第4章 データサイエンスに必須! ビック・データを数学の力でさばいて必要なデータのみ取り出す

データの特徴抽出

切通 恵介

● 特徴をつかみたいけどデータが多すぎ

近年ビッグ・データやAIといったキーワードで大量のデータを扱うことが増えています．例えば製造業においては大型のプラントの計器から数百個のセンサ・データが取得されます．また，アンケートから得られたユーザの年齢，性別，趣味，年収などのプロフィール情報を分析することでマーケティングを行う例もあります．しかし，このような大量のデータを人間の目で全て把握することは難しいでしょう．これらの課題を解決するためには大量データの特徴を抽出する必要があります．

● 特徴抽出と線形代数との関係

データ分析における特徴抽出はよく次元圧縮とも言われます．その代表的な手法が，主成分分析，線形判別分析，独立成分分析，非負値行列因子分解の4つです．これらの手法はざっくりと次の手順で計算されます．

(1) 手法が与える条件に応じた最適な変換（例えば主成分分析であれば分散を最大化させる変換）を求める問題を解くべく，関数を定義し，それを最小化，または最大化する問題に帰着させる．
(2) 設定された最小化，最大化問題を解き，最適な変換を求める．

これらの手順において，実は線形代数の基礎的な計算を必要とします．というのも，データとその変換はそれぞれ行列と定義できるためです．目的関数の定義において行列やベクトルの計算の他，固有値，固有ベクトル，行列式の算出などを用います．また，その目的関数の最適化においてもベクトルや行列の微分を活用します．

4-1 データの特徴抽出が必要な理由

機械学習の勉強やサンプルとして利用されるデータではカラム数（特徴数，次元数）の少ないデータが使われます．例えばデータセットとして有名なIris Dataset[1]のカラム数は4つ，Titanic[2]のカラム数は10です．一方で，実世界のデータにおいてはこのような少ないカラム数ではなく非常に多くのカラム数を持つデータが存在します．工場内のセンサ・データなどの例では数百のセンサから集めたデータを使って分析をすることがあります．例えば，カラム数が1000個のセンサ・データがあったとします．このようなデータを分析する場合どのような課題が生じるでしょうか．次に主な2つの課題を示し，特徴抽出が必要なことを示します．

● 理由1：特徴が少ないデータで把握できる

データの可視化は大量のデータを人間が判断したり新たな知見を得たりするために行うデータ分析において最も重要な作業の1つです．カラムごとの統計量や分布を計算してグラフ化したり，カラム間の関係性や相関を確認したりすることでデータの全容を把握できます

しかし，1000次元（1000個のカラム）のデータの場合，平均や分散などの標準的な統計量を比較するだけでも時間がかかります．さらに複数のカラム同士の傾向を調べるために散布図を使うだけでも${}_{1000}C_2 = 499500$通りの散布図ができます．これを全て閲覧することは現実的ではありません．データの理解の観点でも特徴を抽出して次元数を削減し，人間の認知負荷を下げる必要があります．

● 理由2：機械学習のモデル学習がうまくいく

1000次元のデータを機械学習に入れて予測などを行う場合，全てのデータを入れるのではなく，この中から重要なものだけを取り出して利用することがあります．例えば，ある1つのセンサの値を他のセンサか