executive series

機械学習入門

SAS Best Practices e-book

執筆：キンバリー・ネバラ（Kimberly Nevala）

sas best practices
THOUGHT PROVOKING BUSINESS

目次

1

機械学習とは?

機械（**machine** \mə-ˈshēn\）: 所定のタスクを実行するために機械的、電気的、電子的な仕組みで動作する装置。

学習（**learning** \ˈlərniNG\）: 勉強すること、練習すること、教えてもらうこと、何かを経験することによって知識やスキルを身につける活動またはプロセス。

機械は学習する?

はい！ 機械は、データを調べてパターンを検出したり、既知のルールを適用して以下のことを実行したりすることによって学習します。

- **分類：** 人やモノなどを仕訳し整理する
- **予測：** 特定されたパターンにもとづき類似の結果や行動を予見する
- **特定：** これまでは知られていなかったパターンや関係を明らかにする
- **検知：** 異常な行動や想定外の行動を洗い出す

機械が学習するために用いるプロセスは「アルゴリズム」と呼ばれます。アルゴリズムが異なれば、学習の方法も異なります。観測された応答や環境の変化に関する新たなデータが「機械」に提供されると（因果関係などをさらに詳しく学習できるため）、アルゴリズムのパフォーマンスが改善されます。その結果、時の経過とともに学習を重ねるほど、「インテリジェンス」のレベルが向上していきます。

でも……
機械には
創造性がある?
あるいは、
独立した知性が
ある?

ビッグデータ時代の到来により、利用できるデータの量とそれを処理する人間の能力は、ともに飛躍的に増大しています。機械の学習能力も向上しており、ますます「賢く」なっているように思われます。それでも、機械は自主的に思考することはありません（今のところは、ですが）。

もちろん、機械学習は、これまで見落とされていたチャンスや解決すべき課題を特定する手段としては有効です。しかし、機械が自律的に創造性を発揮することはありませんし、人間によるお膳立てや指示がない状態で自然発生的に事実（データ）から新しい仮説を立てることもありません。また、前例のない刺激に反応するための新たな方法を機械が判断できるわけでもありません。

重要ポイント：機械学習アルゴリズムの出力は、学習材料として与えられるデータに完全に依存しています。データが変われば、結果も変わります。

パーソナライズされたマーケティング

今日の企業は、顧客が自社の製品を購入し、サービスを利用し、専門知識を活用する理由について、以前よりも深く理解できるようになっています。大量の消費者データを「機械」にかけることで、消費のパターンや好みのチャネルを検出できるからです。機械は履歴データとリアルタイム・データを駆使し、出張が多くコーヒーが大好きな顧客（＝筆者のことです）は「お気に入りのカフェがその角の先にありますよ」というリアルタイム・メッセージを歓迎してくれるはず、という判断を下すことができます。その一方で、筆者の父はこの種のインタラクション（対話や情報のやり取り）を歓迎しません。父は “自宅コーヒー派” だからです。ですから、メールのクーポンには反応します。クーポンには、父が次に買い物に出たときに購入しそうなコーヒー以外の商品に関する特典や割引券を含めることもできるため、リアルタイム・メッセージとは別の、クーポンならではの利点があります。

機械は、人間のお膳立てがあれば既存のチャネル（デジタル、紙、実店舗）の全てをまたいでデータと状況を分析し、顧客一人ひとりに対するインタラクションを最適化することができます。しかし、現存しない新たな顧客対応チャネルを機械が自主的に創出することは決してありません。

機械学習に適している課題は?

機械学習が特に適しているのは以下のような課題の解決です。

- 適用可能なルールがあると直感的には思えるものの、シンプルなロジックだけでコーディングまたは記述するのは難しい。
- 候補となる出力やアクションは定義済みだが、どのアクションが最良かの判断はさまざまな条件に左右されるため、事象／イベントの発生前に予測したり一意に特定したりすることができない。
- 人間による解釈のしやすさよりも、結果の精度の方が重視される。
- データが、従来の分析手法ではうまく扱えない特徴を備えている。特に、幅が広いデータ（レコード数と比較して各レコードに含まれるデータポイントや属性の数が非常に多いデータセット）や、相関度が極端に高いデータ（類似した値や密接に関連した値を多く含むデータセット）は、従来の分析手法では問題が生じやすい。

