

Table of Contents

Introduction	1.1
2.1 データリテラシーとは	1.2
2.2 データの種類を知る	1.3
2.3 データの適正を判断する	1.4
2.4 データの可視化	1.5
2.5 データの可視化	1.6
3.1 データサイエンスにまつわる基本概念	1.7
3.2 機械学習の各分野の特徴	1.8
3.3 データ処理と特微量エンジニアリング	1.9
3.4 機械学習を応用する上で重要な概念	1.10
3.5 アソシエーション分析	1.11
3.6 線形回帰	1.12
3.7 ロジスティック回帰	1.13
3.8 サポートベクトルマシン(SVM)	1.14
3.9 決定木	1.15
3.10 K近傍法 (KNN: K Nearest Neighbor)	1.16
3.11 ナイーブベイズ	1.17
3.12 ニューラルネットワーク	1.18
3.13 教師なし機械学習	1.19
3.14 モデルの精度評価	1.20
4.1 画像データ・動画データの性質と処理	1.21
4.2 画像分類と一般物体認識	1.22
4.3 音声処理	1.23
4.4 時系列分析	1.24
4.5 状態空間モデル	1.25
4.6 自然言語処理の基礎	1.26
4.7 大規模言語モデル	1.27
4.8 生成AI	1.28
4.9 機械学習における解釈性	1.29
5.1 データ収集	1.30
5.2 データ保管	1.31
5.3 データ加工	1.32
5.4 データのセキュリティ	1.33
5.5 データ分析	1.34
6.1 データ活用プロジェクトの進め方	1.35
6.2 現状の把握	1.36
6.3 課題の定義	1.37
6.4 契約の種別と開発の形式	1.38

Introduction

7.1 法律・倫理をなぜ学ぶ？	1.39
7.2 データ倫理	1.40
7.3 データ倫理に関するガイドライン	1.41
7.4 個人情報とプライバシー	1.42
7.5 公平性・説明責任・透明性	1.43
7.6 生成モデルが社会に与えるインパクト	1.44
7.7 言語モデルと生成AIに関する課題	1.45
7.8 AIによる個人の意識の操作	1.46
8.1 DS検定に出題される数学・統計学	1.47
8.2 統計学とデータ分析	1.48
8.3 記述統計学と推測統計学	1.49
8.4 基本統計量（平均値、中央値、最頻値）	1.50
8.5 データのばらつきを評価する指標	1.51
8.6 不偏推定量と自由度	1.52
8.7 2変数の関係	1.53
8.8 相関関係と因果関係	1.54
8.9 確率・確率変数・確率分布	1.55
8.10 代表的な確率分布の紹介	1.56
8.11 中心極限定理と標本平均の定理	1.57

DS検定学習メモ

2.1 データリテラシーとは

データリテラシーが包括する知識と技術

- 分析に必要なデータを洗い出す
- データを的確に理解するために、可視化し、探索と洞察を行う
- データに存在する問題を発見し、それを解決するためにデータを適切に処理する
- 使えるデータと使えないデータの判別をし、分析の手がかりとなる変数を認識する
- 適切な手法を用いてデータを分析し、インサイトを抽出する
- 分析の結果をわりやすく、正しく説明する

データ分析におけるデータの扱い方

データ分析の大まかな流れ

- データの読み込み★
- データの理解
- モデルの選択★
- 特徴量エンジニアリング★
- モデルの学習
- 精度評価

★を書いた箇所が特にデータリテラシーが重要になるステップ。

2.2 データの種類を知る

連続データと離散データ

- **連続データ** 連続的な数値で表されるデータ

例: 重量、血圧、気圧、気温、成長率、距離、時間

- **離散データ** 飛び飛びの数値で表されるデータ

例: サイコロの目、年齢、試験の得点、学年、販売個数

厳密な区別が難しいデータもある。「売上」は、離断的なデータだが、近似的に連続データとして暑かった方が便利なときがある。逆に、血圧が整数値でしか表されない場合、データは離散的とも言える。

量的データと質的データ

- **量的データ** 数値としての意味を持つ。四則演算を適用できる
- **質的データ** カテゴリや順序を表すデータ。見た目が数字であっても数量としての意味は持たず、ラベルとしての役割しかない。**カテゴリカルデータ**と呼ぶこともある。

尺度水準

- 質的データ
 - **名義尺度** ラベル、名称

例: ID、性別、血液型、電話番号、郵便番号

- **順序尺度** 順序や大小に意味があるが、間隔には意味がなく足し算引き算ができない

例: アンケートの評価値、ランキング、震度

- 量的データ
 - **間隔尺度** 等間隔のメモリに意味がある。差には意味があるが比率に意味はない。

例: 日付、年齢、時間、試験点数

- **比例尺度** 四則演算すべてが可能。原点を0とする

例: 速度、身長、価格、電力

- 尺度水準は**ダミー数値**。等しいか等しくないかのみ表せるので、頻度や最頻値までしか計算できない。
- 順序尺度だと中央値やパーセンタイルが計算できる。
- 間隔尺度になると和と差が計算できるため、平均、標準偏差、相関係数の計算が可能。
- 比例尺度は四則演算すべての計算が可能

順序尺度と間隔尺度の見分け方 等間隔かどうかを考える。震度は、震度4と震度3の差と震度3と震度2の差が同じではないため間隔尺度ではない。

間隔尺度と比例尺度の見分け方 年齢や試験点数は間隔尺度の意味で使われることがおおいが、用途によっては比率を計算することがある。

午後2時などの時刻は間隔尺度。差に意味はあるが午後2時の3倍は午後6時などの比率には意味がない。

また、比例尺度は、原点0に相対的ではなく絶対的な意味がある。午前0時や絶対的な0ではない。

構造化データと非構造化データ

- 構造化データ** 列と行の構造と概念を持つデータ
- 非構造化データ** 上記でない。画像、音声、テキスト、XMLやHTML

オープンデータ

- 無償で利用可能なデータセット
- 機械判読に適したデータ形式で配布されている
- 二次利用を許可している

データを集める方法

- 自社データ
- 調査やデータ販売企業からの購入
- クローリング（HTMLを収集）、スクレイピング（収集したHTMLから必要な情報を抽出）
- APIを利用して収集
- IoTデバイスやセンサーを用いて計測したデータ。画像・動画データ、在庫情報や位置情報、生体情報など

2.3 データの適正を判断する

データバイアス

データバイアスとは、データに含まれる偏った認識や差別・偏見を含む偏りのこと。

グループの特徴が違うことに起因するバイアス ECサイトの広告の効果測定では、広告を出していないユーザーも分析対象にする必要がある

集団の一部が脱落してしまうバイアス 新薬の実験をAとBの2グループで実施するとき、片方のグループで多くの参加者が実験から脱落してしまうと、人数に顕著な差が生じてしまう。

データの定義の違いに起因するバイアス 商品のIDや商品名、カテゴリ名が統一されていないと正確に比較ができるない。

利用継続期間の違いに起因するバイアス クラウドサービスで採取できるデータは、ロイヤリティの高い継続利用者層を中心のデータになるため、好意的なデータが多くなる。利用期間の短いユーザー層のデータも対象にすべき。

顧客の意思の違いに起因するバイアス 店舗利用後の調査に協力してくれる人は、すでに店舗に対して愛着を持っている可能性が高い。

サンプリングバイアス

母集団から標本抽出を行う段階で生じる統計的偏りをサンプリングバイアスという。

対策としては

- 事前にデータの偏りの有無を確認する
- **層化抽出法**によって、サンプルをバランスよく抽出する
- アップサンプリングやダウンサンプリングを用いてサンプルの偏りを調整する。ただし、このときにさらなる人為的なバイアスを加えないよう注意が必要

偏りが精度にどれだけ影響をあたえるかは、データのボリュームに依存。データのボリュームに応じて、データの偏りが精度に及ぼす影響を確認することが大切。

標本抽出とサンプリングバイアス

標本の数が不十分な場合、母集団を代表できていない偏った標本のみが抽出されるリスクが高く、母集団の情報を推定した際の信頼性が低下する。これは**標準誤差**が大きくなるということ。分析の目的に対して、どれくらいの推定精度が必要か（どれくらいの標準誤差が許容か）を定め、そこから逆算した目標を満たす標本サイズを決定するのがよい。

無作為抽出にもいくつかの手法がある

- **系統抽出法** 母集団を構成するすべてのデータを並べた上で、最初の標本のみ無作為に選び、それ以降は決められた等間隔で系統的に標本を選ぶ方法。乱数をふる回数が1回のみのため、単純無作為抽出よりもコストが低い。
- **多段抽出法** 都道府県をランダムに選び、市区町村をランダムに選び、などと、無作為抽出を複数の段階に分けて行う方法。各段階で世帯数に比例数確率で対象を無作為抽出する方法もある。

段階数を多くすることで調査コストを下げるができるが、推定誤差が大きくなるリスクもある。

- **層化抽出法** 母集団をあらかじめ複数の層に分け、各層の中から必要な数だけ標本を無作為抽出する手法。層別抽出法や層化サンプリングとも呼ばれる。サンプリングバイアスの軽減に効果的。

アルゴリズムバイアス

分析の手法やアルゴリズムが特定の変数を強調して学習したことにより、分析結果にバイアスが生じてしまうこと。原因がデータバイアスであることもあり、データバイアスとアルゴリズムバイアスは併存することもある。

アルゴリズムバイアスへの対策としては、複数の手法で機械学習モデルを構築し、結果に違いが生じているかを比較することが挙げられる。

データの網羅性

モデルは学習データからパターンを発見するため、学習データに全く含まれない事例に関しては、正しく予測することはできない。

外れ値と異常値

- **外れ値** とは、他のデータからみて極端に大きな値、または極端に小さな値のこと
- **異常値** とは、外れ値のうち、原因がわかっているものを指す。

外れ値は分析にとって意味がある外れ値である可能性もある。

平均値の計算では外れ値の影響は大きいため、トリム平均（両端にある外れ値を除いたデータで計算した平均値）を計算することがある。外れ値の判断では、箱ひげ図や四分位範囲を用いることで見つけることができる。

2.4 データの可視化

可視化の目的

- 2つのデータ群の比較
- 2変数の相関関係
- データの分布や偏り、特異点、外れ値
- 各構成要素が占める割合
- 周期性や突発的な変化

様々な場面と目的に対して有効

- データ分析の初期段階において、現状を把握し、課題を見つけ、仮説を立てる
- 時系列分析などにおいて、可視化で傾向を掴んだのちに、将来に対して予測する
- 可視化を通じて分析の結果を客観的に評価し、それを施策、改善につなげる
- 分析結果を相手に説明し、メッセージを伝えやすくする

様々なグラフ

- **折れ線グラフ** 横軸に日付や時刻などの系列、縦軸は量的なデータを取る。
- **棒グラフ** 一般的には横軸に離散的な値（数字またはカテゴリ）、縦軸にデータ量をとる
- **円グラフ** データの構成割合を表現するためのグラフ
- **度数分布表** データを複数の区間に分割し、それぞれの区間に属するデータの個数を表したもの。1つの量的データの分布を見る。区間のことを**階級**、区間の幅を**階級幅**、階級を代表する値（ほとんどの場合、中央値）を**階級値**という。
- **相対度数** 度数を比率で表したもの。
- **累積度数** 特定の階級までのデータの合計頻度
- **累積相対度数** 特定の階級までの相対度数の合計
- **相対度数分布表** 相対度数による分布表
- **累積相対度数分布表** 累積相対度数による分布表
- **ヒストグラム** 度数分布表を棒グラフにしたもの。横軸は階級、縦軸は度数
- **スタージェスの公式** データ個数に対する階級数の目安

階級数の目安を $1 + \log_2 N$ とする。（Nはデータの個数）

階級幅の目安は、「(データの最大値 - データの最小値) ÷ 階級数」

- **ヒストグラムの平均値** 階級値 × 度数 の和 ÷ 度数の合計
- **クロス集計表** 2つのシス的変数に対して、両方に該当するデータ数または割合を記す。クロス集計表からは確率を計算できる。

箱ひげ図と四分位範囲

- 区間の幅が短いところはデータが詰まっている。区間の長さが長いところはデータの密度が疎
- **四分位範囲 (IQR)** Q3-Q1を四分位範囲という。

Introduction

上限境界値を $Q3+1.5\times IQR$ 、下限境界値を $Q1-1.5\times IQR$ と定め、この範囲にないデータを外れ値と判断する。

2.5 多次元情報の可視化

- ヒートマップ
- 散布図行列
- 平行座標プロット データ間の類似性や異常値を識別しやすくしてくれる
- ポロノイ図 **母点**を基準として平面上の座標空間を分割してできる図。**ポロノイ領域**、**ポロノイ境界**。各母点の勢力範囲を表すと解釈できる。
- アニメーション

3.1 データサイエンスにまつわる基本概念

人工知能の定義

- 人間が持つ知的な情報処理能力を機械に持たせること
- 周囲の状況（入力）によって行動（出力）を変える能力を持つ機械

※本書での考え方、定義

人工知能と機械学習の関係性

- 人工知能
 - ルールベース
 - 機械学習

機械学習とは、明示的にプログラムしなくても学習する能力をコンピュータに与える研究分野（アーサー・サミュエルによる定義）

機械学習においては学習データを用いる、学習のプロセスを必要とする。

なにかを判断（例えばクレジットカードの審査）するには、判定軸としきい値が必要。ルールベースでは、判断基準を人間が決める。機械学習では、判断基準の最適値を学習によって見つける。

機械学習の仕組みとデータの扱い方**

- 特徴量 データの特徴を表す変数
- 正解データ 正解データ、正解ラベル
- 教師あり学習の目的は、特徴量と正解を関連づける法則を見つけること

機械学習技術の応用例

- 回帰予測 (Regression) 連続的な値を取りうる数値を予測するタスク。
 - 売上高の予測
 - サービス会員が一ヶ月の間に購入する商品の総額の予測
 - 株価の予測
- 分類 (Classification) データの特徴に基づいてどのカテゴリに属する可能性が高いかを判断する（主に、確率的に判断する）。
 - 病気の有無や陽性・陰性を判定する
 - スパムメールの判定
 - ニュース記事をカテゴリに振り分ける（確率が最も高いカテゴリに分類する）
 - 農作物の品質や大きさを自動仕分けする
 - 商品の推薦・レコメンドをする
- 異常検知 (Anomaly Detection) センサーなどの測定データから、モデルの学習に効果的な特徴量に基づいて、通常と異なる挙動を示すデータから機器の異常や故障を検知する。

機械学習におけるディープラーニングの立ち位置

Introduction

人工知能 ↴ 機械学習 ↴ ニューラルネットワーク ↴ ディープラーニング（深層学習）

ニューラルネットワークの中に多層パーセプトロンがあり、層が多くなったものをディープニューラルネットワーク (DNN) という。ディープラーニングは、DNNを含む深層モデルを学習・運用するためのアルゴリズム、技術、応用など全体を指す。例えば、訓練データの前処理、ハイパーパラメータの調整、GPUによる高速化、モデルの評価などの要素もディープラーニングに含まれる。

3.2 機械学習の各分野の特徴

- 教師あり学習(Supervised Learning) ... 学習データは特徴量と正解ラベルのセット。分類問題と回帰問題の2種類のタスクを扱う
- 教師なし学習(Unsupervised Learning) ... 学習データに正解ラベルがついておらず、特徴量のみ。クラスタリングと次元削減の2種類の多数を扱う。
- 強化学習(Reinforcement Learning) ... 与えられた条件下で、プレイヤーのようなエージェントが最大限の報酬を得るよう行動を最適化する手段
- 半教師あり学習 ... 教師あり学習と半教師あり学習を組み合わせた手法。学習データとして正解ラベルがついているデータとついていないデータの両方を使う。ラベルのついていないデータにラベル付け（分類問題）するため、ラベルのついたデータで教師あり学習をおこない、信頼度の高いデータのラベルを予測する。新しくラベルづけられたデータセットを用いて再度も出るを訓練する。

正解ラベル付きデータが十分に用意できず、正解ラベルをつけるコストを削減したい場合に利用する。

- アクティブラーニング ... 一部のデータにのみ正解ラベルがついている状態でモデルを構築し、この初期モデルを用いて残りのラベルなしデータから、モデル学習に効果的なデータ（初期モデルが判断を誤り、モデルの精度向上に有用と考えられるデータ）を抽出し、人が手動で正解ラベルを付与する。半教師あり学習と違い、アノテーションが人為的に行われる。

教師あり学習

- 分類問題 ... 各データが所属するカテゴリを推定するタスク。データが持つ属性の違いに基づいてデータ群を分ける境界線を引くイメージ。
- 回帰問題 ... 連続値を予測するタスク。データ点の分布をうまく再現できる関数を見出すイメージ。
 - シングル出力回帰 ... 1つの数値のみを予測する
 - マルチ出力回帰 ... 複数の値を出力する。「感情分析」も回帰タスクの1つで、マルチ出力回帰になる。

※感情分析は、ポジ／ネガや喜び、悲しみなどを判定する場合は分類タスクになる。感情の強度を数値化する場合は回帰タスクになる。

回帰問題を扱うモデル

- 線形回帰
 - 単回帰分析 ... 説明変数が1つ
 - 重回帰分析 ... 説明変数が複数
- 決定木にも回帰を扱う回帰木がある

分類問題を扱うモデル

- ロジスティック回帰 ... ある事象が発生する確率を求める。二値分類やマルチクラス分類に使われる
- サポートベクトルマシン
- 決定木（分類木）
- K近傍法
- ナイーブベイズ
- ニューラルネットワーク ... 多くの場合、画像や音声などの非構造化データの分類に使われる
 - 単純パーセプトロン ... 0/1の判断

