコメンデーションは、利用者に合う商品やコンテンツを推薦する仕組みです。主な方式は2つ。

- 協調フィルタリング: 似た利用者や似た商品の行動をもとに推薦する。「あなたと好みが似た人はこれも買っています」
- 内容ベース: 商品の属性（ジャンル、出演者、素材）の類似をもとに推薦する

協調フィルタリングは、商品の中身を知らなくても、行動の履歴だけで推薦できます。商品の属性を使う内容ベースは、新しい商品でも属性が分かれば推薦できます。協調フィルタリングは、行動の履歴が無い新しい利用者や新商品に推薦しにくい弱点があり、これをコールドスタート問題と呼びます。

SAMPLE

類似度の計算例。3つの商品 a、b、c への利用者の評価（5段階）が、利用者 A は（5, 4, 1）、利用者 B は（4, 5, 2）、利用者 C は（1, 2, 5）だとします。評価を数の並びとして、2つの並びの向きの近さをコサイン類似度で測ると、A と B は約0.966、A と C は約0.507です。A に近いのは B なので、B が高く評価して A がまだ見ていない商品を、A に推薦します。1に近いほど似ていて、0に近いほど似ていません。この計算は普通電卓の範囲を超えるので、試験で問われるのは「似た利用者の行動で推薦する」という考え方のほうです。

間違えやすい点です。協調フィルタリングは、商品の属性を使いません。属性で似たものを探すのは内容ベース。また、推薦はおすすめの偏りを生みやすく、似た情報ばかりに囲まれる状態（フィルターバブル）の指摘があります。

深層生成モデル・敵対的生成ネットワーク（GAN）・生成 AI

深層生成モデルは、ニューラルネットワークでデータの分布を学び、そこから新しいデータを作り出すモデルです。代表は、変分オートエンコーダ（VAE）、敵対的生成ネットワーク（GAN）、拡散モデル、自己回帰モデルです。

GAN は、2つのネットワークを競わせて学習する方式。生成器は本物らしい偽物を作り、識別器は本物か偽物かを見分けます。生成器は識別器をだまそうと偽物の質を上げ、識別器はだまされまいと見分ける力を上げます。この競い合いで、生成器は本物そっくりのデータを作れるようになります。原論文は、識別器が、データが学習データから来たか生成器から来たかの確率を推定するとして、この枠組みを説明しています[7]。GAN の悪用として問題になっているのが、実在の人物の顔や声を精巧に作るディープフェイク。

生成 AI は、文章、画像、音声、コードなどの新しいコンテンツを生成する AI の総称です。深層生成モデルは、その中核の技術です。「生成 AI」は応用の呼び名、「深層生成モデル」は仕組みの呼び名で、呼び名の軸が違います。

SAMPLE

場面で判断する例。「偽札を作る人と、偽札を見破る人が競い合って、どちらも腕を上げる」という説明がある技術はどれかと問われたら、GAN です。生成器が偽札を作る人、識別器が見破る人にあたります。競わせる構造があるかどうかが、他の生成モデルとの区別の決め手です。

生成モデルの区別を表にまとめる。

モデル	生成の仕組み	覚える手がかり
VAE（変分オートエンコーダ）	データを少数の値に圧縮し、そこから復元して作る	圧縮と復元
GAN	生成器と識別器を競わせる	偽物を作る側と見破る側の競い合い
拡散モデル	ノイズを少しずつ取り除いて作る	ノイズから画像へ
自己回帰モデル（LLM など）	次の要素を1つずつ予測して並べる	次の語の予測

説明文から選ぶ問題は、手がかりの語で決まります。「競い合い」は GAN、「ノイズを取り除く」は拡散モデル、「次の語を予測する」は大規模言語モデルです。

間違えやすい点は、GAN の識別器の役割。識別器は完成した作品を評価して終わりではなく、学習の途中で生成器に「どこが偽物っぽいか」の手がかりを与え続けます。学習がうまく進むと、識別器は本物と偽物を見分けられなくなり、正解率は偶然と同じ0.5に近づきます。

