身近になる機械学習
グローバル版

デロイトは2018年において、大企業および中堅企業は機械学習の活用にさらに注力すると予測している。機械学習技術の実装数および当技術を利用したパイロットプロジェクト数は2017年から倍増し、2020年までにはさらに2倍になるだろう。また、機械学習アプリケーションプログラムインターフェース（API）やクラウドで使用できる専用ハードウエア等の技術により、大企業だけでなく小規模企業にも広く利用されるようになるだろう。

機械学習とは、明示的にプログラムしなくても、システムがデータを読み込み、その経験の過程で自ら学習し、改善を図ることができる人工知能（AI）、またはコグニティブテクノロジーのことである。

アナリスト達は機械学習の投資と導入は世界規模で大きく拡大すると予測している。International Data Corporation（IDC）は、AIおよび機械学習への投資は2017年の120億ドルから2021年までに576億ドルへ拡大するとみている。しかし今はまだ、機械学習の導入は初期段階にすぎない[87]。機械学習のツールやフレームワークはまだ確立されておらず発展途上にある[88]。

デロイトでは、従業員500人以上の規模で、コグニティブコンピューティングを積極的に導入している米国企業の「コグニティブに理解のある」経営幹部を対象に調査を行った[89]。回答者の半数の企業が従業員5,000人以上の規模であった。回答者は当技術について中程度か、それ以上の理解があり、自社での利用について熟知している。

回答者らはコグニティブテクノロジーの可能性に高い興味を示していたものの、進行中の実用化件数およびパイロットは5件以下との回答が過半数（60%）を占めた[90]。

しかし本章で扱う5つの主要領域の進化（後述）によって、機械学習ソリューションはもっと簡単に早く開発できるようになり、このパワフルな技術の導入を阻んできた障壁もいくらか取り除かれるだろう。これらの方向性に沿った進化は、機械学習への投資拡大と企業内での機械学習の積極的な利用につながるにちがいない。その結果、2018年末までに企業の機械学習パイロット件数と実用化件数は倍増し、機械学習を利用する大企業の2/3以上でそれぞれ10件以上になると考えられる。

しかし企業における機械学習の活発な利用を導く、重要な進化には5つの方向性がある。

これら5つのうちの3つ、すなわちデータサイエンスの自動化、必要な学習データ量の低減、学習の高速化は機械学習をより簡単に、より安く、より早くするため、機械学習市場を拡大するだろう。残りの2つ、モデルの結果説明およびローカルデバイス上での機械学習は、新領域で応用・活用され、これもやはり市場を拡大するだろう。

機械学習は他の方法でも進化しつづけているため、2018年内にも新たに大きな改良が生まれる可能性がある。

以下にデロイトが特定している5つの進化の方向性の詳細を、適用範囲が広いものから順に説明していこう。

1. データサイエンスの自動化：データエクスプロレーションやフィーチャーエンジニアリングなど、多大な時間を必要とする機械学習タスクは一般にデータサイエンティストの活動時間の80%を占めるが、これを次第に自動化できるだろう[91]。

データサイエンスはしばしば専門家領域と誤解されがちだが、実際は芸術とサイエンスの混合である。データラングリング（wrangling）から探索的データ解析、フィーチャーエンジニアリング、フィーチャー選択、予測モデリング、モデル選択などに至るまで、データサイエンティストが時間を費やすタスクの多くは、そのすべてまたは一部が自動化できる。例えばAirbnbのデータサイエンティストたちは、ゲストおよびホストの顧客生涯価値（lifetime value：LTV）モデルを構築するかたわら、自動化されたプラットフォームを使用して複数のアルゴリズムとフィーチャーエンジニアリング手順を検証した。自動化の機能を併用しなければ、LTVモデルを構築するのに十分な時間を割けなかっただろう。データサイエンティストらは自動化によってアルゴリズムにある変更を加えれば5%以上精度を向上できることを発見できた[92]。これは重大なインパクトである。

数多くのデータサイエンスの自動化ツールや技術を、大手企業だけでなく起業間もないベンチャーも提供するようになり、機械学習の概念実証にかかる期間は月単位から日単位へと短縮できるはずである[93]。データサイエンスの自動化は、データサイエンティストの生産性のさらなる向上を意味する。データサイエンティストの深刻な人手不足[94]は解消され、企業は機械学習を利用した活動を倍増できるだろう。

2. 必要な学習データ量の低減：機械学習モデルのトレーニングには数百万のデータ要素が必要なこともあり、これが導入の大きな障壁となりうる。学習で使用するデータの取得およびラベル付けには膨大な時間とコストがかかる[95]。例えば、MRI画像に診断内容をラベル付けしなければならないプロジェクトがあるとしよう。一人の放射線技師を雇い、1時間に6枚のペースで1,000枚の画像を精査しラベル付けしてもらうには3万ドル以上のコストがかかると推定される。そもそもプライバシーや機密情報保護の問題があって、データ取得自体が難しいこともある[96]。