Introduction

- 多層パーセプトロン
- ディープニューラルネットワーク (DNN)
- ディープラーニング

その他の分野

- 画像処理
 - 画像分類 (画像分類)
 - 畳み込みニューラルネットワーク (CNN; Convolutional Neural Network)
 - 一般物体認識
- 音声処理
 - 音声生成
 - 混同正規分布モデル
 - 隠れマルコフモデル
 - DNN-HMM
 - WaveNet
- 時系列分析
 - 古典的な手法
 - MA (移動平均)
 - AR (自己回帰モデル)
 - ARMA (自己回帰移動平均モデル)
 - ARIMA (自己回帰和分移動平均モデル)
 - SARIMA (季節変動自己回帰和分移動平均モデル)
 - 状態空間モデル
- 自然言語処理
 - 古典的な手法
 - RNN (再帰型ニューラルネットワーク)
 - Seq2Seq
 - 大規模言語モデル
 - Transformer
- 生成AI
 - 画像生成AI
 - 拡散モデル (Diffusion Model)
 - 音声生成AI
 - 大規模言語処理AI

教師なし学習

- クラスタリング ... データをいくつかのクラスタに分けることでデータ特性を浮かび上がらせる手法
- 次元削減 ... データを低い次元に圧縮することで重要な情報を際立たせる手法

クラスタリングを扱うモデル

- 階層クラスター分析
 - デンドログラム (樹形図) を用いた手法
- 非階層クラスター分析
 - K平均法

次元削減を扱うモデル

- 主成分分析

強化学習

エージェントが最大の報酬を得られるように最適な行動を学習するタスク。囲碁などのゲームで活用されていた。近年は自動運転やロボティクスなどにも応用されている。

強化学習は、「状態・行動・報酬」に基づいて、意思決定ルールを見出す。エージェントは、環境が与えた状態に對して、行動の試行錯誤を繰り返し、最適な報酬が得られる行動のルールを学習していく。

従来の方法は、組み合わせをすべて計算するため、現実的な速度で課題に対応することができなかった。これが、深層強化学習によって解決された。

深層強化学習は、深層学習（ディープラーニング）と強化学習を組み合わせた手法。代表的な手法は **DQN (Deep Q-Network)**。DQNは、行動価値関数を近似するために、画像認識にも使用されている畳み込みニューラルネットワーク(CNN)を用いる。ニューラルネットワークを用いることで、状態の数が膨大になっても学習を現実的な時間内で終了することができる。

Atari社のブロック崩しが2013年に人間のスコアを超えた。2015～17年にDeepMind社がAlphaGoを開発。2017年に完全自己対局で学習できるAlphaGo Zeroも開発。DQNをベースに多くの深層強化学習モデルが開発され、自動運転、ロボティクスなどにも活用されている。

3.3 データ処理と特徴量エンジニアリング

特徴量エンジニアリングとは

- 予測変数として採用する列の選別
- データに前処理を施し、予測に効果的な形に加工する

文字列データを数値データに変換

- **カテゴリカルデータ** カテゴリを区別するために用いられるデータ。カテゴリカルデータは質的変数。数値で表現することもあるが、意味を持たないダミー数値。
- **One-hotエンコーディング** カテゴリ毎に列をつくり、1つの列項目だけを1、それ以外を0にする。スパースな行列になりやすい。列が増えすぎると特徴量として扱いにくい場合もある。（列の数=特徴量の数が多くなり、過学習しやすくなる）
- **ラベルエンコーディング** 1つのカテゴリが1つの数値に対応するように数字に置き換える。

データ欠損の処理

- **リストワズ法** 欠損値があるサンプルをそのまま削除する方法
- **回帰補完** 欠損列と非欠損列の間に相関が強い場合、回帰を利用して欠損値を埋める方法
- **統計量で補完** 欠損部分が全体に影響しないよう、平均値、中央値、最頻値などの統計量で埋める方法

数値データが抱える問題

- 変数間にスケールが異なる
- 変数間で何等かの関係（依存関係や因果関係など）が存在する
- 分布に偏りがある
- 外れ値や異常値を含む

分析結果への影響は、分析手法やモデルの種類によって異なる。SVMやK近傍法は外れ値の影響を受けやすいが、決定木やアンサンブル学習器は外れ値に比較的頑丈。

対数変換、正規化と標準化

- **対数変換** 対数変換は、偏りのある裾の長い分布をしたデータを正規表現に近づけるために有効
- **正規化(Normalization)** 最小値が0、最大値が1の無次元の量になるようにデータを変換する
- **標準化(Standarization)** 平均が0、分散が1の無次元の量になるようにデータを変換する

正規化と標準化の効果

- スケールが揃うことで比較しやすくなる
- モデルの学習がより高速かつ高精度になる

正規化

正規化は、0-1正規化などとも呼ばれる。以下の式で変換する。

$$\frac{\text{データの値} - \text{最小値}}{\text{最大値} - \text{最小値}}$$

注意点として、正規化は**外れ値に敏感に反応してしまう**という特徴がある。極端に小さい値または大きい値がデータに含まれている場合、変換後のデータも偏ってしまう。

標準化

各データから平均値を引いて標準偏差で割ることで標準化する。

$$\hat{x}_i = \frac{x_i - \mu}{\sigma}$$

機械学習を用いた予測モデルを作る際、特微量に標準化を適用することが推奨されている。

標準化は**外れ値に頑丈**である。

使い分け

- 正規化の方が計算コストが低い。最大や最小がわかっている場合は正規化が使われることもある。
- 大量なデータに対して高速に演算を繰り返す必要のある画像処理が代表例
- データに外れ地がある場合、標準化の方が適している。データの最大値や最小値が未知である場合も標準化が使われる。

量的変数の離散化

離散化とは、**量的変数を質的変数に変換すること**。離散化によって、データの特徴を強く反映する特微量に変換でき、以下のような効果が期待できる。

- データのスケールを揃える
- データの変動による影響を緩和する
- 結果の解釈が簡単になる

固定幅による離散化がシンプル（年齢データを10歳区間で区切る）。

3.4 機械学習を応用する上で重要な概念

過学習

- **汎化性能** … 答えが未知な新しいデータに対しても予測性能を出せること
- モデルは、複雑であればあるほど過学習しやすい（多項式回帰の次数が多い状況をイメージ）
- 過学習の代表的な原因
 - データ数が少ないので、**特徴量の数**が多すぎる
 - 相関が強い特徴量が多く存在する
 - モデルが複雑すぎる
- バイアスとバリアンス
 - バイアスは、**推定値と実測値の差**を表す
 - バリアンスは、**推定値のばらつき**を表す
 - バイアスとバリアンスはトレードオフの関係にある
- **次元の呪い** 機械学習モデルの変数の数を増やしすぎることで学習がうまくいかなくなること。また、必要な訓練データが指数関数的に増えてしまうこと。

※ 「バイアス」という用語はいろいろなところで出てくる

- **バイアスとバリアンス**のバイアスは、推定値と実測値の差を表す。統計・予測モデルの偏りを表す言葉
- ニューラルネットワークや回帰モデルの定数項もバイアスやバイアス項と呼ばれる
- 偏りを一般に表す言葉として「バイアス」という言葉が使われる。サンプリングバイアス、データバイアス、アルゴリズムバイアスなど。

過学習への対策

- 学習データの量を増やす
- ハイパーパラメータを調整してモデルの複雑さを抑える
- 正則化を実施する

損失関数

損失関数(Loss Function)とは、予測値と正解値とのズレの大きさを計算するための関数。コスト関数とも呼ばれる。機械学習の学習とは、この損失関数を最小化するようモデルのパラメータを更新していくこと。

ノーフリー・ランチ定理

あらゆる問題において高い精度を出せる汎用的なモデルは存在しない、というもの。

万能の機械学習モデルやアルゴリズムは存在しない。しない。例えば、画像データを分析対象とする画像認識にはニューラルネットワークが最も強力なアルゴリズムと知られている。一方で、表形式の購買履歴データを分析する場合は決定木などの手法が適切。目的、データの種類、処理方法、分析のコストなど様々な要素を踏まえて、ケースバイケースで判断することが重要。

3.5 アソシエーション分析

機械学習というよりは従来型のデータマイニング手法。

- 条件Aのときに、事象Bが起こる可能性が大きい
- 事象Aが発生すると別の事象Bが一緒に発生しやすい

このようなルールをアソシエーションルールという。こういったルールを見つける手法。

バスケット分析

ある商品と一緒に購入される可能性が高い商品はどれかを分析する手法。アソシエーション分析の1つだが、商品の購入のみを対象としていて、「ウェブサイト訪問」といった行動履歴など他の情報と一緒に分析する手法ではない。

ABC分析

売上の占める割合でA、B、Cにランク付けする。典型的には70%、25%、5%でわかる。

アソシエーション分析で使われる指標

信頼度 ... Aが発生した条件のもとでBも発生する割合（条件付き確率）

$$\frac{n(A \cap B)}{n(A)}$$

- 信頼度が大きい場合、同時に発生する可能性は大きい（Aを購入した顧客がBも購入する可能性が大きい）
- ただし、Aの購入者が全体に対して少ない（レアケース）場合や、Bが人気の高い商品であるときも信頼度が大きくなるため、注意が必要。

支持度 ... AとBが同時に発生する割合

$$\frac{n(A \cap B)}{n(U)}$$

- 支持度は全体に対する割合になるため、この値が大きいということは全体に対して影響が大きい。
- 信頼度も支持度も両方とも高い場合は、どちらもすでに人気がある場合があるため注意が必要

計算例

ウェブニュースサイトに1日1000人が訪問した。

- A: スポーツのカテゴリを見た人が500人。そのうち旅行カテゴリも見た人が300人
- B: 旅行カテゴリを見た人が50人。そのうちグルメカテゴリも見た人が40人
- C: グルメカテゴリを見た人が200人

	信頼度	支持度
A	$\frac{300}{500} = 0.6$	$\frac{300}{1000} = 0.3$
B	$\frac{40}{50} = 0.8$	$\frac{40}{1000} = 0.04$

Bの信頼度は高いが支持度が低い。この2つであれば、Aを重要視して、スポーツカテゴリを見た人に旅行カテゴリを推薦するのが効果的。

リフト値 ... 事象Aが事象Bをどの程度引き上げる効果があるのか

$$\frac{\frac{n(A \cap B)}{n(A)}}{\frac{n(B)}{n(U)}}$$

全体に対するBの割合よりもAに占めるBの割合の方が大きい場合に値が大きくなる。つまり、AはBを引き上げる効果を持っている、捉える。

3.6 線形回帰

線形回帰とは、1つ以上の説明変数を利用して、連続値である目的変数を直線モデルで予測する方法。

予測変数 Y が、説明変数 X の線形結合で表されるモデル。

単回帰分析

説明変数が1つの線形回帰の手法

単回帰分析モデルを表す線形式は以下:

$$Y = aX + b$$

ここで、 a, b は回帰係数やパラメータ、重みなどと呼ばれる。

重回帰分析

説明変数が複数

重回帰分析モデルを表す線形式 :

$$Y = a_1 \cdot X_1 + a_2 \cdot X_2 + \cdots + a_n \cdot X_n + b$$

$a_i (i = 1, 2, \dots, n)$: 偏回帰係数

- **多重共線性 (Multicollinearity)** : 相関が強い説明変数（特徴量）を組み合わせた時に、それらが干渉し合うことを という。

最小二乗法(Least Square Method)

以下の L が 一番小さくなるように $f(x)$ の係数を求める。

$$L = \sum_{i=1}^N \{y_i - f(x_i)\}^2$$

L は 損失関数 (Loss Function) という。

線形回帰の場合、回帰係数は損失関数を、 a と b で偏微分した式がともに0となるように方程式を解くことで a と b が求まる。

決定係数 R^2

決定係数 とは、回帰式が実データをどの程度再現しているかを評価するための値。（線形ではないデータ、曲線ようなデータを線形で近似していないかどうかをチェックする）

決定係数 R^2 は、以下で定義される。

$$R^2 = 1 - \frac{\sum_{i=1}^n [y_i - f(x_i)]^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

- 分子は、実測値と予測値の二乗誤差で、予測値の分散を表す。
- 分母は、実測値と実測値の平均との二乗誤差で、実測値の分散になっている。（分散の式は、さらに n で割る）

決定係数 R^2 は0から1の間をとり、**1に近いほど良いモデル**（1に近いほど回帰式がデータとの重なりが大きくなる）。

3.6 正則化

損失関数にペナルティ項を追加することで、過学習を防ぐ。

正則化では、損失関数にペナルティ項を追加した次の式を最小化する。

$$E(w) + \lambda \frac{1}{p} \sum_i |w_i|^p$$

ここで、パラメータ λ は、ペナルティの重さを制限する役割を持っている。 λ を大きくすると過学習に陥りにくくなるが、逆に学習不足になることもある。

- **L1正則化**: $p = 1$ とする

特徴選択と次元圧縮に効果的

- **L2正則化**: $p = 2$ とする

パラメータの大きさをゼロに近づける効果があり、汎化性の高いモデルが得られる。（過学習防止に効果的）

- **ラッソ回帰**: L1正則化を取り入れた線形回帰のこと
- **リッジ回帰**: L2正則化を取り入れた線形回帰のこと
- **Elastic Net**: ラッソ回帰とリッジ回帰を組み合わせた手法

3.7 ロジスティック回帰

複数の説明変数をもとに、目的変数である「ある事象が発生する確率」を求める非線形回帰の手法。

ロジスティック回帰は、二値分類とマルチクラス分類の両方に適用できる。（回帰なのに分類となっているが、ロジスティック回帰は、カテゴリになる確率を予測する）。

例

- 顧客が商品をリピート買いする確率
- 病気の発症確率
- スパムメールである確率
- 発行されたクーポンを使用する確率

二値分類を扱うロジスティック回帰

説明変数を X_i 、目的変数を Y としたとき、ロジスティック回帰は以下の式になる。

$$Y = \frac{1}{1 + e^{-(a_1 X_1 + a_2 X_2 + \dots + a_n X_n + b)}}$$

この式は、シグモイド関数に説明変数 X_i の線形結合を代入した式になっている。

シグモイド関数は、以下で表される

$$f(x) = \frac{1}{1 + e^{-x}}$$

シグモイド関数の値域は0から1になるため、確率として扱いやすい。

ロジスティック回帰の偏回帰係数も最小二乗法により求めることができる（そうだが）、損失関数として交差エントロピーを使うものもあるとのこと。

（具体例で考える）

線形回帰は、学習データとして実測値があり（イメージは、説明変数Xと目的変数Yによる散布図）、未知のデータに対する値を予測する、というもの。

ロジスティック回帰は二値分類で、確率を求めるもの。目的変数が量的変数ではなく質的変数になる。

例）商品をリピート買いするかどうか

- 購買履歴（説明変数）から、ある顧客が商品をリピート買いするかどうかを分類する（リピート買いする=1、しない=0）
- 学習データとして、実測値（購買履歴）が与えられる
- 実測値を下に、ロジスティック回帰式のパラメータを求める
- 未知のデータに対して、ロジスティック回帰式を適用し、予測値（リピート買いするかどうかの確率として、0から1の値）を得る

マルチクラス分類を扱うロジスティック回帰

<https://datawokagaku.com/multinomial/>

マルチクラス分類に適用する方法は2つ

One vs Rest (OvR): 1つのクラスとそれ以外の全クラスの二値分類器をクラスの数だけ作って、最も高い確率のクラスを予測結果とする。OvRはロジスティック回帰だけでなく、他の二値分類のアルゴリズムにも応用できる。

多項ロジスティック回帰:

ロジスティック回帰によるマルチクラス分類を多項ロジスティック回帰という。

多項ロジスティック回帰では、シグモイド関数の代わりに**ソフトマックス関数**を使う。

ソフトマックス関数は、以下で表される

$$p_k(x) = \frac{e^{x_k}}{e^{x_1} + e^{x_2} + \dots + e^{x_K}}$$

ここで、 K はクラス数（赤、青、黄色に分類したいなら、 $K = 3$ ）になる。

この関数は、各クラス毎の値が各クラスの確率を表し、各クラス毎の値を計算して足し合わせると 1 になる。

実際の予測値を算出する式は複雑になるが、簡単のため、説明変数が 1 つとして展開してみる。

線形回帰では、予測値を説明変数の線形式で表した。今、説明変数は 1 つとしているので、予測値 Y に対する線形回帰は以下になる。

$$Y = aX + b$$

二値分類のロジスティック回帰では、右辺の線形結合をシグモイド関数に代入していた。観測データに対する予測値が 1 になる確率は、以下で与えられる。

$$P(y = 1) = \frac{1}{1 + e^{-(aX+b)}}$$

マルチクラス分類の多項ロジスティック回帰では、各クラスごとに回帰係数があるとイメージ。赤、青、黄色などの 3 クラス分類を想定すると、確率としては $P(y = 1)$, $P(y = 2)$, $P(y = 3)$ の 3 つの確率が考えられる。

多項ロジスティック回帰で、観測データがあるクラスに分類される確率は以下で与えられる。

$$P(y = k) = \frac{e^{a^{(k)} X + b^{(k)}}}{e^{a^{(1)} X + b^{(1)}} + e^{a^{(2)} X + b^{(2)}} + \dots + e^{a^{(K)} X + b^{(K)}}}$$

