研究員の眼

「次元の呪い」への対処

モデルの精度を上げるにはどうしたらよいか?

保険研究部 主席研究員　篠原　拓也
(03)3512-1823　tshino@nli-research.co.jp

ビッグデータという言葉が世に広まって久しい。これは、音声データや動画データのように2次元の表形式に変換できないようなデータ(非構造化データ)や、グラフや電子メールのように一定の規則性はあるものの表形式にはなっていないデータ(半構造化データ)が中心を占める膨大なデータを指す。

今世紀初頭から、IT化の進展を背景に世界的にその概念が徐々に広がっていった。日本では、この言葉は2010年頃から一般に使われ始めた。2013年には、新語・流行語大賞の候補として選ばれたが、大賞には選ばれなかった。それから10数年が経過するなかで、DX(デジタルトランスフォーメーション)、AI(人工知能)、生成AIなど、ビッグデータをベースとした展開が次々にあらわれてきた。

ビッグデータについては、「次元の呪い」と言われるデータの複雑さゆえの問題が、当初から指摘されてきた。この問題にはどのような取り組みがなされているのか。本稿では、この点を中心に見ていくこととしたい。

◇「次元の呪い」とは

まず、言葉の定義から見ていこう。「次元の呪い(curse of dimensionality)」とは、数学の問題で、空間の次元が増えるにつれて、問題を解くのに必要となる計算の量やアルゴリズムが指数関数的に大きくなる現象をいう。

アメリカで制御理論等の応用数学者であったリチャード・ベルマン博士によって生み出された。例として、1メートルの空間を隣接する点が1センチ幅となるよう、格子点で埋めようとすると、1次元(直線)ならば100個の点で足りる。2次元(平面)だと1万個(＝100の2乗個)、3次元(立体)だと100万個(＝100の3乗個)の格子点が必要となる。(視覚的に捉えるのは困難だが頭の中でイメージするとして)10次元(超立体)では、1垓(がい)個(＝100の10乗個)もの格子点が必要となる。

このように、次元が増すごとに、空間を埋める格子点の数は指数関数的に増えていく。

◇「次元の呪い」がもたらす問題

次元の呪いは、数学を数値的に取り扱う上でさまざまな問題をもたらす。組み合わせ論では、離散的な値をとる複数個の要素から生じている状態を対象に分析や検討が行われる。いくつか例を見ていこう。

(1) 複数の色付き電球が作り出す状態の解明

例えば、0 と 1 でオフとオンを表す、異なる色付きの電球が複数個ある場合を考える。電球の数が r 個のときに、それらがオフまたはオンになることで、重なり合ってできる色の種類は 2 の r 乗個となる。ここで、このようにして重なり合った色の状態が 1 つ与えられたとしよう。どの電球がオフでどの電球がオンとなってその状態が生じているのかを明らかにしようとすると、2 の r 乗個の状態それぞれについて確認が必要となる。r の数が増えれば、確認のための手間は膨大となる。

(2) 複数の食材から作ることのできる料理の検討

冷蔵庫の中にいくつかの食材があったとしよう。それらを使って、夕食をつくることにした。食材が 5 つであれば、つくることのできる料理の数は限られるため、それほど悩むことなく料理を決めてその準備に取りかかることができる。しかし、食材が 100 種類もあると、さまざまな料理が可能となり、どの料理をつくろうか、下味はどうするか、出汁は何でとるか等々、料理を決める段階で大いに迷ってしまう可能性がある。