しかし機械学習に必要な学習データ量を減らすために、多くの有望な技術が生まれている。一つはリアルデータの特性を模倣するようアルゴリズム的に生成された合成データを利用する技術である[97]。デロイトのチームがあるツールを検証したところ、当ツールは従来必要とされた学習データのわずか1/5で正確なモデルを構築できた。残りの4/5のデータは合成したものだ。

合成学習データによってデータサイエンスソリューションのクラウドソーシングの可能性も広がる。多くの組織がサードパーティを巻き込んで機械学習の問題解決モデルの構築を行ってきた。その際には外部のデータサイエンティストが使用できるよう、共有に適したデータセットを提供している[98]。MITの研究者らは、オリジナルデータセットを公開せずに、リアルデータを使用して、予測モデル開発のクラウドソーシングに使用可能な合成データを作成した。15回のテストのうち11回で、合成データボルトから開発されたモデルは、リアルデータで学習したモデルと同等のパフォーマンスを見せた[99]。

他にも必要な学習データ量を減らす技術に、転移学習がある。この手法を用いた機械学習のモデルは、言語翻訳や画像認識など、似たような領域（ドメイン）のデータセットを使用して事前学習することで、新たなデータセットを早く学習することができる。機械学習ツールベンダーの中には、自社が提供する転移学習を利用することで、顧客が提供すべき学習例の数を桁違いに削減できると主張する企業もある[100]。

3. 学習の高速化：「次世代機械学習チップの勢いが加速」の章で詳述したように、ハードウエアメーカーの大手企業もベンチャー企業も、チップ内の計算とデータ転送を高速化して機械学習モデルの学習時間を削減できるよう、専用ハードウエア（GPU、FPGA、ASIC等）を開発している。これらの専用プロセッサにより、企業は機械学習の学習と実行を何倍も高速化でき、その結果関連コストを削減できる。

例えば、マイクロソフトの研究チームはGPUを使用して1年で人間のように特定の会話音声を認識できるシステムを完成させた。同じことをCPUを使って実行しようとすれば5年はかかったであろう[101]。

グーグルはTPUと呼ばれるニューラルネットワーク実行用のAIチップを自社設計し、TPUをCPUとGPUアーキテクチャに加えることによって十数もの追加データセンターの構築にかかるコストを抑えられたとしている[102]。

このような専用AIチップを早く採用したのはデータサイエンスや機械学習関連の大手技術ベンダーや研究機関等であるが、採用の動きは小売業、金融サービス、通信といったセクターにも広がりつつある。主要な全てのクラウドプロバイダ（IBM、マイクロソフト、グーグル、AWS）が提供するGPUクラウドコンピューティングサービスを利用することで、学習の高速化が主流となり、機械学習に取り組むチームの生産性は向上し、企業が取り組むアプリケーション数は倍増するだろう。

4. モデルの結果説明：機械学習は日々進化している。しかし機械学習モデルの重大な欠点がよく問題とされる。機械学習モデルの多くはブラックボックスであり、どう意思決定が行われているのか、確信を持って説明することができない。そのため、モデルが導いた回答への信頼性（例えば顧客にインセンティブを提供する場合など）や、法規制の順守のためなど、さまざまな理由により、多くのアプリケーションが不適切とされたり、容認されずにいる。例えば、米国の金融サービス業界が遵守する連邦準備制度（Fed）の監督文書SR11-7の「モデルリスク管理指針（Guidance on Model Risk Management）」では、とりわけモデル行動の説明が求められている[103]。

特定の機械学習モデルについては、そのブラックボックスを明らかにする技術が数多く生み出され、説明可能性と精度は向上してきている。正確な予測とその予測の根拠を提示するニューラルネットワークのトレーニング方法もMITの研究者が実証している[104]。

すでに製品化されたデータサイエンス技術もある。例えばH2Oのデータサイエンス自動化プラットフォーム「H2O Driverless AI」[105]、DataScience.comの新Pythonライブラリ「Skater」[106]、DataRobotの保険料率算定用の機械学習予測モデリング[107]などである。説明可能な機械学習モデルの構築が可能になれば、金融サービス、ライフサイエンス、ヘルスケアといった規制の厳しい業界の企業も機械学習の使用を増やすことができ、今後数年のうちにパイロットプロジェクト数および実用化件数も大幅に増すものと思われる。