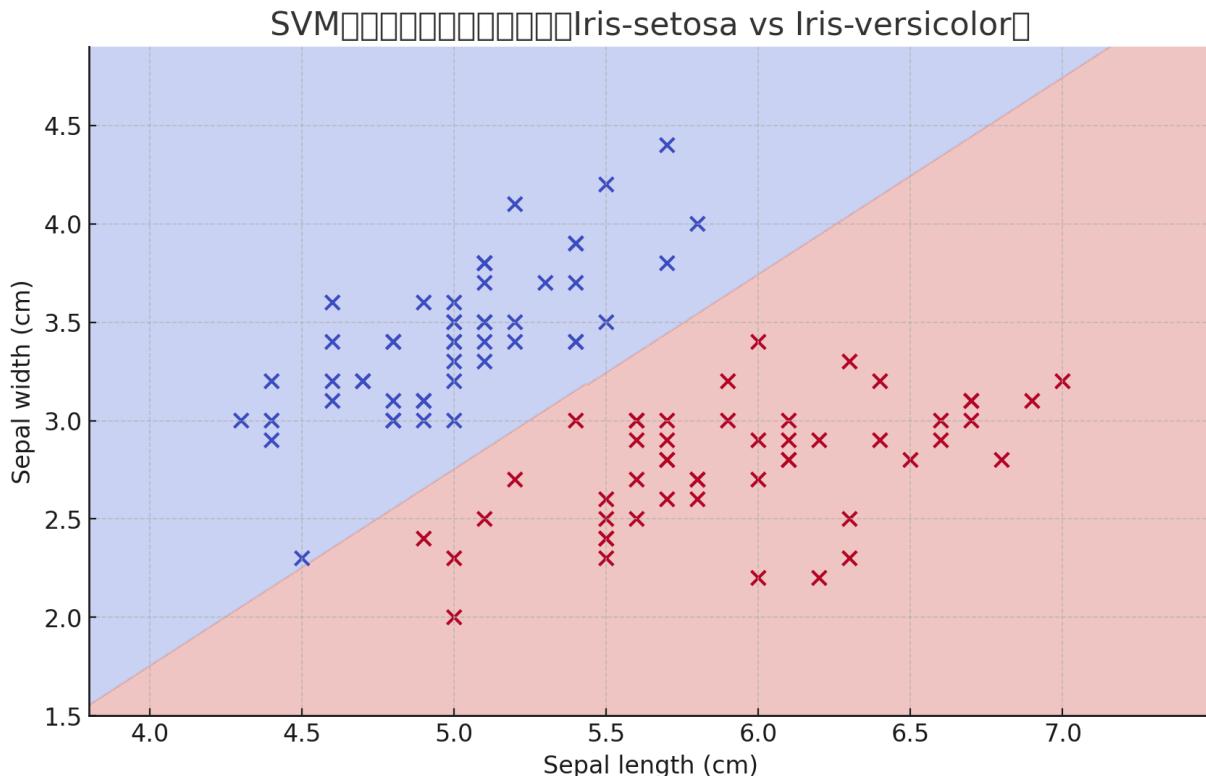
ここで、 K はクラス数を表す。

回帰係数も、各クラスごとにあるようなイメージとなり、以下のようにたくさんのパラメータが出てくる。

$$\begin{aligned} & a^{(1)} X + b^{(1)} \\ & a^{(2)} X + b^{(2)} \\ & a^{(3)} X + b^{(3)} \end{aligned}$$

3.8 サポートベクトルマシン(SVM)

サポートベクトルマシンは、通常、二値分類問題に適用するため、目的変数は二値となる。例えば花の形状を表す特徴量から花の種類（種類）を分類するなど。



(アイリスデータセットで、花の特徴と種類をプロットした例。XY軸は目的変数、花の種類は色で表現されている)

- **決定境界:** 境界線のこと
- **分離ハイパープレーン:** 境界線のこと
- **サポートベクトル:** 決定境界から最も近い訓練データの点。(イメージ的には、決定境界から点への線)
- **マージン:** サポートベクトルと境界線の距離
- **マージン最大化:** SVMでは、サポートベクトルと境界線の距離を最大化するように境界線を決定する
- **カーネル:** 非線形なデータを高次元空間に変換するための関数。線形分離不可能なデータも線形に分離できるようになる。
 - 線形カーネル: データをそのまま利用する
 - 多項式カーネル: データを多項式の形で変換
 - ガウス基底関数カーネル(RBF): データを無限次元空間にマッピングする。
- **カーネルトリック:** カーネル方による計算量を削減する技術
- **ソフトマージン:** データのノイズや誤分類を許容するアプローチ。モデルの柔軟性が増す。どの程度、データ点がマージン内へ侵入してもよいか表すパラメータをスラック変数という

サポートベクトルマシンでは、目的変数は1か-1にエンコードする。新しい観測データについて、決定境界を表す式 $f(x)$ が正を返したら1、そうでなければ-1と予測する。

$$\text{予測クラス} = \begin{cases} 1 & \text{if } f(x) > 0 \\ -1 & \text{if } f(x) \leq 0 \end{cases}$$

3.9 決定木

特徴量（説明変数）を軸に複数のグループに分割し、最下段の各ノードがなるべく同じ属性（同じクラス）のデータのみで構成されるような分割軸としきい値を探す。

分類（目的変数が質的データ）で用いられるものを分類器と呼ぶが、連続値を扱う回帰器もある。

学習は、ノード内の不純度を最大限減らすように、分割軸としきい値が最適化されていく。不純度は、**ジニ不純度**や**エントロピー**を評価指標として用いる。

モデルの構造とデータ処理がシンプルだが、過学習しやすい。外れ値やノイズの影響を受けやすく、データ分割に偏りが生じやすい。データ量が少ないので特徴量の数が多いときに起こりやすい。

アンサンブル学習器

アンサンブル学習器とは、単純なモデル（弱学習器）を多数組み合わせて使うことで精度を改善する手法。バギングとブースティングという手法がある。

- **バギング**: 弱学習器を多数作り、並列に学習されたあと、すべての結果を結合する。（例：ランダムフォレスト）
- **ブースティング**: 弱学習器を逐次的に作ることで、以前に構築された弱学習器の結果を用いる。（例：勾配ブースティング木）

ランダムフォレスト

- データと特徴量の両方をランダムにサンプリング
- データは、重複を許すサンプリングを行う（ブートストラップ法）
- 特徴量は、全特徴量が M 個なら、 \sqrt{M} 個程度を抽出する
- 分類問題の場合は多数決、回帰問題の場合は予測値の平均値を出力する。

ランダムフォレストは、並列処理ができるため高速。実装しやすく、同時に高い精度を見込める。

勾配ブースティング

- 1つの弱学習器モデルを作成し学習する
- 最初の学習器で誤識別したデータに重みを増やして、次の弱学習器ではその部分を重点的に学習する。
- 上記を繰り返す
- XGBoost, AdaBoost, lightGBMなどがある

ブースティングはチューニングが難しく、機械学習上級者向け。学習に時間がかかる。

一方で、高い精度をだすこともでき、KaggleやSignateなどのコンペでは人気がある。

分類木は、多種多様な特徴量を扱う柔軟性があるので好まれる。

決定木モデルの強み

- 構造化データを扱うため、特徴量が扱いやすく、モデルによるデータ処理の仕組みが直感的にわかりやすい
- 決定木の仕組みをツリー構造で可視化するツールもある。
- ランダムフォレストでは、特徴量の重要度を可視化することができる。（説明がしやすい）

3.10 K近傍法 (KNN: K Nearest Neighbor)

K近傍法は、主に分類問題に用いられる手法。特徴量をベクトル空間にプロットし、未知のデータに対して、距離が近い順にK個を取得し、その多数決でデータが属するクラスを推定するという手法。Kはユーザーが設定するパラメータ。

K近傍法は仕組みが単純で幅広いタスクに適用できるが、以下のような問題がある。

- 学習データのクラス間の偏りに影響を受けやすい
- **Kの設定**に影響を受けやすい
- 高次元データには不向き（計算量が多くなる＝次元の呪い）

K近傍法は、精度が常に高くなくても致命的ではないケースに手軽に使用される手法。

使用例：レコメンドエンジン

- Aさんと各ユーザーの間の類似度を計算する（似た評価をする場合に類似度が高くなるように計算する）
- 類似度が上位のK人を取り出す
- Aさんが未評価の映画の評価値を、類似度を重みとした加重平均で算出し、推定評価値とする
- 推定評価上位10件の映画をAさんにレコメンドする

3.11 ナイーブベイズ

ナイーブベイズは、与えられたデータから各クラスの確率を計算し、**事後確率**が最も高いクラスに観測データを分類する、という手法。

例えば、メールのスパム判定や良い知らせか悪いしらせかを、文章中に含まれる単語の出現頻度を使って推測する。

使用例

- リアルタイム予測
- テキスト分類
- レコメンド
- センチメント分析

無相関の仮定

ナイーブベイズでは、説明変数が独立して予測対象に影響を与えると仮定する。説明変数同士は互いに無相関であると仮定される。

ベイズの定理

ベイズの定理

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

この式自体は条件付き確率の定義から簡単に出てくるが、事前確率、事後確率、原因、結果などの捉え方をしっかり理解する必要がある。

具体例

病気の検査の例がイメージしやすい。

- A : 病気に罹患している（これを「原因」の事象と捉える）
- B : 検査結果が陽性であった（これを「結果」の事象と捉える）

ベイズの定理は、以下のようにかける

$$P(\text{罹患している} | \text{検査結果が陽性}) = \frac{P(\text{検査結果が陽性} | \text{罹患している}) \cdot P(\text{罹患している})}{P(\text{検査結果が陽性})}$$

ここで、左辺は、「検査結果が陽性だったときに、実際に病気に罹患している確率」といえる。左辺の確率を直接的に計算することは難しいが、右辺に出てくるそれぞれの確率があらかじめ算出されていれば、左辺の確率が計算できる。

https://www.tech-teacher.jp/blog/statistics_4_conditional/ の例を参考にして、具体例を示す。

罹患率が0.1% (0.001)の病気の検査を考える。

この検査では、陽性、陰性が以下のように判定されることがわかっているとする。

Introduction

- 病気に罹患している人に対して陽性とでる確率が95%、陰性とでる確率が5%
- 病気に罹患していない人に対して陽性とでる確率が20%、陰性とでる確率が80%

表にすると以下になる。

	陽性	陰性
病気に罹患している	95%	5%
病気に罹患していない	20%	80%

ここで、陽性と判定されたときに実際に病気に罹患している確率を求めたい。

求めたいのはベイズの定理の左辺の値になる。そのため、右辺に出てくる確率を事前にわかっている情報から求める。

$$P(\text{検査結果が陽性} \mid \text{罹患している}) = 0.95$$

$$P(\text{罹患している}) = 0.001$$

$$P(\text{検査結果が陽性})$$

$$= P(\text{罹患していて陽性}) + P(\text{罹患していないのに陽性})$$

$$= 0.001 \times 0.95 + 0.999 \times 0.2$$

$$\simeq 0.2$$

最終的に求めたい確率は

$$\frac{0.95 \times 0.001}{0.2} = 0.00475$$

結果としては 0.475% になる。

事前確率、事後確率

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

ベイズに定理において、 $P(A|B)$ を事後確率、 $P(A)$ を事前確率、 $P(B|A)$ は尤度という。

BのもとでAが起こる確率を、AのもとでBが起こる確率やA、Bの確率から計算できる、というのがベイズの定理。

病気の検査の例だと、検査結果が陽性のときに罹患している確率を、罹患しているときに陽性である確率や、罹患している確率、陽性である確率などから求めている。

ナイーブベイズ分類器への応用

https://qiita.com/kazuya_minakuchi/items/c3a0066e90cfdfc8a859

文書内に含まれる単語の出現頻度からカテゴリを分類するタスクを考える。

カテゴリはAとBの2つのみとする。

あらかじめ、以下のデータが与えられているとする。

出現する単語	カテゴリ
a,b,c	A
a,c	A
c	A
a	A
a,b	B
a	B
b	B

このとき、

出現単語が a のみである記事のカテゴリを推定したい。つまり、出現単語が a のみであるとき、カテゴリがAである確率とカテゴリがBであるときの確率を計算し、確率が高い方を推定値としたい。

求めたい確率は以下の 2 つ

- $P[A|(a, \bar{b}, \bar{c})]$ … 出現する単語がaのみであるとき、カテゴリがAである確率 … (1)
- $P[B|(a, \bar{b}, \bar{c})]$ … 出現する単語がaのみであるとき、カテゴリがBである確率 … (2)

まず、1 つ目の確率を考える。ベイズの定理から、1 つ目の確率は以下で計算できる。

$$P[A|(a, \bar{b}, \bar{c})] = \frac{P[(a, \bar{b}, \bar{c})|A] \cdot P[A]}{P[(a, \bar{b}, \bar{c})]}$$

$P[(a, \bar{b}, \bar{c})|A]$ は、ナイーブベイズの前提としている説明変数の独立性より、 $P[a|A], P[\bar{b}|A], P[\bar{c}|A]$ の積で計算してよい。

$$\begin{aligned} & P[(a, \bar{b}, \bar{c})|A] \\ &= P[a|A] \cdot P[\bar{b}|A] \cdot P[\bar{c}|A] \\ &= \frac{3}{4} \cdot \left(1 - \frac{1}{4}\right) \cdot \left(1 - \frac{3}{4}\right) \end{aligned}$$

$P[A]$ は、カテゴリがAである確率で、4/7 である。

$P[(a, \bar{b}, \bar{c})]$ は、計算しなくてよい ((1)と(2)の両方に出てくるが、値は定数となるため、比較するときは無視して良い)

上記から、 $P[A|(a, \bar{b}, \bar{c})]$ は以下になる

$$P[A|(a, \bar{b}, \bar{c})] = \frac{4}{7} \cdot \frac{3}{4} \cdot \left(1 - \frac{1}{4}\right) \cdot \left(1 - \frac{3}{4}\right) \simeq 0.080$$

同様に、 $P[B|(a, \bar{b}, \bar{c})]$ を計算すると 0.094 になる。

B である確率の方が高いため、カテゴリとしては B と推定する。

3.12 ニューラルネットワーク

多くの場合、画像や音声などの非構造化データの分類に利用される。

入力層、隠れ層、出力層で構成される。分類問題に用いる場合、出力層には**予測クラス**の数だけノードがあり、各ノードからの出力値は該当クラスにデータが属する確率を表す。

例：0から9までの数字が手書きで書かれた画像を識別（分類）する場合出力層は10個のノードがあり、各ノードは、画像がその数字である確率を表す。

- **全結合層**: 隠層のニューロンが前後の層すべてのノードとエッジで結合されていること。
- **単純パーセプトロン**: 構成要素の最小単位
- **多層パーセプトロン**: 単純パーセプトロンをたくさんつなぎあわせたもの
- **活性化関数**: 各ニューロンで、入力データ（1つ前の層の数だけある）と出力データ（1つの数値）を対応付ける関数。

単純パーセプトロン

- 入力層と出力層の2層のみのシンプルな構造。
- 出力層のノードは0か1の二値のみ出力。
- **線形分離可能な問題にしか使えない** (活性化関数が線形結合で表現されるため、直線や平面で分離できる問題にしか使えない)

多層パーセプトロン(MLP)

- 入力層と出力層の他に、**中間層（隠れ層）**を持つ、3層以上のパーセプトロン
- 活性化関数に非線形関数（シグモイド関数などと思う）を用いることで、複数クラスの分類と線形分離不可能な問題にも対応可能

ディープニューラルネットワーク(DNN)

ディープニューラルネットワークは多層パーセプトロンと構造的には似ているが、隠れ層の数が多いニューラルネットワークを指す。使われる活性化関数もMLPと異なるものが使われることがある。

	MLP	DNN
定義	少なくとも1層の隠れそうを持つフィードフォワード型のニューラルネットワーク	MLPと構造は似ているが、一般的には「深い」層構造を持つニューラルネットワーク
構造	全結合層で構成される	MLPと同じだが、層が深くなることで複雑なパターンをより細かく表現できるようになる
活性化関数	非線形活性化関数（ReLUやシグモイドなど）を使って、層間で非線形変換を行う	より深い層で学習を当たさせるため、ReLUやLeaky ReLUなどが多く用いられる
層数	1～数層程度	通常5層以上を持つ
用途	比較的単純な分類や回帰タスク。画像認識には限界がある	画像認識、自然言語処理、音声認識などに用いられている

ディープラーニングは、DNNを含む多層構造のニューラルネットワークを用いて複雑なパターンや特徴を学習するための機械学習の一分野。ディープラーニングはDNNだけでなく、**CNN**、**RNN**、**生成モデル**などを含む。

DNNはディープラーニングの手法の1つ。ディープラーニングは、DNNを含む深層モデルを学習・運用するためのアルゴリズム、技術、応用など全体を指す。例えば、訓練データの前処理、ハイパーパラメータの調整、GPUによる高速化、モデルの評価などの要素もディープラーニングに含まれる。

重みとバイアス

ニューロン間の結合の強さを線形結合で表現する。

$$x = w_1x_1 + w_2x_2 + \cdots + w_Nx_N + b$$

ここで、 (w_1, w_2, \dots, w_N) を重み、 b をバイアスという。

活性化関数

重みとバイアスで線形結合された値を、**活性化関数**にわたすことでの出力が決定される。

活性化関数としては、シグモイドやReLUなどが使われる。なお、活性化関数にシグモイドを使うと、ロジスティック回帰と同じ式になる。

誤差逆伝播と勾配消失問題

一般的な機械学習の最適化プロセスと同様、ニューラルネットワークでも出力データと正解ラベルの誤差を小さくするようパラメータを最適化していく。つまり、**損失関数を最小化し、最適解に近づけるために、ネットワークの各層の重みやバイアスを繰り返し調整する**のがニューラルネットワークにおける学習になる。

入ラルネットワークの最適化は、出力層に近い層から更新されていく。これを**誤差逆伝播 (Back Propagation)**と呼ぶ。

勾配消失問題とは、この誤差逆伝播を計算する仮定で、**勾配**に関する情報が小さくなってしまい、伝播ができなくなり、学習が進まなくなる問題。**勾配**とは、**勾配降下法**において、各パラメータに対して損失関数の値がどのように変化するかを示す**微分の値**。勾配の値が大きい場合、パラメータを大きく変化させて損失を減らす方向へ進むことができる。勾配の値が小さいと、パラメータの更新量が小さくなり、学習が遅くなる。層を越るたびに勾配が小さくなっていくことがある。入力層に近い層では勾配が非常に小さくなり、重みがほとんど変化しなくなるため、学習が進まなくなる。このことを「勾配消失」という。