具体例で納得！

画像や映像に含まれている人やモノの識別

十分な学習を積んだ機械学習アルゴリズムは、混雑した空港の監視カメラ映像から既知の「容疑者」を認識することができ、航空便への搭乗や、さらに悪い事態を阻止するために役立ちます。

ソーシャルメディア・プラットフォームでは、アップロードされた写真に機械学習を適用し、人物へのタグ付けや、ランドマークなどの有名オブジェクトの識別を自動的に行っています。

この課題が機械学習に適している理由

画像データは複雑です。1枚の画像は大量のピクセルで構成されているため、深さに比べて幅が極端に広いデータセットになります。近くにあるピクセルは類似した色（＝類似した値）になることが多いため、相関度が高いデータとなります。同じ対象物の画像でも、微妙な（あるいは、かなり明白な）違いを含んだ複数のバリエーションがありえます。

もちろん、人間なら写真を見れば、知っている人（および知らない人）を、たとえ表情、ポーズ、服装が違っていても簡単に識別することができます。また、人間は概念的なレベル（動物、鉱物、野菜など）と具体的なレベル（犬、猫、魚など）のどちらに関しても、「類似した」項目を判別できます。しかし、こうした照合判断をする際に人間が利用している知識を、理路整然とした形で、単純な処理ステップや個別のルールに変換することは、人間自身でさえ困難です。

ゲノミクス

機械学習は、特定の疾病の発現経路に関与している遺伝子を発見する取り組みにも役立ちます。

また機械学習は、個々の患者にとって最も効果的な治療法を、患者の遺伝子構造、デモグラフィックス（人口統計学的属性）およびサイコグラフィックス（心理学的属性）上の特性にもとづいて判断する目的にも利用できます。

この課題が機械学習に適している理由

遺伝子データは幅が広く、１人の人間には２万個を超える遺伝子があります。その結果、遺伝子のデータセットでは常に、そこに含まれる個人（レコード）の数に比べ、各レコードに含まれる遺伝子（データポイント）の延べ総数の方が圧倒的に多くなります。

また、複雑性を増大させる数々の要因が存在します。例えば、2万個を超える遺伝子のそれぞれの中身には、非常に幅広いバリエーションが見られます。もちろん、血縁者同士のゲノム（全遺伝情報）が類似しており、高度な相関関係にあることは事実です。この比較的少数の集団がかかりやすい特定の疾病について分析する場合は、データプールが極端に浅くなることもあります。なお、遺伝子だけを分析しても、将来の健康状態や病気の発現を高い精度で予測することは困難です。生化学的要因や環境的要因などの幅広い要因も考慮しなければならず、そのためには多種多様なソースのデータを統合する必要があります。

ナビゲーションと自動運転車

この分野で機械学習を活用すると、A 地点から B 地点までの最良ルートの特定、乗継条件や移動時間の予測、最新の道路状況にもとづく最良ルートの予測などが可能です。

また、ドライバーが何もしなくても機械が全てを行う自動運転車や無人自動車も実現間近です。

この課題が機械学習に適している理由

自動車の運転は複雑な課題ですが、その対象範囲は明確です。実際、車両が取りうるのは、発進、停止、前進、後退、ハンドル操作（方向調整と右折／左折）、加速、減速などの限られた動作です。ただし、どの動作を行う場合でも、その意思決定には多くの要因が影響します。道路条件、気象条件、他の車両の存在と挙動、二足歩行する人間と四足歩行のペット、交通規則などは、そうした要因のごく一部です。人間のドライバーは、こうした入力の全てを本能的かつ瞬時に評価しますが、考えうる全ての組み合わせを個別のルールとして記述するのは不可能です。