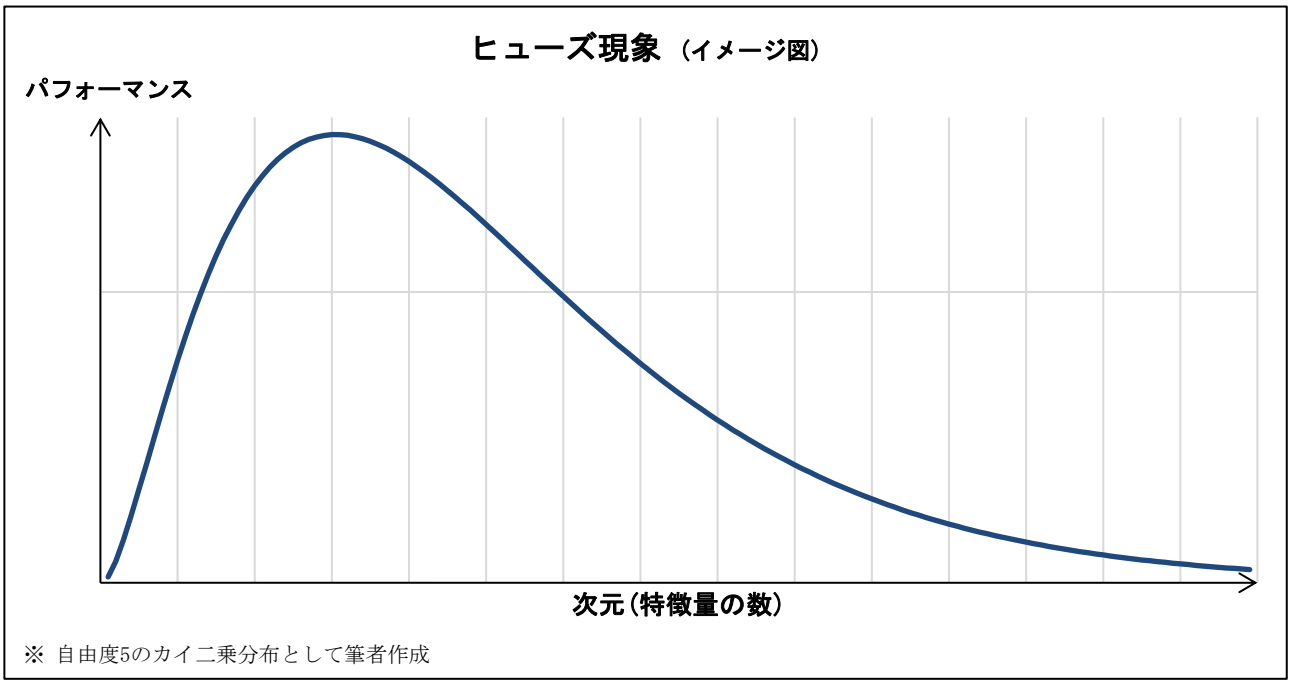
(3) お店で買うシャツを選ぶ場合の検討

アパレルショップでシャツを買う場合について考える。「明るい色のシャツでカジュアルとして着られるもの」という程度の条件で選んでいけば、いくつか見たり試着したりして買うシャツを決められるだろう。しかし、色は白かピンクか水色、デザインは無地かストライプかチェック、素材はコットンかポリエステルかリネン、用途はカジュアルかアウトドアかスポーツ、値段は安いものか中価格帯かセール、などと条件を設定すると、これらの条件を満たすものは 243 通り(=3 の 5 乗)にもなる。選択肢が増え過ぎてしまい、買うシャツがなかなか見つけられなくなる。

◇「次元の呪い」が機械学習にもたらす弊害

こうした次元の呪いは、AI の機械学習でも問題となることが多い。次元の呪いを端的に表すグラフとして、ヒューズ現象(Hughes phenomenon)が知られている。これは、横軸に特徴量(要素)の数(次元)、縦軸に分類のパフォーマンスをとったときに、次のようなグラフが得られることをいう。次元が増えると、ある程度まではパフォーマンスが向上する。しかし、さらに次元が増えていくとパフォーマン

スは低下することを表している。(図では、次元が 3 以上になるとパフォーマンスが低下)



※ 自由度5のカイ二乗分布として筆者作成

AI を用いて、特徴量(要素)の数(次元)を増やして予測や分類を行おうとするときに問題が起こりやすい。具体的には、つぎのような問題が挙げられる。

〈1〉学習データの不足

次元を増やすと、取りうる状態の数は指数関数的に増える。その結果、ほとんどの状態に対する学習データが存在せず、機械学習が困難となる。

〈2〉計算の増加

次元を増やすと、データを処理するための計算量と時間が増える。このため、機械学習の負荷が大きくなる。

〈3〉過学習

次元を増やすと予測や分類のためのモデルが過度に複雑となる。その結果、モデルが、データに潜むパターンではなくノイズに適合してしまう可能性がある。これにより、新しいデータをもとに予測や分類を行ったときに、モデルの能力が低下してしまう。