活性化関数にReLUやその変種は、勾配がゼロになりにくく、勾配消失を軽減する。その他にもいくつかの対処法がある。

3.13 教師なし機械学習

- クラスタリング（クラスター分析）
 - 階層クラスター分析
 - 非階層クラスター分析
- 次元削減

クラスタリング（クラスター分析）

類似性を持つデータ同士をグループに分類する方法。

正解データは与えられないため、データの特徴に基づいてグループに分ける。

階層クラスター分析

データ集合から最も近いデータを順にグルーピング。テンドログラム（樹形図）では、距離を分歧点までの高さで表す。適切な高さで線を引くことで、データをクラスター（グループ）に分ける。

階層化されているため説明しやすい。ブランドや商品など視点を定めて顧客情報をクラスター化するなど。

データ数が多すぎると、テンドログラムが複雑になり、クラスターを決めにくくなる

非階層クラスター分析

あらかじめ決めておいた数のクラスターにデータを分離する。

代表的な手法はK-means (K 平均法)。

- Kはユーザーが指定する
- データをランダムにK個のクラスターに分ける
- 各クラスターのデータ点の重心を求める
- 各データ点とK個の重心の距離を計算する
- K個の重心のうち、距離が一番近い重心を含むクラスターに各データを割り当て直す
- 上記を、重心の位置が変化しなくなるまで繰り返す

高速かつ大規模データの分析に適している。SNSの投稿を分類するなど。

次元削減と主成分分析

主成分分析(PCA: Principal Component Analysis)は、次元削減の代表的な手法。

例：国語、数学、理科、社会、英語という5教科の点数（5次元のデータ）を、「理系力」と「文系力」という2つの軸で評価する（2次元のデータ）。

相関のある多くの変数を、相関が少ない合成変数（主成分）に要約する。

次元削減と教師あり学習の併用

学習データに対して次元削減を施することで、精度を向上させることができる。

3.14 モデルの精度評価

データ分割の方法

ホールドアウト法 学習データを訓練データとテストデータの2つに一度わけるだけ。訓練:テストを7:3または8:2とする場合が多い。

- データ量が大きいときでも比較的短時間で精度スコアを算出できる
- 一方で、分割後のデータに偏りが生じると、精度検証結果の信頼性が落ちる。特にデータ量が少ないとときは注意が必要。

交差検証法

- データ全体をK個のグループに分ける
- 1つをテストデータ、残るK-1個を訓練データにして、学習と精度の評価を行う
- 別のグループをテストデータとして、同じように学習と精度の評価を行う
- K個の精度評価の結果を平均してモデルの精度とする
- データが少なくとも信頼できる精度評価が得られる
- 計算の量やかかる時間は大きくなる

精度評価のための指標

混同行列

	予測結果が0	予測結果が1
正解クラスが0	真陰性のデータ数(TrueNegative)	偽陽性のデータ数(FalsePositive)
正解クラスが1	偽陰性のデータ数(FalseNegative)	真陽性のデータ数(TruePositive)

- 真陰性** 予想結果が0（陰性）で、予想結果が正しい
- 偽陰性** 予想結果が0（陰性）で、予想結果が誤っている
- 真陽性** 予想結果が1（陽性）で、予想結果が正しい
- 偽陽性** 予想結果が1（陽性）で、予想結果が誤っている

精度指標

正解率(Accuracy): 全体に対する真陰性、真陽性の割合（予想結果が正しい割合）

$$\frac{\text{真陽性}(TP) + \text{真陰性}(TN)}{\text{全体}}$$

適合率(Precision): 陽性判定されてデータのうち、実際に陽性だったデータの割合。偽陽性を避けたいときに注目。

$$\frac{\text{真陽性}(TP)}{\text{真陽性}(TP) + \text{偽陽性}(FP)}$$

- 陽性判定の正確性を示す指標。
- 適合率を重要視する場合、偽陽性を少なくしたくなるため、陽性判断の基準が厳しくなる。そのため、陽性的可能性があっても陰性と判断されやすくなり、本来は陽性なのに陰性と判断される（偽陰性）データが多くなる。

偽陰性のデータが多くなると、再現率は下がることになる。(※)

- 病気診断では、偽陰性は避けたいので、適合率を重視すべきではない。

再現率(Recall): 陽性データのうち、陽性として検出されたデータの割合。陽性を見落としてしまうリスクが高いときに注目。

$$\frac{\text{真陽性}(TP)}{\text{偽陰性}(FN) + \text{真陽性}(TP)}$$

- 陽性だったデータのうち、陽性と検出できた割合
- 再現率を重要視する場合、陽性の検出をしやすくしたくなるため、陽性判断の基準が緩くなる。そのため、本来は陰性なのに陽性と判断される（偽陽性）データが多くなる。

偽陽性のデータが多くなると、適合率は下がることになる。(※)

- 病気診断では、偽陽性は再検査すれば誤診を避けられるため、偽陰性ほど深刻ではない

F値(F-measure): 適合率と再現率の調和平均

$$\frac{2 \times \text{適合率}(Precision) \times \text{再現率}(Recall)}{\text{適合率}(Precision) + \text{再現率}(Recall)}$$

(※) 適合率と再現率はトレードオフ関係になる

混合行列や適合率、再現率は分類タスクの評価に使う。回帰タスクなどのほかのタスクには使えない。

ROC曲線、AUCを用いた評価

正解率(Accuracy)、適合率(Precision)、再現率(Recall)は、二値分類問題のしきい値の設定に依存して評価が変わりやすい、という課題がある。「ROC曲線」は、二値分類のしきい値に依存しない手法。(ROC: Receiver Operating Characteristic)

ROC曲線は、真陽性率と偽陽性率を用いて定義される。二値分類問題のしきい値を少しづつ変化させた際の、真陽性率と偽陽性率をプロットすることで形成された曲線。

真陽性率

$$\frac{\text{真陽性}(TP)}{\text{偽陰性}(FN) + \text{真陽性}(TP)}$$

- 実際に陽性であるもののうち、陽性と検出された割合。
- 再現率(Recall)と同じ

偽陽性率

$$\frac{\text{偽陽性}(FP)}{\text{真陰性}(TN) + \text{偽陽性}(FP)}$$

- 実際に陰性であるもののうち、誤って陽性と判断した割合

AUC

ROC曲線の下の面積をAUCという。

- 完全にランダムに分類する場合、原点から点(1,1)までの直線になり、AUCは0.5になる。

Introduction

- AUCが0.5というのは、あてずっぽうで判断したことと変わらない、という意味（つまり、精度が全くない、という意味）
- AUCは大きいほど精度がよい

点の意味を考えてみる

- 原点(0,0)は、真陽性率も偽陽性率も0という意味。

真陽性率が0ということは、陽性のデータに対して、予測値が陽性と判断されたデータが0件で、すべて陰性（偽陰性）と判断された、ということ。

偽陽性率が0ということは、陰性のデータに対して、予測値が陽性と判断されたデータも0件ということ。

陽性判定の基準を著しく高く設定して、すべて「陰性」と判断するところとなる。

- 点(1,1)は、真陽性率も偽陽性率も1という意味。真陽性率が1ということは、陽性のデータに対して、すべてが「陽性」と判断された、ということ。

偽陽性率が1ということは、陰性のデータに対しても、すべてが「陽性」と判断されたということ。

陽性判断の基準を著しく低く設定して、すべて「陽性」と判断するところとなる。

- 途中の点は、陽性判断の基準（しきい値）を変えながら採取していく。

真陽性率の方が偽陽性率の方よりも高いとAUCは大きな値になる（ROC曲線が上に膨らむ）。

真陽性率の方が偽陽性率よりも高いとはどういう状態だろうか？

陽性データに対して、陽性判定の割合が大きく（正しく判定できている）、陰性データに対して、偽陽性判定の割合が小さい（正しく判定できている）という状態。つまり、正解率（Accuracy）が上がれば真陽性率があがり偽陽性率もあがることになる。

ROC曲線とAUCは、しきい値を変えながらプロットするため、特定のしきい値によらずにモデルの性能を評価できる、という特徴がありそう。

回帰問題の精度評価

回帰問題の精度評価は、予測値と正解のズレ・誤算で評価する。小さいほど精度が良い。

平均二乗誤差 (MSE: Mean Squared Error)

$$MSE = \frac{1}{N} \sum_{i=1}^N N(y_i - y_{0i})^2$$

ここで y_i は予測値、 y_{0i} は正解値。

二乗平均平方根誤差 (RMSE: Root Mean Squared Error)

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N N(y_i - y_{0i})^2}$$

二乗平均平方根誤差は、MSEの平方根を取った値。平方根を取ることで、測定値と同じ単位になるため、直感的に捉えやすくなる。

- 誤差値が二乗されているため、大きな誤算ほど大きな重みが付与されている
- 誤差が大きいデータの影響を受けやすいため、外れ値に対して脆弱である

平均絶対誤差 (MAE: Mean Absolute Error)

$$MAE = \frac{1}{N} \sum_{i=1}^N |y_i - y_{0i}|$$

予測値と正解値の差の絶対値の平均。

誤差を二乗していないため、RMSEに比べると外れ値の影響を受けにくい。

赤池情報量基準 (AIC: Akaike's Information Criterion)

$$AIC = -2\ln L(\theta) + 2k$$

$\ln L(\theta)$ は、**最大対数尤度**といい、**対数尤度関数に最尤推定量**(θ)を優したものの。最尤推定量は、モデルのパラメータについての最も尤もらしい推定量のこと。

- 第1項は、モデルがデータとよく一致しているほど小さくなる。対数尤度は、観測値（正解値）と推定量の残差を用いて計算され、モデルがデータとよく一致しているほど、尤度は大きくなる。尤度の対数を取って、正負を逆転させた第1項は、モデルがデータとよく一致しているほど小さくなる。
- 第2項は、「モデルの複雑さ」を表し、説明変数が多いほど大きい値を取り、説明変数が少ないほど小さい値を取り。（過学習を防ぐペナルティ項のようなもの）

AICは以下であれば良いなどの基準ではなく、相対的な評価として用いられる。複数のモデルを比較して評価する必要がある。

ハイパーパラメータのチューニング

学習前にユーザーが手動で設定するパラメータを**ハイパーパラメータ**という

- ランダムフォレストの弱学習器の数
- ニューラルネットワークの隠れ層の数

ハイパーパラメータは、モデルの複雑さや学習の進行を制御する役割を持っている。汎化性能を高めるために、ハイパーパラメータを調整（チューニング）することがある。

ハイパーパラメータの調整は、**バリデーションデータ（検証データ）**を用いて行われる。

- 学習データを訓練データとテストデータに分けたあと、訓練データを更に、実際の訓練に使うデータと検証データに分ける。

Introduction

- ハイパーパラメータを変えながら、実際の訓練データでモデルの学習を行い、検証データで評価する

チューニング方法としては**グリッドサーチ**と**ランダムサーチ**の2つの方法がある。

グリッドサーチ: 与えられたパラメータの組み合わせを総当たりで試し、ベストな精度を実現する組み合わせを探索する方法

ランダムサーチ: パラメータの値の範囲と施行回数を指定し、ランダムな試行により最適なパラメータを探索する方法

4.1 画像データ・動画データの性質と処理

標本化、量子化、符号化

A-D変換: アナログ画像をデジタル情報に変換すること

標本化（サンプリング） アナログ画像を等間隔の格子状に区切る処理。格子1つ1つが画素（ピクセル）。1つの画素は1つの単色を表す。

- **解像度** 1インチあたりの画素数（ピクセル数）。ppi(Pixels per inch)。
- **ジャギー** 解像度が低いときに、輪郭に現れるギザギザ
- **エイリアシング** 解像度が低いときに現れる本来存在しない縞模様（PCモニタを写真で取るとよく出てくる）

量子化 画素ごとに数値を割り当てる。量子化された値は、画像の色の濃淡（階調）を表す。

量子化レベルが1ビットの場合、各画素は0か1の2つの階調で表現される。グレースケール画像でよくつかわれるのは8ビットで、各画素を白から黒までの256段階で表現する。

符号化 画素が保持している値を2進数にして、コンピュータで扱える形のデータに変換すること。

画像データの構成と性質

- グレースケール画像は「縦×横」の2次元配列（値は8ビットで、ピクセルごとの輝度（明るさ）を表す）
- カラー画像は色という次元が加わって、3次元配列とされている。色はRGBの3つ（チャネルと呼ぶ）が必要で、それぞれ8ビット、合計で24ビットで表現される。

画像データの保存フォーマット

	PNG	JPG	GIF
色数	フルカラー（24bit, 1677万色）	フルカラー（24bit, 1677万色）	256色
メリット	保存を繰り返しても画像が悪化しない	圧縮率を調整することでファイルサイズを小さくできる	ファイルサイズが小さい。動画を作れる
デメリット	ファイルサイズが大きい	保存を繰り返すたびに画質が劣化	色数が制限されている

画像データの典型的な前処理

画像の特徴を抽出しやすくするための処理

例

- カラー画像をグレースケールに変更する処理
- 被写体の色が際立つように、画像の明るさ、コントラスト、色相、彩度などを補正する補正処理
- サイズの変更、統一する処理

フィルタ処理: さまざまなフィルタがある

ノイズ除去: 画像認識では、ノイズがあると、ノイズを重要な特徴として認識してしまうことがある。

ノイズ除去の手法として**平滑化**がある。平均フィルタ、メディアンフィルタ、重み付き平均値フィルタ、ガウシアンフィルタなどがある。平均フィルタは3x3領域の画素の平均をとって、中心画素を入れ替える方法。

リサイズとトリミング: リサイズは拡大・縮小。トリミングは一部を切り落とすこと。

0-1正規化: 画素値を0と1の間に揃えることで学習効率を上げる。

標準化: 平均を引いて標準偏差で割る処理（標準化）。深いニューラルネットワークでは、何層も演算を繰り返すうちに分布が偏ってくることがある（**共変量シフト**と呼ぶ）。そこで、角層において定期的に標準化処理を行うことで、データの分布の偏りを補正する。学習を早くするとともに、過学習を防止する効果がある。

標準化と似た処理として、主成分分析を組み合わせた**白色化**という処理もある。

パディング: 画像認識で使われるCNN（畳み込みニューラルネットワーク）では、1つの層から出力される特徴マップが元の画像よりも小さくなる。画像の縮小を回避したい場合、事前に画像の周りを0で埋める**パディング**を行う。

グレースケールへの変換: 機械学習ではカラー画像をグレースケール画像に変換することがある。

画像処理のためのツール

PythonにはOpenCVやPillowなどのツールが用意されている。

動画データの処理

動画データは画像データの時系列データ。

- **フレーム** 動画を構成する一枚一枚の画像
- **フレームレート** 1秒間あたり何フレームを表示できるか。(fps)
- **エンコード** 映像データと音声データを圧縮すること
- **コンテナ** エンコードした映像データと音声データを1つにまとめる動画フォーマット
- **デコード** エンコードされた映像データと音声データをデコードすることで再生される。
- **コーデック** エンコード、デコードを双方向に行うことができるソフトウェア

フォーマット	特徴
MP4	多様なOSで再生可能。高画質ながら軽量
AVI	Windowsにおける動画の標準フォーマット。高画質だがデータサイズがおおきくなりがち
MOV	MacOSにおける動画の標準フォーマット。高画質だがデータサイズがおおきくなりがち。ビデオカメラにもよく使われる
FIV	YouTubeやニコニコ動画で使用
WebM	Googleによって開発されてウェブ向けの軽量フォーマット

4.2 画像分類と一般物体認識

画像分類はニューラルネットワークが主流。ニューラルネットワークは特徴量を自動的に抽出できることから、非構造化データである画像の分析に向いている。

活用例

- 顔認識
- 製造業の現場における品質管理
- スマート農業

画像分類の技術の進歩

画像認識の最も強力な手法は畳み込みニューラルネットワーク (CNN: Convolutional Neural Network)。

畳み込みニューラルネットワークでは、前段での抽出結果を下流へ入力するということを繰り返すことで、層が深くなるについてれ、高度な特徴を抽出できるという性質がある。

CNNは、畳み込み層、プーリング層、全結合層の3種類のレイヤーから構成されている。

- 畳み込み層 画像から特徴を抽出する
- プーリング層 重要な情報だけを残して情報圧縮する
- 全結合層 抽出された特徴に基づいて、画像分類の結果を確率を表す数値として出力する

畳み込み層では、画像に対してカーネルと呼ばれる正方形行列 (3×3 や 5×5) がフィルタとして適用されることで、元の画像の特徴が強調された特徴マップが作成される。フィルタの処理は、元画像のピクセルとフィルタ行列の積和が計算される。エッジ検出用のカーネル、ぼかし用、シャープ化用など様々なカーネルがあるようだが、学習の仮定で複数のカーネルが訓練され、データに適した特徴を抽出するように自動で最適化される。

プーリング層では、重要な特徴を残しつつ画像の情報量を圧縮する。最大プーリングでは、画像の小領域毎に最大の画素値だけを残す。平均プーリングは、平均値を残す。

一般物体認識

画像認識のなかでも画像分類は、1つの画像に対して被写体が○○というクラスに分類される確率を出力する。

画像の中に複数の物体があるときに、物体の予測を行う技術を 一般物体認識 (Generic Object Recognition) といいう。

一般物体認識は、 物体検出(Object Detection / Localization) と、 画像分類(Classification) の組み合わせとなる。

物体認識の手法としては、以下の方法がある。

- 物体領域を矩形で切り出す手法
- 物体領域を画素単位で精密に切り出す手法 (セマンティックセグメンテーション)
- 両方を組み合わせた手法 (インスタンスセグメンテーション)