強化学習・転移学習

強化学習は、エージェント（行動する主体）が環境の中で行動し、得た報酬の合計を最大にする行動の決め方（方策）を試行錯誤で学ぶ方法です。正解のラ

SAMPLE

ベルは与えられません。与えられるのは、行動の結果としての報酬だけです。囲碁、ロボットの制御、ゲームの操作が代表的な応用です。画面の画素だけを入力にして、家庭用ゲーム機のゲームの操作を強化学習で身につけた研究もあります[8]。

報酬は、直後の報酬だけでなく、先まで含めた合計で見ます。行動を4回とったときの報酬が順に (1, 0, 5, −2) なら、合計は4です。遠い先の報酬ほど割り引いて数える流儀では、割引率を0.9として、$1 + 0.9 \times 0 + 0.9^2 \times 5 + 0.9^3 \times (-2) = 3.592$です。目先の小さな報酬を取らずに先の大きな報酬を狙う判断が学べる点が、強化学習の特徴。

教師あり学習との違いは、正解を教えるか報酬で評価するか。教師あり学習は「この入力にはこの答え」と正解を与え、強化学習は「この行動は良かった」と結果で評価します。

転移学習は、ある課題で学習済みのモデルを、別の関連する課題に流用し、少ないデータで追加学習する方法です。たとえば、大量の一般画像で学習した画像認識モデルを、数百枚の医用画像の分類へ応用する場合。ファインチューニングは転移学習の一つで、学習済みモデルの重みを新しい課題のデータで微調整します。

転移学習が効く場面は、手元のデータが少ない課題で、似た課題の大規模なデータで学習済みのモデルがあるときです。ゼロから学習するより、少ないデータと短い時間で性能が出ます。

場面で判断する例。あるチームが、自社の少量の製品画像で傷の判定モデルを作ります。ゼロから学習すると精度が出ず、ネット上の大量の一般画像で学習済みのモデルを借りて、自社の画像で追加学習したら精度が上がりました。この手法が転移学習。もし、正解ラベルを使わず、行動の報酬で学ぶ構造なら強化学習です。

SAMPLE

間違えやすい点です。強化学習は、「ラベルの無いデータをまとめる」教師なし学習とは別物。報酬という評価の信号が与えられる点が違います。転移学習は、同じモデルを複数の課題で同時に使う方法ではなく、一度学習したものを別の課題へ移す方法です。

大規模言語モデル・基盤モデル・拡散モデル

大規模言語モデル（LLM）は、大量のテキストで、次に来る語（トークン）を予測するように学習した巨大なモデルです。多くは、2017年に提案されたTransformerという構造を使っている。Transformerは、系列の処理で再帰や畳み込みを使わず、注意機構（どの語に注目するかを重みづけする仕組み）だけで組み立てられています[9]。文章の生成、要約、翻訳、対話を1つのモデルでこなします。

基盤モデルは、大量で幅広いデータで事前に学習し、追加の学習やプロンプトで、多くの下流の課題に適応できるモデルの総称です。この呼び名は、スタンフォード大学の研究者らの報告で広まりました[10]。大規模言語モデルは、基盤モデルの代表例です。基盤モデルのほうが広い概念で、画像や音声を扱うものも含みます。

学習の流れは2段階です。まず、大量で幅広いデータで事前学習を行い、言語の使い方や画像の一般的な特徴を身につける。次に、特定の用途に合わせて、少量のデータで追加の学習（ファインチューニング）をするか、プロンプトで指示して使います。事前学習のあとに追加の学習をする点は、転移学習と同じ考え方です。

拡散モデルは、データに少しずつノイズを加えていく過程を逆にたどり、ノイズから画像などを生成するよう学ぶモデルです。ノイズだらけの画像から少しずつノイズを取り除き、きれいな画像を作る流れ。この方式で質の高い画像を生

SAMPLE

成できることが示されました[11]。文章の指示から画像を作る画像生成 AI でも使われています。

3つの関係と違いは次の表のとおりです。

用語	何を指すか	主な用途
大規模言語モデル	文章を扱う巨大なモデル	文章生成、要約、対話
基盤モデル	幅広い課題に適応できる事前学習済みモデルの総称	文章、画像、音声の各課題
拡散モデル	ノイズから生成する方式のモデル	画像、音声の生成

大規模言語モデルは「何を扱うか」（文章）、拡散モデルは「どう生成するか」（ノイズを取り除く）、基盤モデルは「どう使うか」（幅広い課題に適応）を軸にした呼び名です。軸が違うので、1つのモデルが複数の呼び名にあたることがあります。