〈4〉誤差距離の意義の薄弱化

機械学習においては、実際の値とモデルの出力の差を誤差として扱い、これを小さくすることでモデルの精度を向上させていく。しかし、高次元空間では、2 つのデータポイント間の距離が極端に大きくなることがあり、誤差の距離が意義が薄れてしまうことがある。

◇「次元の呪い」への対処

これらの弊害を克服するためには、次元を削減することや、有効な次元を抽出することが必要となる。これは、データ内の最も重要な情報を保持しながら、冗長な特徴量や重要度の低い特徴量は破棄しようとする対処の方法だ。具体的に、見ていこう。

〔1〕主成分分析(Principal Component Analysis, PCA)

主成分分析は、元の変数を、元の変数の線形結合である新しい変数に変換する統計手法だ。この新しい変数は、主成分と呼ばれる。主成分分析により、主成分に置き換えて、残りの次元を減らすことで、データを簡明に分析することができる。

〔2〕線形判別分析(Linear Discriminant Analysis, LDA)

データの差異が大きい特徴量を統計的に特定する手法だ。特に、分類を行うモデルで有用となる。線形判別分析によって特定された特徴量の線形結合により新たな変数を作成することで、精度の高い分類モデルを作ることができる。

〔3〕t-分布型確率的近傍埋め込み(t-distributed Stochastic Neighbor Embedding, t-SNE)

低次元での距離の分布を、正規分布ではなく、裾野の厚いt-分布に従うものと仮定する。そのうえで、データがまばらで距離が大きくなりがちな高次元での距離分布が低次元にも合致するよう、データの変換を行う。条件付確率を用いて類似度を表すことで、高次元でのデータの局所的な構造を低次元でも維持する(類似しているデータを低次元上でも近くに保つ)ことを可能としている。

〔4〕自己符号化器(オートエンコーダー) (Autoencoder)

次元の削減や有効な次元の抽出を、ニューラルネットワークを用いて行う方法だ。入力されたデータを一度圧縮し、重要な特徴量だけを残した後、再度もとの次元に復元処理をするアルゴリズムを意味する。特に、画像認識などの分野で威力を発揮するとされる。

PCA や LDA は、新たな変数の作成を線形に行う線型手法とされる。複雑なデータでは、予測や分類の一定の精度低下が避けられない場合がある。

一方、t-SNE や自己符号化器は、非線形手法と位置づけられる。t-SNE は、計算負荷が大きく、大規模データセットでの適用が難しい場合がある。また、自己符号化器は復元処理の誤差が大きくなり、分類の精度が低下する場合がある。

◇ AI 開発の動向からは目が離せない

以上、次元の呪いの概要とそれへの対処について見てきた。高次元のデータを扱う際には、モデル

が優れたパフォーマンスを発揮するために次元の呪いに対処することが重要とされてきた。

実は、最新の AI のニューラルネットワークモデルでは、予測や分類の誤差の評価に、距離による測定を用いないものが多くなっているといわれる。それによって次元の呪いから逃れることができるという。ただし、それに伴って別の問題が生じる可能性は無視できないかもしれない。

例えば、そうした距離を用いない評価が人間の肌感覚に合うものかどうか。また、そもそも人間のアナログな肌感覚など最初から考慮せずに、結果のパフォーマンスのみを追求することとなるのか…。

AI 開発は日進月歩の状況が続く。その動向について、引き続き注視していくこととしたい。

(参考資料)

“The Curse of Dimensionality in Machine Learning: Challenges, Impacts, and Solutions” Abid Ali Awan (datacamp / WRITE FOR US, Sep 13, 2023)

“Curse of Dimensionality in Machine Learning”(Geeks for Geeks, Last Updated: 11 Dec, 2024)

“What Is the Curse of Dimensionality?” Badreesh Shetty(Updated by Jessica Powers)(Built in, Aug 19, 2022)

“Curse of Dimensionality”(Wikipedia)