画像処理のための計算リソース

GPU(Graphics Processing Unit): 大規模なテンソル (行列やベクトル) 計算のような単純な処理を高速に行うことができ、並列演算処理が得意。GPUの演算能力を画像処理以外に汎用化されたGPGPUも開発された。

CUDAは、NVIDIAが提供しているGPGPUで並列演算を行う開発環境。他にGoogleは、テンソル計算に特化したTPUを開発している。

データ拡張、転移学習

CNN（畳み込みニューラルネットワーク）では、パラメータが数千万個以上あり、言語処理用では数億個、数十億個を超える。

学習データを用意するコストを減らすための工夫

- **データ拡張**: データにランダムな変更（回転や平行移動など）を行うことで、人工的に訓練データのバリエーションを増やす
- **転移学習**: 膨大なデータで訓練した学習済みモデルの汎用的なパラメータ値を別のタスクに応用する。

転移学習における再訓練については以下のようなやり方がある

- (1) 新たに追加した「専用の層」のパラメータのみ訓練する
- (2) 出力層に近い側のパラメータのみ訓練し、上流のパラメータは固定する
- (3) 訓練済みモデルのすべてのパラメータを再訓練する

(2)や(3)、特に(3)は**ファインチューニング**と呼ばれる。時間はかかるが高い精度を出す可能性がある。

画像認識については、多くの学習済みモデルが公開されていて、転移学習に利用可能。（VGG16、ResNetなど）

モデルの軽量化

- **プルーニング**: 枝刈り。比較的に汎化性能への寄与度の低い（ニューラルネットワークなどの）ノード間の接続を切る
- **蒸留**: 大規模モデル（教師モデル）への入力とその出力を用いて、小さいモデル（生徒モデル）を学習する
- **量子化**: パラメータをより小さいビット数で表現する

4.3 音声処理

- **音声生成(Speech Synthesis)**: 人間の音声を人工的に合成すること
- **音声認識(Speech Recognition)**: 音声波形からスペクトル（波形を短時間で切った周波数成分）を作成し、数理モデルを使って内容を推定すること

音声認識の基本プロセス

1. 音声をデジタル情報に変換する
2. 音声波形から周波数や時間変化などの特徴を抽出する
3. 言葉の最小単位である**音素**を特定する
4. 辞書と照合することで音素列を単語に変換する
5. 単語間のつながりを解析して、文章を生成する

音声のデジタルデータの変換

デジタルデータの変換は、画像の場合と同じで**標本化(サンプリング)**、**量子化**、**符号化**の流れになる。

- 標本化: 連続的な音波を一定の時間間隔ごとに切り出す（画像であれば格子状に切り出して画素とした）
- 量子化: 波の強さを離散的な値に近似する
- 符号化: 量子化された値をビット列で表現する

パルス符号変調(PCM: Pulse Code Modulation) は、音声のAD変換でよく用いられる手法。アナログ信号の強度を一定間隔で標本化（サンプリング）し、整数値として量子化し、最後にビット列で表現する。

音声の標本化

- **サンプリングレート**: 1秒間に音波の情報を数値に変換する回数。CDの音声は、**サンプリング周波数44.1kHz**で、**1秒あたり44100回信号を測定し記録すること**を意味している。
- **サンプリング定理**: CDを再生するとき、再生装置では半分の22.05Hzまでしか再現ができない。サンプリング定理とは、AD変換でデジタル信号に変換する際、再現したい信号に含まれる最大周波数の2倍を超えるサンプリング周波数で標本化を行う必要がある、というもの。

人間が20kHzよりも高い周波数は聞こえないとされているため、CDはその2倍を超える周波数で標本化されている。

音声の量子化

量子化ビット数で表されるビット数で数値化される。

CDは16ビット、ハイレゾ音源は24ビット。

音声の符号化

量子化された値をビット列で表現するが、さらにデータ圧縮を行ってデジタルデータに変換する。

音声のデータ量

サンプリングレート、量子化ビット数、音声の長さによってデータ量は決まる。

例) 50kHz、量子化ビット16ビット、30秒の音声のデータ量

$50,000\text{kHz} \times 16\text{bits} \times 30\text{秒} = 24,000,000\text{bits} = 3\text{MB}$

データ量は1秒あたりのデータ量として表すこともある。CDの例では $44,100 \times 16\text{bit} = 705.6\text{kbps}$ になる。CDはステレオ音源のため、左右のデータ量をあわせると倍になる。

音声データの保存フォーマット

- WAV ... 非圧縮のフォーマット。高音質だがデータ量が大きい
- MP3 ... 圧縮されているためデータ量が小さい。非可逆。
- FLAC ... 可逆圧縮を用いる

周波数成分の抽出

高速フーリエ変換は、音声信号を、周波数成分の分布を表す周波数スペクトルに高速に変換する手法。周波数ごとの振幅の大きさに分解する。

スペクトル包絡は、波数スペクトル上で各周波数成分のピークを滑らかに結んだ曲線。色や音声の特徴を表す上で重要な役割を果たす。スペクトル包絡は、短時間フーリエ変換やケプストラム分析などの手法で周波数成分に分析した後、得られたスペクトルの傾向を滑らかに結んで生成される。

フォルマント周波数は、スペクトル包絡においてピークが立っている複数の周波数。

音韻は、言語に依存せずに人の発生を区別できる音の要素。音韻が近ければフォルマント周波数も近い値を取る。

音響モデル

音響モデルとして長く用いられてきたのは隠れマルコフモデル(HMM: Hidden Markov Model)。音素（母音や子音など）ごとに学習を行い、音素列がどの単語に対応するかを判断するために、事前に用意された音素列と単語を対応させた辞書を使ってパターンマッチングを行う。

音声生成の技術

text-to-speech TTSの手法

- 波形接続TTS: 文字回音声の断片の集合体から必要なものを結合して音声を合成する。声を変えたり抑揚や感情を加えることが難しい傾向にある。
- パラメトリックTTS: 音声の波形をモデル化して、文法、口の動き、高さ、抑揚などの特徴に関するパラメータに基づいて波形を生成する

パラメトリックTTSの方が低コストかつ高速に処理が可能だが、人間らしさの観点から波形接続TTSに劣る。

従来の音声合成では、声の高さと音色の2つの特徴を空いていするため、発生メカニズムに基づいた数理モデルが使われていた。数理モデルによる音声合成では、確率的なアプローチを用いていた。代表的なモデルは「混合正規分布モデル」と「隠れマルコフモデル」である。

2010年代以降、ニューラルネットワークを用いたモデルが普及。Microsoftが発表したDNN-HMMは、HMMとDNNを組み合わせたもの。その後、ディープラーニングを用いた音声処理ネットワークが主流に。2016年にDeepMindが発表したWaveNetは音声合成にブレークする一をもたらした。

4.4 時系列分析

時系列データ分析の活用

- ・時間とともに変化するデータ
- ・一定の間隔で取得したデータ
- ・連続的に記録するものもあれば、離散的に記録するものもある

例

- ・音声、動画
- ・各種センサーの測定値
- ・気温や降水量などの気象データ
- ・売上推移、株価変動

時系列データ分析の例

- ・需要予測
- ・異常検知
- ・文章生成

時系列データの構成成分

大きく分けて、ノイズ、周期性、トレンドの3種類

ノイズ 一般的に、分析によって抽出したい情報と関連しない不要な情報をノイズという。

時系列データ分析では、解析の目的によってどのような成分がノイズに該当するかが変わる。長期的な傾向を知りたいときは、短期的な変動がノイズになりうる。ノイズの影響を減らすために、移動平均がよく使われる。短期的な変動のパターンを捉えたいときは、長期的な傾向をノイズとして除くことになる。

トレンド 一般的に、細かな変動やノイズを除いた、長期にわたって持続的に上昇する、ならかな傾向を指す。一般には、季節性の周期性の変動よりも周期が長いのが特徴。

周期性・季節変動 周期性とは、一定期間ごとに繰り返される変動パターンのこと。

循環変動 周期的に繰り返されるが、周期が一定ではないものを言う。例えば、景気指数変動は好況、不況の繰り返しだが、周期は一定ではない。

短期的な変動と差分系列 時系列データから周期性やトレンドを除いた後に残る局所的で細かい変動が**短期的変動**。時系列データを**差分系列**（ある時刻と1つ前の時刻のデータの差分、階差）に変換することで、長期的な変動が除去され、局所的な変動成分を抽出することができる。差分系列をとると、急激な変動による外れ値をみつけるのにも有効。

移動平均系列

移動平均 (Moving Average) は、区間を移動しながら一定区間ごとに平均値を計算して得られる系列。平滑化やスムージングとも呼ばれる。

- ・移動平均から得られる移動平均系列は、長期的なトレンドを見やすくできる
- ・区間の幅を**ウィンドウ**や**ウィンドウサイズ**と呼ばれる。ウィンドウサイズを広く取りすぎると、本当に捉えたい傾向まで平坦にしてしまうため注意が必要。

- **単純移動平均**と呼ばれるものは以下の種類がある
 - 中央移動平均: 該当時刻とその前後の時刻を使う
 - 後方移動平均: 該当時刻とそれ以前の時刻を使う (tとt-1で平均を取る)
 - 前方移動平均: 該当時刻とそれ以降の時刻を使う (tとt+1で平均を取る)
- 遠い過去よりも直近の値に比重を置く**指數平滑移動平均**もある

定常性と自己相関

定常性 時系列における定常性とは、**平均と自己共分散が時間の経過とともに一定である**という性質。定常性を持たない系列は、**非定常仮定**に従う系列という。

時系列分析モデルによっては定常性のデータに対して使用可能であるため、定常性を満たすようにデータを変換（差分変換や対数差分変換）することがある。

自己相関: 事例列データのある時刻の値と、その時刻から一定時間ずらした時刻の値の間に相関関係があることをいう。

精度評価の注意点

機械学習の精度評価では、学習データを訓練データとテストデータに分割して評価するが、**時系列分析の場合、データをランダムに分割すると精度を正しく評価できないことに注意が必要。**

時系列データ分析では、過去のデータを用いて未来のデータを予測するため、訓練データとテストデータを分割するときも、過去のデータと未来のデータになるよう分割する必要がある。

時系列分析の具体的な手法

4つの手法を紹介

- **MA (移動平均, Moving Average)**
- **AR (自己回帰, Auto Regressive) モデル**
- **ARMA (自己回帰移動平均, AutoRegressive Moving Average) モデル**
- **ARIMA (自己回帰和分移動平均, AutoRegressive Integrated Moving Average) モデル**

MA (移動平均, Moving Average) : 移動平均モデルは、現在の値を、過去のホワイトノイズ項（ランダム誤差）の線形結合として表現するモデル。

一次MAモデル MA(1) では、以下の式になる。

$$y_t = c + r_t + \theta_1 r_{t-1}$$

r_t は、時点 t において発生するホワイトノイズ（ランダムな誤差情報）を表す。

時点 t における値 y_t は、時点 t におけるホワイトノイズ r_t と、1 時刻前のホワイトノイズ r_{t-1} から算出される、という式になっている。パラメータ θ_1 は、1 時刻前のノイズの影響を受ける程度。 c は定数項となっている。

1 次MAモデル MA(1) は、「現在の時刻は 1 つ前の時刻での誤差の影響を受ける」というモデルである。

q 次MAモデル MA(q) は、「現在の時刻は q 時点前以降の誤差の影響を受ける」というモデルで、以下の式で表される。

$$y_t = c + r_t + \sum_{i=1}^q \theta_i r_{t-i}$$

この式は、過去の各時刻でのホワイトノイズをパラメータ θ_i で重み付けした和が現在の値に影響を与えていると会社kうできる。

AR（自己回帰, Auto Regressive）モデル: 自己回帰モデルは、一定期間に遡る過去の自分のデータの線形結合で表される。過去のデータに対して回帰を行うことで現在の値を予測する。

1次ARモデル AR(1)の式は次のようになる。

$$y_t = c_0 + c_1 y_{t-1} + r_t$$

ここで r_t はホワイトノイズを表している。

c_0 と c_1 は、通常の回帰分析と同様に最小二乗法で求めることができる。

q次ARモデル AR(q)は以下になる。

$$r_t = c_0 + \sum_{i=1}^q c_i y_{t-i} + r_t$$

ARMA（自己回帰移動平均, AutoRegressive Moving Average）モデル: ARとMAを組み合わせたモデル。

p次ARモデルとq次MAモデルを組み合わせたARMAを、ARMA(p,q)モデルと記述することが多い。ARMA(p,q)の式は次のようになる。

$$y_t = c + r_t + \sum_{i=1}^p c_i y_{t-i} + \sum_{i=1}^q \theta_i r_{t-i}$$

ARIMA（自己回帰和分移動平均, AutoRegressive Integrated Moving Average）モデル: ARMAに和分モデルを加えたモデル。

和分は、時系列データの階差を何回取れば定常になるかを意味する。時系列データが非定常な場合、ARMAモデルは適用できない。ARIMAモデルでは、差分を取ることによって定常過程に変換し、ARMAモデルを適用可能な形に変換する。

SARIMAモデル: ARIMAの発展版に **SRIMAモデル（季節変動自己回帰和分移動平均モデル, Seasonal AutoRegressive Integrated Moving Average）** がある。ARIMAに季節性の周期変動を加えたモデルになっている。SARIMAでは、データを季節周期に分割し、その差分を取り、そのデータに対してARMAモデルを適用する。

4.5 状態空間モデル

状態空間モデルも時系列データ分析のモデル。

従来のモデルでは、観測値のみ着目し、現在の観測値を過去の時刻の観測値から予測するという考えに基づいている。

状態空間モデルでは、データの生成過程を状態と観測値の2種類の系列としてモデリングし、状態から観測値を予測する。

状態を表すモデル: $s_t = f_t(s_{t-1}, u_t)$

観測値を表すモデル: $y_t = g_t(s_t, r_t)$

時刻tにおける状態 s_t は、1時点前の状態と状態ノイズ u_t の関数として表される。

時刻tにおける観測値 y_t は、同時刻の状態 s_t と観測されるノイズ r_t の関数として表される。

状態を、以前の状態と状態の誤差（ノイズ）の関数とし、観測値を、状態と観測の誤差（ノイズ）の関数と考えている。

ここでのノイズは、平均が0の正規分布に従うと考える。

カルマンフィルター

カルマンフィルターは、状態を効率的に推定するための手法。

状態空間モデルの具体例

みかんの糖度は寒暖差や降雨量といった天候要因によって影響されることが知られている。単純化のため、「状態」を「天候の変化」とし、「観測値」を「みかんの糖度」とする。

みかんの糖度については測定が可能。要因となる天候の変化について、状態空間モデルでは、「天候」と「観測の誤差」に分けて考える。天候は観測しづらいものだが、状態空間モデルを用いて、状態である天候についての推定を行う。

状態空間モデルの強み

- 非定常なデータを扱える
- モデルの次数を固定しなくてよい
- 欠損値を含むデータも扱える（直接的に観測値を推定しないため）
- 観測値の変化の要因を説明できる

4.6 自然言語処理の基礎

一般的な流れ

1. 形態素解析 文章の意味を持つ最小単位（形態素）に分割し、品詞を推定。MeCab、JUMAN、Janomeなど
2. 構文解析 形態素間の関係性を解析。主語、目的語、述語などの係り受け構造を推定。CaboCha、KNP、GiNZAなど。
3. 意味解析 同じ文の中の意味構造を見出す。文法的な要素や文中にある単語が別の単語と関連性が高いかを判断
4. 文脈解析 複数の文の意味や関係性を解析。代名詞の特定など。
 - 照応解析 文章内の代名詞などの照応表現が指している場所を推定する。
 - 談話構造解析 因果と背景などを解明することが目的。

自然言語処理のためのデータ前処理

- データクレンジング
 - 名寄せ ... 全角と半角など言葉のゆらぎを統一
 - ストップワードの除去 ... a や the、句読点の、や。の除去
 - ステミング ... 語幹の抽出 (playing, played -> play)
- Bag-of-Words ... 単語の出現頻度を考慮しながら数値ベクトルに変換する
- TF-IDF ... 単語の重要度を付与する

文章を数値化する手法について、当初はOne-hotエンコーディングが使われていたが、現在はディープラーニングを用いた高度な数値ベクトル化手法であるWord2Vecが主流。

代表的な言語タスク

- 機械翻訳(Machin Translation) 翻訳
- 感情分析(Sentiment Analysis) テキストからポジ／ネガの感情を特定する。
- 文章要約(Summarization) 与えられた文章やテキストの主要なポイントを抽出し、それを短い要約にまとめる
- テキスト分類(Text Classification) 文章を 1 つまたは複数のカテゴリに分類する
- 固有表現検出(NER: Named Entity Recognition) 文章から、人名、組織名、地名、日時表現、金銭表現などの固有表現を特定する
- 質問応答(Question Answering) 特定の質問に対する答えを出力する選択問題や、文章から問題文の答えを抜き出す機械読解、対話形式の質問応答などがある
- 意味的類似度(Semantic Similarity) 2つの文が同じ意味かどうかを判定する
- 自然言語推論(NRI: Natural Language Inference) 2つの文の間の論理的な関係を推論する。矛盾があるか、他奥を含意するかを判定。

GLUE

GLUE (General Language Understanding Evaluation) は、言語タスクのベンチマーク（評価基準）の 1 つ。2022年には日本語版のJGLUEも開発されている。

9つの公開されている言語理解タスクから構成されている。一般公開されている様々なデータセットを組み合わせてベンチマークが作られていて、これを用いて言語能力のスコアを算出する。