適用できる領域としては、クレジットスコアリング、レコメンデーションエンジン、離脱顧客対策、不正検出、病気の診断と治療などが考えられる[108]。

5. ローカルデバイスへの搭載：ニーズに応じてデバイス等に搭載されるようになれば機械学習の利用も増すだろう。デロイトが昨年予測したように、モバイルデバイスおよびスマートセンサへの機械学習の搭載が増えてきており、技術の適用もスマートホーム、スマートシティ、自動運転車、ウエアラブル技術、産業IoTへと拡大している[109]。

Google、Microsoft、Facebook、Apple等の技術ベンダーは、ポータブルデバイス上で画像認識や言語翻訳等のタスクを実行できる小型の機械学習ソフトウエアモデルを開発している。Googleは「TensorFlow Lite」を、Microsoftは組み込み学習ライブラリを、Facebookは「Caffe2Go」を、そしてAppleはオンデバイス処理用の「Core ML」を、それぞれ使用している[110]。Microsoft Research Labは機械学習モデルの圧縮に取り組み、10〜100倍の小型化に成功した[111]。

GoogleやMicrosoftはもちろん、Intel、Qualcomm、NVIDIA等を含む半導体ベンダーは機械学習をモバイルデバイスに搭載できるよう電力効率の優れたAIチップを自社開発している[112]。スマートフォンデバイスへの機械学習の搭載がますます現実味を帯びてきているため、多様な用途での利用が進み企業の機械学習パイロット数および実用化件数も多くなるだろう。

用語解説

データサイエンス：複雑なデータセット（往々にして大量または体系化されていないことが多い）から洞察を得るための、データ管理、アナリティクスモデリング、ビジネス分析を一般に扱う学際領域。

学習データ：インプットデータと対応するアウトプットデータ（またはラベル）との関係を発見しモデル化する目的で使用されるデータ。例えば住宅販売履歴に関するインプットデータとして面積、建築年、学区の3つの属性データがあり、アウトプットとして販売価格がある場合、これら3つの属性データと販売価格との関係性を見出すべくアルゴリズムを用いる。何らかのモデルによってその関係性を見つければ、これらの3つの属性インプットがあれば他の住宅の販売価格を予測することが可能になる。学習データまたはラベル付けされたデータからそのようなモデルを作成もしくは学習することを「教師あり学習（supervised machine learning）」という。

ブラックボックス：内部の仕組みが明らかになっていないもの。ブラックボックス型の機械学習モデルでは医療診断や与信などの回答が論理的根拠を説明することなく生成される。対照的に、その内部の仕組みを明らかにしているホワイトボックス型のモデルであれば、どうしてその結果に至ったかを理解することが可能となる。

説明可能性（Interpretability）：ここでは、あるシステムが決定を行う理由と方法を説明する能力のこと[113]。

データラングリング（Data wrangling）：複雑で体系化されていないデータセットを使いやすく、分析しやすくするために、データをクリーニングしソートするプロセスのこと。

データエクスプロレーション（Data exploration）：データセットを理解しデータの主要な特徴をまとめるために行うデータ分析の最初のステップ。

フィーチャーエンジニアリング：特定の領域において蓄積された知識を活用して、機械学習モデル向けに、既存のデータを基に関連するデータの特徴を表形式でまとめるプロセス。

ニューラルネットワーク：人間の脳の神経細胞の仕組みを模した、つながり合ったノードの層構造をもつネットワーク。システムが自らの学習データを分析してタスクの実行を学ぶ形式の機械学習に使用される。

要点

機械学習の5つの方向性における進化が加速し、企業での機械学習の使用は2018年末までに倍増するはずである。そして長期的には、これらのベクトルが一助となって機械学習をメインストリームの技術に押し上げると考えられる。これまでは機械学習モデルをトレーニングする人材、インフラ、データに制約のあった業界でも、機械学習の進化により新たなビジネスへの適用が可能になるだろう。

そのために企業は以下を実行すべきである：

- データサイエンティストの業務を一部自動化できるような機会を模索し、どうしたらデータサイエンスの自動化を活用できるか専門家に相談する。
- 学習データの取得というボトルネックを緩和できるような新進の技術（データ合成や転移学習等）に注目する。
- どのような機械学習向けコンピューティングリソースをクラウドプロバイダが提供しているかを調査する。自社データセンターですでに一定の作業を実施しているのであれば、機械学習の専用ハードウエアを追加して併用するよう検討することになるかもしれない。
- 説明可能性を向上できる最新技術がないか調査する。ただし説明可能な機械学習はまだ初期段階のため、市場ではまだ主流となっていない可能性がある。
- デバイス上での機械学習の利用が実現可能となる時期を予測するため、次世代チップのメーカーが公表している性能ベンチマークを定点観測する。