一般的な用途


初期段階での感染検出

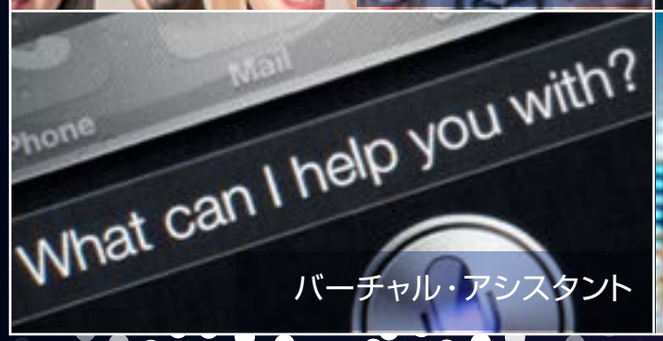

予防保守

不正検知


自動運転車


リソース／資源
の最適化


顔認識・モノ認識



バーチャル・アシスタント


大量の
トレーディング


セールス／
マーケティング

2

基本的な手法

機械学習の4大タイプ

教師あり学習

教師あり学習の場合、機械は実例を教えてもらうことができます。人間が機械に対し、望ましい入力と出力の例を与えます。「機械」（実際にはアルゴリズム）は、この入力を学習材料として使い、答えの予測に利用できる相関関係やロジックを判断します。

これは、学生に例題と正解のみを示し、「解き方を説明しなさい」と指示するのと同じです。教師あり学習では、問題と答えのサンプルが与えられます。機械は、Ａという問題からＢという答えを導き出す方法を探します。適切な論理パターンが見つかったら、類似した問題の解決にそのパターンを適用できます。

一般的な手法

- ベイジアン統計解析
- 決定木
- 予測
- ニューラル・ネットワーク
- ランダムフォレスト
- 回帰分析
- サポート・ベクター・マシン（SVM）

現実の用途

半教師あり学習

一般的な手法

- 「教師あり学習」のページを参照

半教師あり学習は、教師あり学習の場合と同様の問題を解決するために使われます。ただし、半教師あり学習で機械に与えられるデータの場合、答えが示されているデータ（=ラベル付きデータ）は一部のみであり、それ以外のデータには答えが示されていません。言い換えると、入力データの一部は「望ましい出力（答え）」でタグ付けされているのに対し、残りのデータはタグ付けされていません。

半教師あり学習は、データ量が多すぎるため、あるいはデータ内の微妙なバリエーションが多すぎるために、データ全体について入力／出力の実例を完全に用意できない場合に使われます。この場合、機械は、与えられた入力と出力から一般的なパターンを特定し、それを残りのデータに適用することで、それらのデータの出力を推定します（=外挿法による推定）。

現実の用途

画像認識／画像分類

音声認識（スピーチ認識）

Web ページの分類

教師なし学習

教師なし学習では、機械自身がデータを調べてパターンを特定します。この手法では、例題と正解は与えられません。機械は、利用できるデータを解析することで、相関や関係を判断します。

教師なし学習では、人間が周囲の世界を観察するときに自然に使っている方法に準じてモデルが作成されます。すなわち、「制約のない観察と直感にもとづいて推論し、類似する事物をグループ化する」という方法です。人間は経験を積むにつれて（機械の場合は、与えられるデータの量が増えるにつれて）、直感力や観察力が変化および／または向上していきます。

一般的な手法

- アフィニティ分析
- クラスタリング
- クラスタリング：k 平均法
- 近傍法マッピング
- 自己組織化マップ（SOM）
- 特異値分解（SVD）

現実の用途


マーケット・バスケット分析


異常検出／侵入検知


類似した顧客やモノの特定

強化学習

一般的な手法

- 人工ニューラル・ネットワーク（ANN）
- 学習オートマトン
- マルコフ決定過程（MDP）
- Q学習

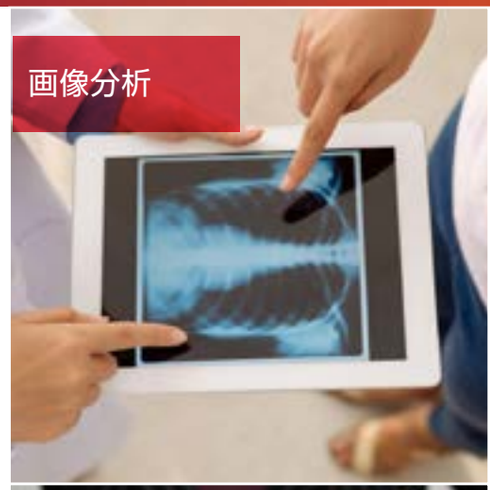
強化学習の場合、機械は、アクション、ルール、取りうる最終状態に関して、許容される条件のセットを与えられます。言い換えると、「ゲームのルール」のみが定義されます。機械はルールの適用や異なるアクションの実行を試しながら、結果として生じる反応を観察することで、どのようにルールを活用すれば望ましい成果を生み出せるかを学習します。そして、どのような一連のアクションをどのような状況で実行すればよいかの判断を積み重ねることで、やがては最適な結果にたどりつきます。