例

Introduction

- CoLA: 文が英語文法として正しいかどうかを判定。
- SST-2: 映画レビューの感情分析(文章分類)
- QNLI: 質問とその答えのペアが与えられ、答えが質問から論理的に導かれるかどうかを判断（テキスト分類）
- MRPC: オンラインニュースからの類似度判定
- SQuAD: ウィキペディアから質問の答えとなるテキストを見つける（質問応答）

SuperGLUEは、更に難易度が高いものとして導入されている。

4.7 大規模言語モデル

- Transformer ... 文章解析用のニューラルネットワーク
- BERT ... Googleから公開された大規模言語モデル
- GPT ... OpenAIから公開された大規模言語モデル
- 基盤モデル

事前学習とファインチューニング

- 事前学習は、教師なし学習として実行される。

インターネット上の大量のテキストデータを学習。自動的に統計的なパターンや意味を抽出する。

- コーパス ... 学習用のテキストデータ。数十億件のトークンに及ぶ

事前学習を終えた学習済みモデル（ニューラルネットワーク）は、隠れ層で、多くの文章に共通する汎用的な特徴量を習得した状態になる。

新しい言語タスクへの転移学習に利用することができる。

ファインチューニングは、事前学習済みのモデルのパラメータを調整することで、特定のタスクに適用できるようになること。特定のタスクに特化したデータセットを一定量用意し、主にネットワークの下流の部分のパラメータのみ、追加学習と値の更新を行う。

事前学習が教師なしであったのに対して、ファインチューニングは教師あり学習。通常、ファインチューニングに必要なデータは数百～数千と、事前学習に比べると少ない。

スケール則

モデルのパラメータ数と訓練データの計算量が増加するにつれて、モデルの性能もほぼ同じ割合で向上するという経験則

GPTモデル

Generative Pre-trained Transformer の略。

GPTの事前学習は、与えられた文章（単語系列）の次に来るべき単語を予測し、文を自動完成できるように基礎訓練されている。事前学習は教師なし学習（自己教師学習）を行う。ランダムな一文の後半を隠し（マスキング）、前半部分から後半の単語を当てる穴埋め問題を大量に解くことでパラメータが最適化される。

GPTは、数少ない事例を与えるだけで、次に来るべき単語を逐次的に予測しながら、文章を自動的に完成できることが特徴。人間に近い自然な文章を生成できる。

GPTにファインチューニングを適用することによって、文章の生成だけでなく、翻訳、質疑応答、校正、ブレインストーミング、自然言語からソースコードを生成するなど様々なタスクに使用することができる。

GPTは基盤モデル、ChatGPTはGPTを基盤モデルとした対話型の文章生成AI。

大規模言語モデルの背景にある技術

Transformer以前は、**RNN（リカレントニューラルネットワーク、再帰型ニューラルネットワーク）** が自然言語処理に用いられていた。RNNは時系列データを扱えることを特徴としたニューラルネットワーク。RNNには「フィードバック」という仕組みを持つことで、過去の隠れ層と現在の隠れ層に繋がりを持っている。こうすることで、過去の情報に基づいて、前の時刻の情報を記憶し、それらを考慮しながら現時刻の情報を処理することができる。

2つのRNNモデルを繋いだモデルを**Sequence-to-Sequence(Seq2Seq)** という。時系列データを入力して処理し、時系列データを出力することからこの名前がついている。

RNNを用いた言語モデルには課題があった。

- 入力系列が名が唸ると、遠く離れた単語間の関係や文脈を正しく把握できなくなる
- 入力データを1ステップごとに処理する必要があり、並列処理ができない

2017年、Googleの研究チームが開発した**Transformer**の登場により、上記の問題が改善され、長い文章の解析精度が飛躍的に向上した。背景にあるのは、内部にもつ**Attention機構**が重要な役割を担っている。

RNNは、過去のある時刻の情報が現在の予測に"どれだけ影響するか"を算出することができない。これに対して、Attention機構は、入力された系列の各単語が、他のすべての単語とどの程度関連しているかを計算する機能を持っている。重要度の高い情報に注意を向けて学習できるようになっている。

4.8 生成AI

テキスト以外にも、画像、動画、音声などがあることに注意（つい、ChatGPTの大規模言語モデルによるテキスト生成を思い浮かべるが）

生成AIの特徴

従来のAI（機械学習）は、回帰、分類といった、学習データから特徴やパターンを見つけ、新しいデータに対する予測を行う、というもの。出力されるのは予測値であって、新しい形でデータを作り出すことではない。

生成AIも、値の予測をしている。GPTは単語を逐次的に予測することで文を完成させる。画像生成AIであれば、画像のピクセルの値を予測している。

生成AIは、実質的に予測を行うものの、新しいデータを生成することを目的としているという点で、従来の機械学習とは異なるといえる。

主なサービス

- 画像生成AI
 - Style-GAN(NVIDIA)
 - DALL-E (OpenAI)
 - Imagen (Google)
 - Stable Diffusion (Stability AI)
 - Midjourney (Midjourney)
- 音声生成AI
 - Jukebox (OpenAI)
 - MusicLM (Google)
- 大規模言語処理AI
 - GPT, ChatGPT (OpenAI)
 - PaLM 2, Bard, Gemini (Google)
 - LLaMA (Meta AI)

生成AIに命令を与える

プロンプト、プロンプトエンジニアリング

- Zero-shot prompting
- Few-shot prompting
- Chain-of-Thought prompting
- 知識生成prompting ... プロンプトの一部に知識や情報を組み込む
- 方向性刺激prompting ... 正しい方向に誘導するためのヒントを与える

文章生成AI

ChatGPTなど

ChatGPTのファインチューニングは以下の3つのステップで行われる

- 教師あり学習
- 報酬モデルの学習

- 強化学習

RLHF (Reinforcement Learning from Human Feedback)。Reinforcement Learningは強化学習のこと。RLHFは人間のフィードバックに基づいた強化学習。

画像生成AI

変分オートエンコーダー(VAE: Variational Auto-Encoder) は生成モデルの一種。入力データの背後にある潜在的な構造やパターンを学習し、新しいデータを生成するために使われる。エンコーダーとデコーダーの2つのネットワークから構成され、エンコーダーは入力データを潜在空間と呼ばれる低次元の潜在変数に変換する。デコーダーは、潜在変数から元のデータ空間に近いデータを再構成する。サンプリングされた潜在変数から元データに近い新しいデータを生成する役割も果たす。

敵対的生成ネットワーク(GAN: Generative Adversarial Network) は、データを生成するための深層学習モデルで、2つのニューラルネットワークが互いに競い合いながら学習する仕組みを持っている。

Diffusion Modelは、「データをノイズで破壊し、元のデータに復元する」という手法で生成を行う。学習時は、データにノイズをジョジョに加えていく「拡散過程」と、その逆の「生成過程」を学習することで、ノイズからリアルなデータを精製できるようになる。初期状態のノイズから段階的にノイズを除去することで、最終的にはデータを精製するという手順。

画像生成の分野ではDiffusion Modelが急速に採用されつつある。GANやVAEよりも精製品質が高いと評価されている。

4.9 機械学習における解釈性

モデルの解釈性とその応用

ブラックボックス性 … モデルの判断の根拠が解釈しにくいこと

大域的な説明(Global Surrogate) … モデル全体の各特徴量の寄与度

局所的な説明(Local Surrogate) … 予測対象の 1 つのサンプル単位での各特徴量の寄与度

局所的な説明

特定の 1 つの入力データに焦点をあてて、そこで得られた予測結果や予測プロセスを説明する手法。

対象サンプルの周辺データに対するモデルの挙動を、別の単純で可読性の高いモデル（線形回帰モデルなど）を利用して説明しようとする。

代表的な技術に **LIME (Local Interpretable Model-agnostic Explanations)** と **SHAP (SHapley Additive exPlanations)** がある。どちらも、特定のデータサンプルに着目し、単純で解釈しやすい線形回帰モデルで近似することで、予測に寄与する因子を推定するツール。

大域的な説明

学習済みモデルがどのようにして予測をするのかをモデル全体の単位で説明する手法。

大域的な説明でも、解釈しやすいオデルで説明する。（説明変数はデータ、目的変数はブラックボックスモデルの予測）。例えば、回帰分析モデルで金獅子、回帰分析モデルを解釈することで、元のニューラルネットワークの解釈を試みる。

画像認識モデルに対しては代表的な技術として**Grad-CAM**がある。CNNモデルの勾配情報を活用し、画像認識を行うモデル全体に対して予測根拠を可視化することを目指す手法。CCNが分類において注文していると推定される範囲をヒートマップで表示することができる。

5.1 データ収集

ウェブサイトからのデータ収集

- クローリング ... HTMLファイルを取得すること
- スクレイピング ... HTMLファイルから必要なデータを取得すること

アプリケーションからのデータ収集

APIを公開している場合があり（例：X）

- REST
- SOAP

オープンデータ

機械判読に適した二次利用が可能なデータセット。定められた規約に従えば無償で利用ができる。

5.2 データ保管

データの形式

構造化データに対しては、CSVやTSVがよく使われる。

ストレージ

非構造化データのように形式が統一されていない様々なデータを保管する場合は、ファイルストレージにデータを保存する。

クラウドストレージだとAmazon S3、Google Cloud Storage、IBM Cloud Object Storageなど。

オンプレミス型では、FTPサーバーやファイル共有サーバーなど。

データベース

リレーションナル・データベースと非リレーションナル・データベースがある。

RDB

テーブル形式にデータを保管するデータベース。SQLを用いて操作する。

NoSQLデータベース

テーブル形式でない多種多様なデータベースをまとめて非リレーションナルデータベースと呼ぶ。キーバリュー型、階層型、ドキュメント指向型など。

データベースの分散処理

HadoopとSpark

Hadoopは、分散ファイルシステムHDFSと分散処理フレームワークMapReduceから構成される。巨大なデータを処理するのが得意だが、反復的なデータ処理は不得意。

Sparkは、Hadoopの後継の分散技術で、ビッグデータや機械学習など、大規模なデータを扱うことにより長けた分散処理フレームワーク。

データの流れ

データソース -> データレイク -> データウェアハウス -> データマート

データレイク 加工されていない生のデータを一元管理するストレージやデータベース。

データウェアハウス データレイクのデータを加工して構造化データとして保存するデータベース。ペダバイト規模のデータに対して高性能なSQLが実行可能。代表的な製品にはOracle Exadata、IBM Integrated Analytics System、Teradataなどがある。

データマート 具体的な活用目的に特化した形でデータを保存する場所がデータマート。データウェアハウス内に保存されることもあるが、データウェアハウスよりも規模が小さく、目的に特化した集計済みデータが保存される。

クラウドサービスの例

Introduction

	Amazon AWS	Google Cloud Platform	Microsoft Azure
データレイク	S3	Cloud Storage	Data Lake Storage
データウェアハウス	Redshift	BigQuery	Synapse Analytics
データマート	RDS	Cloud SQL	Azure SQL Database

BIツールを使って、店舗とECサイトのそれぞれの担当者が、売上や顧客層などのデータを確認したいとする。店舗とECサイトでは、取得しているデータの種類や使いたいBIツールが異なることも考えられる。こうした場合、それに特化したデータマートを作成することが適切と考えられる。

5.3 データ加工

テーブルの正規化

第1正規形 ... 1つのセルには1つの値しか含まれない状態。Excelのセル結合はだめな状態

第2正規形 ... 複合キーがあるテーブルで、一部のキーで一意に定まる列がある状態を部分関数従属という。この部分関数従属を別のテーブルに分割したものを第2正規形という。（例：注文情報から商品の情報を商品マスターに追い出す）

第3正規形 ... 主キー以外の列が決まると一意に定まる列がある状態を遷移関数従属という。これを別テーブルに分割したもの第3正規形という。

- 第1正規形 ... 1つのセルに1つの値にする
- 第2正規形 ... 複合キーによる部分関数従属をテーブル分割する
- 第3正規形 ... 非主キーによる遷移関数従属をテーブル分割する

ER図

- エンティティ
- アトリビュート
- リレーション
- カーディナリティ

SQL

- DDL ... CREATE, ALTER, DROPなど
- DCL ... GRANT, REVOKEなど
- DML ... SELECT, INSERTなど

UNIONは重複が排除される、UNION ALLは重複が排除されない

正規表現

正規表現

5.4 データのセキュリティ

セキュリティの3要素

- **機密性** 許可されたユーザーだけがデータにアクセスできることを保証する。

例: パスワード認証、アクセス権限制御、暗号化

- **完全性** データが不正に改ざんされておらず、正確で完全であることを保証する。

例: 電子署名

- **可用性** データに対してアクセスを許可されたユーザーが要求したときに、いつでも利用可能であることを保証する

例: システムの二重化、バックアップ

マルウェア

セキュリティの3要素を脅かすような悪意のあるソフトウェアの総称をマルウェアという。（ウィルスはマルウェアの一種）

アクセス制御

認証、認可

暗号化

公開鍵暗号は公開鍵で暗号化、秘密鍵で復号。 電子署名は逆で、秘密鍵で署名生成、公開鍵で署名検証。

バックアップ

- フルバックアップ
- 差分バックアップ
- 増分バックアップ

5.5 データ分析

プログラミングの基本的な概念は問われる。

フローチャート

記号の意味はJIS X 0121で定義されている。

データ型

文字コード

オブジェクト志向

デバッグ

テスト

ホワイトボックステスト、ブラックボックステスト

カバレッジ

- 命令網羅: すべての命令を少なくとも1回は実行している
- 判定条件網羅: 条件分岐のすべてをテストしている
- 条件網羅: 複数条件があるときに、それぞれの条件の分岐をすべてテストしている
- 判定条件・条件網羅: 判定条件網羅と条件網羅をどちらも網羅している
- 複数条件網羅: 判定条件のすべての可能な結果の組み合わせを網羅し、かつ、すべての命令を少なくとも1回は実行している

データ分析を助けるサービスや手法

- BIツール
- ノーコードツール、ローコードツール (PowerApps, AppSheet, Honeycodeなど)
- AutoML ... コードを書くことなく機械学習を可能にするサービス。モデルの選定やハイパーパラメータの調整が不要だったり自動化・簡略化されている。Azure Machine Learning、Google Cloud AutoML、Amazon SageMaker Autopilotなどがある。
- 機械学習のマネージド・サービス
- MLOps 機械学習を運用してビジネスに活用していく手法
- AIOps AI技術をビジネス、特に、IT運用管理に用いる手法を指す

6.1 データ活用プロジェクトの進め方

ビジネス力

- 現状の把握 … クライアントから経営課題をヒアリングし理解する
- 仮説の立案 … 解決すべき課題を具体的な分析業務に落とし込む
- データを保管し分析するための環境や基盤を整える
- 仮説の検証 … データを分析する（データ収集や加工も含む）
- ドキュメンテーション・プレゼンテーション … 分析結果を報告書にまとめ、わかりやすくプレゼンする
- 開発手法 … データを適切に分析する手法、それを実行するシステム、あるいは、データ分析に基づいて提供するサービスのためのシステムを開発する
- データ利活用システムの運用・保守

データ活用プロジェクトの一般的な進め方

1. 課題定義と仮説立案
2. 仮説検証
3. 仮説結果の評価と報告

6.2 現状の把握

現状把握の切り口

- ・ 業界構造: 業界に応じたサプライチェーンの構造や顧客の特性
- ・ 市場規模: 業界全体の市場規模とのその企業のシェア
- ・ 経営課題: リピーター率の向上や離職率をセグメントなどの解決・達成したい課題
- ・ 経営指標: 売上、成長率、利益率などの各分野で用いられる指標
- ・ ビジネスモデル: 誰に、何を、どのように、提供して価値を生み出すか、なぜ価値を生み出せるのか
- ・ プレイヤー: 事業に関わる企業や人
- ・ ポジショニング: 競争優位を保つまでの強み、リスク

市場規模

- ・ 政府や官公庁が発行する統計データ
- ・ 業界団体発行のデータ
- ・ シンクタンクやコンサルティングファームが公表しているデータや分析データベース

経営課題や経営指標

財務指標

指標	算出方法	意味
ROE (自己資本利益率)	当期純利益÷自己資本	株主が投下した資金でどれだけ利益が上がったか
ROA (総資本利益率)	当期純利益÷総資産	総資産にたいして どの程度利益が上がったか
PER (株価収益率)	株価÷1株あたり純利益	(収益面から見て) 株価が割安か割高かの判断材料
PBR (株価純資産倍率)	株価÷1株あたり純資産	(資産面から見て) 株価が割安か割高かの判断材料

ビジネスモデルとプレイヤー

ビジネスモデルを見るときはWho、What、Why、Howの4要素を見る

- ・ Who (誰に) : 既存顧客はどんな特徴があるか、開拓したい新規顧客のターゲットとなるのはどんな層か
- ・ What (何を) : 顧客にどのような価値を提供するのか、自社の強みとはなにか
- ・ Why (なぜ) : なぜそれが収益をもたらすのか
- ・ How (どうやって) : どのように価値を提供するのか。競合他社にどう勝てるのか

プレイヤー ... ビジネスに関わる人全般

STP戦略

セグメンテーション、ターゲティング、ポジショニングの略

因子分析 製品やサービスを、価格、スペック、消費者のもつ印象などの複数の軸で、類似度に基づいてマッピングする手法。マップ上の距離が近いものは類似度の近い製品・サービス。

コレスポンデンス分析 ポジショニングマップ上に、商品イメージとブランドをマッピングするもの。商品のイメージや商品同士の類似性を理解しやすくなる。

情報の入手

- **一次情報** 直接体験する、アンケート調査、ヒアリング、実験を行うことで得られるオリジナル情報、研究成果、学術論文など
- **二次情報** 一次情報を所有するソースから取得した情報、一次情報を編集加工した情報、書籍やウェブから得られた間接的な情報。官公庁や研究機関から公開されている情報も二次情報にあたる。
- **三次情報** 情報源が定かではない情報。