場面で判断する例。「ノイズだけの画像から、段階を追ってノイズを取り除き、指示文に合う画像を作る」という説明なら拡散モデル。「大量の文章で次の語を予測するよう学習した」という説明なら大規模言語モデルです。「特定の課題でなく幅広い課題に、追加の学習や指示で対応できる」という説明は基盤モデルです。

間違えやすい点は、大規模言語モデルが「事実を知っている」わけではない点。確率の高い語をつなげて文を作っているので、もっともらしい誤りを出すことがあります。重要な内容は、人が確かめて使います。また、拡散モデルは文章を生成するモデルではなく、おもに画像や音声の生成で使われます。

SAMPLE

サンプルはここまでです

続き(第2章〜巻末)は、統計検定データサイエンス発展 問題集のご購入で、頻出度別ノート・暗記用ノートと合わせて3点すべてダウンロードできます。

Footnotes

1. 同じ範囲表の 1.1「社会で起きている変化」の項目欄（ビッグデータ、IoT、AI、ロボット、AI の非連続的進化、人間の知的活動と AI の関係性）。https://www.toukei-kentei.jp/hubfs/files/grade_range/grade12_hani_20240830.pdf ↩

2. 統計検定 データサイエンス発展 公式サンプル問題の問1（IoT と、超高速・超低遅延・超大容量・超大量接続を実現できる 5G）。https://www.toukei-kentei.jp/hubfs/files/grade/grade12_sample_exercises.pdf ↩

3. Krizhevsky, Sutskever, Hinton「ImageNet Classification with Deep Convolutional Neural Networks」NeurIPS 2012、要旨と本文の結果の表。https://papers.nips.cc/paper_files/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html ↩

4. 総務省 平成30年版 情報通信白書「インダストリー4.0とは」、平成29年版「第4次産業革命を巡る世界的な動き」。https://www.soumu.go.jp/johotsusintokei/whitepaper/ja/h29/html/nc131100.html ↩

5. 内閣府「Society 5.0」。https://www8.cao.go.jp/cstp/society5_0/ ↩

SAMPLE

6. 統計検定 データサイエンス発展 出題範囲表（2024.8.30・202501版）、大項目「社会におけるデータ・AI 利活用」の小項目 1.2〜1.5 の項目欄。https://www.toukei-kentei.jp/hubfs/files/grade_range/grade12_hani_20240830.pdf ↩ ↩2 ↩3 ↩4 ↩5 ↩6 ↩7

7. Goodfellow ほか「Generative Adversarial Networks」arXiv:1406.2661、要旨。https://arxiv.org/abs/1406.2661 ↩

8. Mnih ほか「Playing Atari with Deep Reinforcement Learning」arXiv:1312.5602、要旨。https://arxiv.org/abs/1312.5602 ↩

9. Vaswani ほか「Attention Is All You Need」arXiv:1706.03762、要旨。https://arxiv.org/abs/1706.03762 ↩

10. Bommasani ほか「On the Opportunities and Risks of Foundation Models」arXiv:2108.07258、要旨。https://arxiv.org/abs/2108.07258 ↩

11. Ho ほか「Denoising Diffusion Probabilistic Models」arXiv:2006.11239、要旨。https://arxiv.org/abs/2006.11239 ↩

SAMPLE

統計検定データサイエンス発展 学習用テキスト(無料サンプル)

版

2026年11月版(統計検定 公式「統計検定 データサイエンス発展 出題範囲表」(公式ページ表記 202501版)準拠)

発行

IT資格道場(https://www.shikaku-doujou.com)

お問い合わせ

info@shikaku-doujou.com

「統計検定」は一般財団法人 統計質保証推進協会の登録商標です。記載されているその他の会社名・製品名・サービス名は、各社の商標または登録商標です。

本教材は公式シラバスの内容を基に作成したもので、一般財団法人 統計質保証推進協会および日本統計学会との提携・承認関係はありません。

本文・図表・設問はすべてオリジナルの著作物です。無断での転載・複製・再配布・二次販売はできません。お気づきの点は info@shikaku-doujou.com までご連絡ください。

SAMPLE