6.3 課題の定義

スコーピング

プロジェクトの開始時に、優先順位、予算、納期、技術的難易度を考慮した上で、実現可能な目標やスケジュールを設定する必要がある。

- 要求 やりたいこと
- 要件 やること

ロジカルシンキング

- MECE もれなくだぶりなく。ロジックツリーによる可視化が効果的
- So What? / Why So? だから何なのか、なぜそうなのか。

ファイブフォース分析

競合他社や業界構造を把握することで、自社の置かれた立場を検証し、収益性が高いか否かを把握する手法。

1. 業界内での競争
2. 業界への新規参入者
3. 代替品の存在
4. 売り手の競争力
5. 書い手の競争力

RFM分析

顧客をグルーピングするための手法。

- R = Recency : 最後にいつ商品を購入したか
- F = Frequency : どのくらいの頻度で商品を購入しているか
- M = Monetary : これまでの合計購買金額

KPI、KGI

- KGI: 最終目標を数値で指標化したもの
- KPI: 中間的な数値目標

KGIを頂点として、KPIツリーを作る

仮説の立案と検証

仮説立案 -> 分析 -> 実験 -> 仮説検証 -> 報告

仮説立案

- ビジネス課題と因果関係があると考えられる要素を抽出する
- 目標を具体的な数値で設定する
- 定量的な目標が設定できない場合は、まずはアイデア出しから出発してみる

Introduction

- 手元にある程度のデータがある場合、**探索的データ分析(EDA)** という予備的な分析を行う。探索的データ分析とは、機械学習や統計モデリングを行う前にデータの特性や傾向を確認するプロセス。平均、中央値、分散、標準偏差などの要約統計量を確認し、ヒストグラム、散布図などで可視化する。特徴量同士の相関については散布図行列などを使って確認する。

分析

- 仮説に基づいて必要なデータと手法を検討する
- 基礎統計量を算出し、可視化して、全体の傾向をつかむ
- 複数の種類の分析モデルを試し、データやアルゴリズムにバイアスがないかを確認しながら最適なモデルを見つける

実験

- 分析の結果得られた示唆をもとに具体的な施策を実行する

仮説検証

- KPIに基づいて、実験の結果を評価する
- データ分析の精度向上のために手法や利用データを再検討する
- 実行した施策を拡大させた場合の開発お運用コストを見積もる

ドキュメンテーション、結果の考察と説明

報告書の項目例

- 分析の背景と課題の定義
- 分析のアプローチ
- 分析結果
- 考察と示唆
- ネクストアクションの提案

6.4 契約の種別と開発の形式

請負契約、準委任契約

	請負契約	準委任契約
義務	成果物の完成	業務の遂行
	契約不適合責任	善管注意義務
報酬を請求できるタイミング	成果物の納品時	業務終了時
報告義務	なし	あり
成果物の著作権	受託者	受託者

- **契約不適合責任** 納品されて成果物の種類、品質、数量が契約内容に適合していない場合、受託者が責任を負う
- **善管注意義務** 善良な管理者の注意義務。この義務が果たされていない場合、損害賠償請求が可能となる。

ウォーターフォール開発、アジャイル開発

2つの開発スタイルがある

7.1 法律・倫理をなぜ学ぶ？

倫理とはなにか

- 社会で行動する際に、良し悪しを判断するための根拠
- 社会の秩序を保つために、社会の一員として生活を営む上でのルール

法律と倫理を学ぶ理由

自分の権利を守るため、また、社人の権利を侵害しないため。

企業・社会におけるAI倫理の問題

AIのバイアス問題

AIが人種やジェンダーを軸とする差別的な結果を出力してしまった事例

- Amazonの人材採用システムによる女性差別
- Googleフォトによる黒人の差別的な自動タグ付け

これらは、AIサービスを提供する企業の企業倫理が問われる。

顔認識システムにおける差別の問題

学習データに含まれるバイアスにより、顔認識システムにおいて非白人の認識精度が相対的に低いことが問題視されることがある。これを踏まえ、大手IT企業のGoogle、Amazon、IBM、Microsoftは汎用顔認識技術からの撤退を表明し、警察など法執行期間への顔認識技術の提供を中止すると発表している。

現時点では、顔認識技術は人種差別、性差別を助長するので社会に害を及ぼしうるということが主流の考え方になりつつある。

AIが事故を起こした際の責任体制

AI搭載のシステムが事故を起こした場合、誰が責任を取るのかは法整備がなされていない。

社会に有益なデータ利活用に向けて

Society 5.0 … 内閣府が定めた用語。「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」と定義されている。

7.2 データ倫理

データ倫理を考える意義とは

データ倫理は以下の3つの軸からなるとされている

- データの倫理
- アルゴリズムの倫理
- 実践の倫理

データ倫理とは、データを倫理的に活用するための知識体系、といえる。

データの収集・加工・活用に伴う具体的な注意点

以下は不正行為に該当する

- データの捏造
- データの改ざん
- データの盗用
- データ汚染(Data Poisoning) ... データに微小な変更（摂動）を加えることで、意図的に分析や予測の結果に間違いを起こさせること
- 敵対的攻撃 ... データ汚染は学習データを変更することだが、敵対的攻撃は、学習済みモデルが誤った出力をするように、特定の入力データを恣意的に加工する行為。加工したデータを敵対的サンプルと呼ぶ。

7.3 データ倫理に関するガイドライン

国内外の倫理ガイドライン

- GDPR (欧州一般データ保護規則) ... データ倫理にいち早く着目したガイドラインの代表例
- IEEE Ethically Aligned Desing (EAD)

IEEE (米国電気電子学会) がAIに関する倫理的課題について検討するために作成した報告書。

- Ethics Guidelines for Trustworthy AI (信頼性を備えたAIのための倫理ガイドライン)

EUのAIハイレベル専門家会合によって発表。最終的には国際的なAIガイドラインに発展させる方針。

- 人間中心のAI社会原則

内閣府による「人間中心のAI社会原則検討会議」において原案が公開されたもの。

- Partnership on AI

2016年にMeta、Amazon、Google、IBM、Microsoftの5社によって創立された非営利団体。

民間企業の取り組み

企業内にデータ倫理の専門チームを設け、自主規制のためのAI Principles(AIP, AI原則)を整備するなどを行う企業もある。

ELSI (論理的・法的・社会的な課題)

ELSI (Ethical, Legal and Social Implications) は、科学技術が及ぼす倫理的、法的、社会的な影響を一体として検討する試み。

7.4 個人情報とプライバシー

個人情報の定義と種別

個人情報とは特定の個人を識別するのに使える情報のことで、次の2つの要件を満たすもの

1. 生存する個人に関する情報であること
2. 特定の個人を識別・特定できる情報であること

他の情報と容易に照合することができ、それによって個人を識別できる情報も含まれる。

名刺も個人情報に該当する。メールアドレスについても、アドレスから個人名と組織が特定できる場合は個人情報とみなされる。

一方、プライバシーとは、「個人や仮定内の私事・私生活。個人の秘密。また、それが他人から鑑賞・侵害を受けない権利」を指す。

個人データ・保有個人データ

個人データとは、データベース化された個人情報のこと。

保有個人データとは、自社が保有し、開示などの権限を持つ個人データのこと。

データベース化されていない個人情報は個人データではない。

業務委託を受けて個人データを使用している場合、自社が個人データを保有しているわけではないため、保有個人データではない。

個人識別符号

個人情報のうち、単独で特定の個人を識別可能な文字・番号・符号を個人識別符号と呼び、以下のいずれかが該当する。

- 生存する個人の身体的な特徴に関する情報

例: DNA、顔、指紋、指紋、歩行の様

- 個人に割り当てられる符号

例: 公的な番号（パスポート番号、基礎年金番号、免許証番号など）

要配慮個人情報

以下のように定義される情報は**要配慮個人情報**といわれ、取り扱いの制限が多い

- 本人の人種、心情、社会的身分、病歴、犯罪の経歴、犯罪により害を被った事実その他本人に対する不当な差別、偏見その他の不利益が生じないようにその取扱に特に配慮を要するものとして政令で定める記述等が含まれる個人情報

金融分野ガイドラインにおいて、**機微情報（センシティブ情報）**が、以下のように定義されている

- 政治的見解、信教、労働組合への加盟、人種・民族、本籍地、保健医療及び性生活、犯罪経歴

要配慮個人情報と機微情報に関しては、特定の条件にあてあまる場合を除いて、本人の同意がない限り、取得・利用・第三者への提供が原則禁止されている。

匿名加工情報と仮名加工情報

データ分析において個人情報を利用する際、匿名化を行うことがある。

匿名加工情報とは、特定の個人を識別できず、かつ、復元できないように個人情報を加工したもの。匿名加工情報を作成した場合、その情報に含まれる個人に関する情報の項目を公表する義務がある。

仮名加工情報とは、個人を特定できる情報を削除し、単体では個人を特定できないように加工した情報。仮名加工情報は、匿名加工情報よりも加工の要件が緩和され、他の情報と照らし合わせると個人を特定することが可能。

オプトアウト制度

個人データを第三者に提供する場合、個人情報取扱事業者は、原則として予め本人の同意を得なければならない。これを**オプトイン方式**という。一方、明確な本人の同意を得ずに第三者提供ができる**オプトアウト制度**がある。

オプトアウト制度とは、本人から要求が出た際に、直ちにその要求に応じ、その本人が識別される個人データの第三者への提供を停止するという条件のもとで、本人の同意を得ることなく第三者に個人データを提供できる仕組み。個人データの提供を停止するためには**オプトアウト**という手続きをする必要がある。

カメラ画像とプライバシー

顔画像から目、鼻、口の形や位置関係といった特徴を抽出し数値化したデータセットは個人識別符号に該当する。一方で、カメラ画像から抽出した数値データが、単体では個人と識別できず、かつ本人を判別可能な別の画像や個人識別符号と容易に照合できない場合、個人情報には該当しない。

音声についても、声や通話内容から個人を特定できる可能性がある場合、個人情報に該当する。

IoT推進コンソーシアムが公表している**カメラ画像利活用ガイドブック**において、プライバシー権や肖像権に配慮したカメラ画像の利活用について検討が進められている。カメラ画像の撮影自体にプライバシーや肖像権の侵害が問われる場合がある。合法性を満たすためには、撮影方法や利用目的が正当であること、といった点を考慮する必要がある。

EU一般データ保護規則

GDPRでは、クレジットカード情報、メールアドレス、クッキー情報なども個人情報とみなされている。

データポータビリティ権 GDPRで定められている権利で、各サービスのユーザーが自身の個人データにアクセスできるとともに、持ち出しや移転が可能になること。ユーザーは自身の個人データの管理者に対して、次の権利行使可能。

Introduction

- 自身の個人データを、その管理者から一定のフォーマットで受取、他の管理者に移転する権利
- 別の言い方をすると、自身の個人データを異なる管理者間で自ら直接移転させる権利

EU圏外でも、次のいずれかを満たす場合GDPRが適用される。

- EU域内に拠点を有する管理者、または処理者が、EU域内の拠点の活動の過程において個人データを取り扱う場合
- EU域内に拠点がなくても、域内のデータ主体に対して物品・サービスの提供または行動の監視を行う場合

日本国内にのみ拠点を持つ企業でもGDPRが適用される可能性がある。

さらに、GDPRでは、無断でEU域外への個人データの移転が禁止されている。移転先の第三国が十分なデータ保護の水準を確保していると欧州委員会が判断する必要がある。日本は2019年に十分性認定を受けている。

7.5 公平性・説明責任・透明性

公平性・説明責任・透明性の定義

- 公平性 (Fairness)

AIが不当なバイアスを社会に反映させないように配慮すること。アルゴリズムが不透明である場合、アルゴリズムバイアスが発生している場合、学習データに偏りがある場合に、AIが出力する判断の公平性が損なわれる。

- 説明責任 (Accountability)

AIを用いた業務の内容や目的、不祥事が生じた場合の責任体制を開示する責任を指す

- 透明性 (Transparency)

各プロセスが誰にでもわかるように説明できる状態を指す。

ブラックボックス性の解消に向けた取り組みとして**XAI (Explainable AI, 説明可能AI)** の研究開発が進められている。

人間中心のAI社会原則

内閣府における有識者会議の議論に基づき、2019年に**人間中心のAI社会原則**の原案が公開された。

以下3つの理念を尊重・実現するための内容となっている。

- 人間の尊厳が尊重される社会
- 多様な背景を持つ人々が多様な幸せを追求できる社会
- 持続性ある社会

AIの研究開発や利活用に関して考慮すべき7つの基本原則が含まれている。

- 人間中心の原則
- 教育・リテラシーの原則
- プライバシー確保の原則
- セキュリティ確保の原則
- 構成競争確保の原則
- 公平性、説明責任、及び透明性の原則
- イノベーション

7.6 生成モデルが社会に与えるインパクト

ディープフェイク

ディープフェイクとは、2つ以上の画像（動画）を結合させることで、実在しない対象物の画像（動画）を生成する技術を指す。

初期のディープフェイクは、敵対的生成ネットワーク（GAN; Generative Adversarial Network）を用いて作り出された。GANは2つのニューラルネットワーク、ジェネレータとディスクリミネータで構成される。2つのNNを競合させることで、それぞれの性能が強くなっていき、GANのシステム全体として限りなく本物に近い偽画像データを生成できるようになる。

同様な技術を用いて音声や文章など、マルチモーダルなフェイク技術が展開されている。

フェイクコンテンツの悪用例

ディープフェイクを制作した人が使用したプログラムはオープンソースとして公開されている。現在は、スマートフォンなどを使ってほぼ誰でも簡単にフェイク画像・動画を作成できてしまう。

ディープフェイクの有益な使い方

エンターテインメントやクリエイティブ分野では有用な目的での利用も期待されている。

ディープフェイクの悪用を防ぐための対策

フェイクコンテンツを検出する技術の研究開発への取り組みも始まっている。2020年、Microsoft社がセキュリティ対策ソフトウェアのVideo Authenticatorを発表している。

各SNSサービスでは、ディープフェイクの画像や動画は禁止されている。

7.7 言語モデルと生成AIに関する課題

生成AIに関する法的な問題

- 著作権の侵害
- 秘密情報の漏洩
- プライバシーの侵害

著作権法のもとで、著作物の条件に「十分な創作性を持つ表現物」であることが含まれている。生成AIが生成した文章は、そのまま、著作物として認められる可能性は低い。一方で、AIの支援を受けて生成したものが著作物になる可能性はある。

- AIの生成物をそのまま利用せず、そこからアイデアをもらって新しく創作をする
- AIの生成物の顕著な一部分を創作的に編集して活用する

生成AIと著作権侵害のリスク

AIの学習データとして、他社の著作物を無断で使用することを特例として認める著作権法上の規定がある。一方、生成AIに著作物を学習させた場合、その学習済みモデルによる生成物が著作権侵害を起こしやすいことに注意が必要。生成プロセスにおいて、他人の著作物を入力し、他人の直鎖物に依拠した生成物を出力させる行為は、著作権法で禁じられている複製・翻案に該当するためである。

ただし、私的な範囲での複製は著作権法で認められている。

生成AIと個人情報・機密情報の問題

個人情報を入力することは、OpenAI社への個人情報の提供に該当する。これは、学習に利用されるかどうかとは関係がないことに注意が必要。

生成AIの倫理的な問題

以下のような倫理的な課題が注目されている

- 出力における差別やバイアスをなくす
- モデルの透明性と解釈可能性を改善する
- AIの民主化を促進し、幅広い層にアクセス可能な技術にする

学習データ自体は人間が作成したものもあるので、人間社会に潜むバイアスに影響されやすい。なるべ多様性をもたせるような工夫がされているものの、センシティブな属性は生成された文章の公平性に影響を及ぼす可能性がある。

法規制と国内外のガイドライン

法規制は不十分。OpenAIなどは利用規約で禁止事項を明確にしている。

医療、経済、法律といった高リスクの専門分野では生成AIの利用が厳しく規制されている。

ハードローとソフトロー 法的な拘束力のある法律や条例をハードローという。法律上強制力をもたないようなガイドライン、自主規制、推奨事項などをソフトローという。現在、日本や米国はソフトローを中心に採用し、EUや中国は罰則を伴うハードローを採用する傾向にある。

Introduction

日本では、2023年12月に**AI事業者ガイドライン案**が提出され、2024年1月に正式に公表された。AI開発者、AI提供者、AI利用者に分けて、それぞれが留意すべき事項や必要な取り組みが示されている。

EUのAI規制の動向 2023年5月以降、AI規制法が可決されている。生成AIの提供事業者には以下の義務を課している。

- 透明性確保の義務
- AI生成物に関して、AIによる生成を明記する
- 著作物で保護されたデータをAIの学習に利用した場合は公表し、詳細な要約を提供する
- EU法に違反するコンテンツ生成に対するセーフガードの確率

米国のAI規制の動向 2022年10月にAI権利章典として、AIシステムの開発の非拘束的な5つの原則を公開している。

2023年7月には、AIの開発企業7社と米政府が、AIがもたらす様々なリスクに対処することを自主的な拘束としてまとめた。

7.8 AIによる個人の意識の操作

フィルターバブルとエコーチェンバー

フィルターバブル AIレコメンド機能などにおいて嗜好の分析にもとづくパーソナライズが強すぎることで、特定の分野ばかりに注意を向けさせ、特定の団体の存在だけ強調され、無意識に我々の意識に偏りをもたらしてしまう現象を指す。

エコーチェンバー SNSを利用する際、自分と意見や関心が似ているユーザーのみフォローすることによって、SNS上で投稿すると自分と似たような意見ばっかり帰ってくる、という現象。あたかも小さな部屋（チェンバー）の中で自分の声がこだましてくるかのようなイメージに由来している。

8.1 DS検定に出題される数学・統計学

(省略)

8.2 統計学とデータ分析

省略

8.3 記述統計学と推測統計学

- **記述統計学** データの特徴や傾向をわかりやすく・直感的に説明することを目指す。基礎統計量を算出し、表やグラフで可視化する
- **推測統計学** 母集団から異標本を抜き出して、その標本の特性を調査することで母集団全体の特性を推定する。

記述統計学の詳細

基本統計量 または 要約統計量

- 平均値
- 最頻値
- 標準偏差
- 最大値
- 最小値

記述統計学の弱点

基本的に全数調査であるため、完全なデータをすべて取得できない場合は適用するのが難しい。

推測統計学の詳細

統計的推定に関わる用語

- **母集団**
- **標本**
- **母数** 母平均や母分散など、母集団を決定するパラメータ。（平均、分散くらいしか出てこない）
- **統計的推定** 母数が未知の場合に、標本から観測された値を用いて母数を推定すること。点推定と区間推定の2種類がある。
- **推定量** 母数を推定するために標本から求められる統計量。例えば標本平均など。

点推定と区間推定

- **点推定**は、母数を1つの値で推定すること。
- **区間推定**は、推定した区間に母数が収まる確率を考える。この確率のことを**信頼度(信頼水準、信頼係数)**という。90%、95%、99%がよく使われる。設定した信頼度のもとで推定した母数が存在しうる区間のことを**信頼区間**という。

8.4 基本統計量（平均値、中央値、最頻値）

平均値

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

中央値

すべてのデータを順番に並べてちょうど真ん中の順番に来る値。

データが奇数個($2n+1$)の場合、 $n+1$ 番目のデータが中央値となる。（5個あれば3番目が中央値）。

データが偶数個($2n$)の場合、 n 番目と $n+1$ 番目の平均値を中心値とする。（6個の場合、3番目と4番目を足して2で割る）。

平均値と中央値の使い分け

正規分布のようにデータの分布が左右対象の場合、平均値、中央値、最頻値はほぼ同じ値になる。データに偏りがある場合、平均値、中央値、最頻値は異なる値を取る。

データが左に偏っている場合、中央値 < 平均値となる。

データが右に偏っている場合、平均値 < 中央値となる。

平均値は、**標本全体を代表する値**として使われる。

データに異常値や外れ値が含まれると、平均値は影響を受けやすい。

中央値は、外れ値や異常値の影響をあまりうけずにデータ全体を代表できる。

最頻値

最も頻繁に出てくる値。度数分布表では度数が最も大きい値。

質的データ（商品カテゴリなど）に対しては、平均や中央値は計算できないため、最頻値を使う。

8.5 データのばらつきを評価する指標

標本誤差と標準誤差

標本誤差は、母集団の真の値（例えば母平均）と、標本から得られた統計量（例えば標本平均）の差。母集団全体を調査する代わりに一部の標本を使って統計を算出する際に生じる誤差のこと。標本が母集団を完全に代表していないために発生する誤差。標本誤差は、標本サイズが大きいほど小さくなる傾向にある。

標準誤差は、標本の統計量（例えば標本平均）の標準偏差。特に、標本平均の標準誤差は、標本平均が母平均の真の値からどれだけばらつくかを表す。標準誤差が大きいほど推定量の精度が悪く、標準誤差が小さいほど推定量の精度が良いと解釈する。

標準誤差は、標本サイズが大きくなるほど小さくなり、より正確に母平均を推定できるようになる。

標本平均の標準偏差は、中心極限定理により次の式で求まる。

計算方法:

$$\text{標準誤差} = \frac{\sigma}{\sqrt{n}}$$

ここで、 σ は母集団の標準偏差になる。

母集団の標準偏差が未知の場合、標本の**不偏分散**で代用する。

標本の不偏分散 \hat{S}^2 は次の式で得られる。

$$\text{標本の不偏分散} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

これを使うと、標準誤差（標本平均の標準偏差）は以下で計算する。

$$\text{標準誤差} = \frac{\hat{S}}{\sqrt{n}}$$

標準偏差

標本分散:

$$S^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$$

標準偏差:

$$S = \sqrt{S}$$

- 標準偏差が小さいと、平均に近いデータが多い
- 標準偏差が大きいと、平均から離れたデータが多い
- 標準偏差が0の場合、すべてのデータが同じ値である。
- データが正規分布に従う場合、

1σ の範囲に含まれるデータの割合は **68.27%**

2σ の範囲に含まれるデータの割合は **95.47%**

3σ の範囲に含まれるデータの割合は **99.73%**

になる。

標準誤差と標準偏差の解釈の違い

- 標準偏差はデータのばらつき
- 標準誤差は推定量（主に標本平均）の標準誤差、母集団に関する推定の精度

8.6 不偏推定量と自由度

母集団と標本に関する記号と計算式

標本、母集団の統計量、推定量について、本書では以下の記号を使う

	平均	標準偏差	分散	不偏分散	共分散
母集団	μ	σ	σ^2		σ_{xy}
標本	\bar{x}	S	S^2	\bar{S}^2	S_{xy}

標本平均(\bar{x}):

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

標本分散 (S^2):

$$S^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$$

標本の標準偏差 (S):

$$S = \sqrt{S^2}$$

標本の不偏分散 (\hat{S}^2):

$$\hat{S}^2 = \frac{1}{n-1} \sum_{i=1}^n n(x_i - \bar{x})^2$$

母平均 (μ):

$$\mu = \frac{1}{N} \sum_{i=1}^N x_i$$

母分散 (σ^2):

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2$$

母集団の標準偏差 (σ):

$$\sigma = \sqrt{\sigma^2}$$

良い推定量の条件

点推定において良い推定量となるための条件がいくつかある。

- **一致性** 十分に標本を集めれば推定結果が母集団の値に一致する。すなわち、

$$\hat{\theta} \rightarrow \theta(n \rightarrow \infty)$$

- **不偏性** 推定結果がバイアスしていない。すなわち、

$$E(\hat{\theta}) = \theta$$

この式は、推定量の期待値が母集団に一致することを表している。

標本から計算した統計量（平均や分散など）は、母集団の値からはずれる可能性がある。分散に関しては、標本分散は母分散よりも小さくなることが証明されている。されている。そのため、標本文さの値を使ってそのまま母分散を推定すると、母分散を過小評価してしまうことになる。

このような偏りを補正した統計量を**不偏推定量(Unbiased Estimator)**という。

分散については、次のような不偏分散を使う。

$$\hat{S}^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

この不偏分散は、普遍性を満たしている。つまり、標本に対する不偏分散の期待値は、母集団の分散に一致している。

- 分散の不偏推定量は不偏分散になる
- 平均の不偏推定量は標本平均になる。（標本平均は普遍性を持つことが示されている）

なお、書籍では紹介されていないが、期待値 E の定義は以下となる。

確率変数 X が離散型であり、その確率関数が $P(\cdot)$ であるとき、期待値 $E(X)$ は次で計算される。

$$E(X) = \sum_{i=1}^n x_i \cdot P(X = x_i)$$

確率変数 X が連続型であり、その確率密度関数が $f(x)$ であるとき、期待値 $E(X)$ は次で計算される。

$$E(X) = \int_{-\infty}^{\infty} x \cdot f(x) dx$$

自由度

自由度とは、自由に値を取れるデータの数と定義される。

標本平均の自由度は n 、不偏分散の自由度は $n-1$ 。

不偏分散の場合、偏差の合計は 0 になるという制約があるため、自由に扱える数の値が 1 つ制約されてしまう ($n-1$ 個のデータを決めると、残り 1 つは必然的に値が決まってしまう) というイメージ。

8.7 2変数の関係

相関関係の解釈

- 正の相関 ... 一方が増えれば他方も増加するような関係
- 負の相関 ... 一方が増えれば他方が減少するような関係
- 無相関 ... 正の相関とも負の相関とも言えない
- 偽相関 ... 相関があるように見えるが、別の要因で相関があるように見えているだけの状態

共分散と相関係数

共分散(S_{xy})は、以下の式で得られる:

$$S_{xy} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

共分散では偏差の積を取っている。正*正なら正、負*負も正になる、つまり、正の相関がある場合に正の値を取る。正*負または負*正は負になるため、負の相関がある場合に負の値を取る。

共分散を各変量の標準偏差の積で割り算することで、単位や標準偏差の影響を受けない値になる。これが相関係数になっている。

相関係数 r_{xy} :

$$r_{xy} = \frac{S_{xy}}{S_x S_y} = \frac{1}{n S_x S_y} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

相関係数は共分散／標準偏差の積になっている。

このように定義される相関係数は、正式には**ピアソンの積率相関係数**という。

相関係数の目安

- $r \leq 0.6$: 負の相関が強い
- $-0.6 \leq r \leq -0.2$: 負の相関がある程度みられる
- $-0.2 \leq r \leq 0.2$: 相関はなさそう
- $0.2 \leq r \leq 0.6$: 正の相関がある程度みられる
- $0.6 \leq r$: 正の相関が強い

スピアマンの順位相関

2変数の間に線形な関係が存在しないデータや、データの値の順位しか情報がない場合、ピアソンの積率相関は使えないが、**スピアマンの順位相関係数**を使うことができる。

ピアソンの積率相関は、例えば2変数が曲線のように相関している場合に使うことができない。スピアマンの順位相関係数は、単調増加や単調減少の傾向だけを見るため、曲線のような非線形な関係にも適用することができる。

スピアマンの順位相関係数 ρ :

$$\rho = 1 - \frac{6}{n(n^2 - 1)} \sum_{i=1}^n (x_i - y_i)^2$$

ここで、 x_i, y_i は、 i 番目のデータにおけるそれぞれの変数の順位を表す。

Introduction

- ρ が1に近い : 2変数間で順位の正の関連付けが強い
- ρ が0に近い : 2変数間で順位の関連付けがほぼない
- ρ が-1に近い : 2変数間で順位の負の関連付けが強い

8.8 相関関係と因果関係

因果関係と相関関係の違い

相関関係では、出来事の起こる前後関係は考慮されない。

因果関係があるとは、次のことを意味する。

事象Aが起きたことによって事象Bが変化する

または

事象Bが起きたことによって事象Aが変化する

因果関係には方向性がある。

疑似相関と交絡因子

交絡因子とは、以下のような因子をいう。

- 本来関係のない2つの事象に影響を与えて相関係数を高くする第三の因子
- 本来関係のない2つの事象のそれぞれと因果関係がなり立つ第三の因子

疑似相関とは、交絡因子によって相関関係があるように見える状態を言う。

因果関係を特定できるのか

因果関係の厳密な証明は難しい。

因果関係の根拠を集める手法としては、対照実験やランダム化実験などがある。これらはビッグデータとして自動的に収集され蓄積されたデータは使用できず、仮説検証のために最初から計画的なデータ収集を行う必要がある。玄逸的な実験を行うことは科学の基礎研究や医薬開発などを除いて、現実的ではないことが多い。

8.9 確率・確率変数・確率分布

条件付き確率

事象Bが起こる条件の下で、事象Aが起こる条件付き確率を $P(A|B)$ と下記、以下の式で求める。

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

上の式は、次のように書き換えることができる。（ベイズの定理）

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)}$$

確率変数

確率変数とは、確率的に決まる値を保持できる変数。

数学的な定義: 確率変数 X は、確率空間 (Ω, \mathcal{F}, P) から実数の集合 \mathbb{R} への関数として定義される。ここで、

- Ω は、試行のすべての結果からなるサンプル空間
- \mathcal{F} は、 Ω 上の σ -加法族
- P は、確率測度

である。確率変数 X は、サンプル空間の各点 $\omega \in \Omega$ に対して、実数値 $X(\omega)$ を対応させる関数である。

確率変数には、**離散型確率変数**と**連続型確率変数**の2種類がある。

離散型確率変数は、有限または可算無限個の値を取る確率変数。離散型確率変数は、確率関数によってその確率を記述する。

- X が離散型確率変数で、値 x_1, x_2, \dots, x_n を取る場合、その確率関数は次のように表される。

$$P(X = x_i) = p_i \text{ for } i = 1, 2, \dots, n$$

連続型確率変数は、無限に多くの値を取る確率変数で、実数の範囲で値を取る。連続型確率変数は、確率密度関数によってその確率を記述する。

- X が連続型確率変数で、確率密度関数 $f(x)$ を持つ場合、任意の区間 $[a, b]$ における確率は次の積分で求められる。

$$P(a \leq X \leq b) = \int_a^b f(x) dx$$

確率変数を用いた場合の期待値（平均）と分散は次で定義される。

- **期待値(平均)**

- 離散型の場合: $E(X) = \sum_{i=1}^n x_i p_i$
- 連続型の場合: $E(X) = \int_{-\infty}^{\infty} x f(x) dx$

- **分散**

- 離散型の場合: $Var(X) = \sum_{i=1}^n (x_i - \mu)^2 p_i$
- 連続型の場合: $Var(X) = \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx$

8.10 代表的な確率分布の紹介

二項分布（ベルヌーイ分布）

結果が2通りしかない試行をn回実施したとき、ある事象が何回起こるかを示す確率分布

例

- コインを5回なげて裏が3回出る確率
- 100人にある治療を施したとき、80人以上に症状の緩和が見られる確率
- 100名の腫瘍を精密検査したとき、10名以上が悪性になる確率

1と0の2通りの結果が出る試行において、1が出る確率を p とする。 n 回の試行で k 回だけ1となる確率は次になる。

$$P(p, n, k) = \binom{n}{k} p^k (1-p)^{n-k}$$

- 二項分布の期待値: $E(X) = np$
- 二項分布の分散: $Var(X) = np(1-p)$

二項分布は、試行回数nを大きくしていくと正規分布に近づいていく。

ポアソン分布

ある確率で起こる事象が一定の時間内に発生する回数を表す確率分布

単位時間あたり平均 λ 回起こるランダムな減少が、その単位時間の期間中に k 回起こる確率は次になる。

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

- ポアソン分布の期待値: $E(X) = \lambda$
- ポアソン分布の分散: $Var(X) = \lambda$

ポアソン分布は、 λ がある程度大きくなると、 λ を中心とする正規分布に近づく。

ポアソン分布は二項分布を近似した確率分布になっている。二項分布において、期待値($n \cdot p$)を一定に保つつつ、 $n \rightarrow \infty, p \rightarrow 0$ という近似を適用するとポアソン分布になる。

正規分布

正規分布は、平均 μ を中心として、左右対称の鐘形をしている。標準偏差 σ は、平均から山の変曲点までの距離に相当する。

正規分布では、平均から $\pm 1\sigma$ の範囲に全体の約68%、 $\pm 2\sigma$ の範囲に約95%、 $\pm 3\sigma$ の範囲に約99%のデータが収まる。

正規分布の確率密度関数は次になる。

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

標準化と正規化

標準化:

$$\hat{x}_i = \frac{x_i - \mu}{\sigma}$$

標準化は、各データ値から全データの平均値を引いて、標準偏差で割り算する操作。平均を0、標準偏差を1に変換することができる。

正規化:

データを0と1の間に揃える操作で**スケーリング**とも呼ばれる。以下の計算を行う。

$$(データの値 - 最小値) / (最大値 - 最小値)$$

標準化と正規化の目的:

- データを扱いやすい形式に整える
- データ分析の精度や効率をよくする
- 単位や範囲の異なるデータ属性を比較しやすくする

特徴

- 正規化は標準化に比べて計算量が少ない。
- 正規化は外れ値、偏り、最大値、最小値に敏感に反応しやすい。
- 標準化はデータの外れ値や偏りの影響を補正できる。

標準正規分布（Z分布）

正規分布に確率変数 X を標準化した確率変数 Z を**Z値**と呼ぶ。

$$Z = \frac{X - \mu}{\sigma}$$

このZは、平均0、標準偏差1の正規分布で、**標準正規分布**と呼ばれる。

Z値が従う確率密度関数は以下になる。

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

偏差値の計算（Z分布の応用）

偏差値は以下の式で計算される。

$$t_i = 50 + 10 \times \frac{x_i - \mu}{\sigma}$$

偏差値は、点数を標準化することで、平均0、標準偏差1の分布に変換した後、更に10倍して50を加算することで、平均50、標準偏差10の正規分布に変換した値になっている。

正規分布は、 $\pm 1\sigma$ に全体の68%のデータが収まるので、偏差値40から60に全体の68%が収まることになる。

8.11 中心極限定理と標本平均の定理

中心極限定理

中心極限定理: 平均が μ 、分散が σ^2 の母集団から、大きさ n の標本を無作為抽出する。このとき、このとき、もとの母集団がどんな分布であっても、標本の平均 \bar{x} は、平均 μ 、分散 σ^2/n で近似した分布でおよそ表現できる。標本サイズ n がある程度大きければ、 \bar{x} は正規分布に従うとみなせる。

母集団から標本 (x_1, x_2, \dots, x_n) をランダム抽出する。標本サイズ n が十分大きければ、標本平均 \bar{x} およびそれを標準化した統計量について以下が成り立つ。

$$\bar{x} \sim N(\mu, \frac{\sigma^2}{n})$$

平均を引いたり、標準化すると以下が得られる。

$$\bar{x} - \mu \sim N(0, \frac{\sigma^2}{n})$$

$$\frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

これは、次のように解釈できる。

- n が大きいほど、標本平均は母平均に近づき、標本平均の分散が小さくなる（分母に n があるため）
- n が大きいほど、母集団がどんな分布であっても、標本平均の従う分布は正規分布に近づく
- n が大きければ、母集団がどんな分布であっても、Z統計量は標準正規分布で近似できる

正規母集団の標本平均についての定理

平均値 μ 、分散 σ^2 の正規分布に従う母集団から、大きさ n の標本を取り出した場合、標本平均 \bar{x} は、平均が μ 、分散が σ^2/n の正規分布に従う。

中心極限定理との違いは、母集団が正規分布に従う場合、 n が十分な大きさを持たなくても、標本平均は正規分布に従う、という点。