強化学習は、ゲームの遊び方を誰かに教えることに相当します。どのようなゲームでも、ルールと目標は明確に定義されています。ただし、プレイヤーは現在の戦況や対戦相手のスキルとアクションに応じて自分の戦い方を調整しなければならないため、ゲームは毎回、プレイヤーの判断に応じて異なる展開になります。

現実の用途

ホット・
トピックス

ディープ・ラーニング



ディープ・ラーニングは、極めて高度なニューラル・ネットワークを活用する最先端の高度な機械学習手法です。ディープ・ラーニングと呼ばれるのは、生成されるモデルが従来型のニューラル・ネットワークと比べ非常に複雑である（＝階層が深い）からです。また、ディープ・ラーニング・モデルは、従来のモデルに比べ極めて大量のデータを処理できる点も特長です。

重要な理由

ディープ・ラーニングは、今日利用されている数多くの高度な機械学習システムを支える基盤です。恐らく最も重要な点は、ディープ・ラーニングによって画像、音声、ビデオを理解・分析する能力が飛躍的に向上したことです。これが可能になった主な要因は、機械学習の研究が大きく進展したことと、利用できるデータの量とコンピューティング性能の両方が飛躍的に増大したことです。

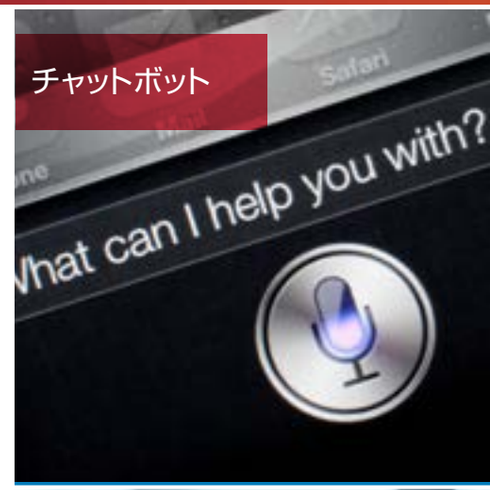
ホット・
トピックス

コグニティブ・コンピューティング

コグニティブ・コンピューティングを搭載したシステムが目指すのは、人間の行動を理解およびエミュレートすることや、人間と機械を橋渡しするインターフェイスをより自然で直感的なものにすることです。こうした取り組みでは通常、人間が「母語」でシステムを操作できる仕組みも展開します。言い換えると、ユーザーがコードを記述または理解する必要性を排除する、ということです。これを実現するため、コグニティブ・コンピューティング・プラットフォームでは、自然言語処理、高度な機械学習アルゴリズム（ディープ・ラーニングを含む）、自然言語生成などの幅広い手法を活用します。

重要な理由

コグニティブ・コンピューティングによって、機械（ソフトウェア・システム）は、人間にとって「よりアクセスしやすく、より直感的に関わり合える存在」となります。そのため、コグニティブ・コンピューティングは、自動化された業務システムや分析ソリューションの普及のカギとなる可能性もあります。最終的な目標は、「人間」対「機械」という対立の構図を超越したレベルで、人間と機械がシームレスに連携できる協調型のシステムを生み出すことです。これは、人々が人工知能について話すときに共通して思い浮かべている近未来のAI機能の、現時点での「前身」と言えます。



ホット・
トピックス

自然言語処理

自然言語処理（NLP）は、機械が人間の書き言葉や音声による指示（またはその両方）を理解できるようにする機能です。NLPには、言語を機械やアルゴリズムが理解できる形式に翻訳する機能も含まれます。自然言語生成（NLG）は、機械が人間に対し、計算結果や反応を「平明な英語」（またはサポートするように設計された任意の言語）で伝達できるようにする機能です。

重要な理由

一部のNLPツールは、単純に翻訳を実行し、人間の指示に含まれる単語を既存の（機械向けの）辞書にマッピングします。より高度なNLPアプリケーションは「理解」することを目指しており、適切なアクションや反応を返すための前段階として、人間の言葉（音声やテキスト）の意味や意図を推論します。方言、比喩表現、口語表現、個人の話し方の癖などにおける振れ幅の大きさや、新たなコミュニケーション形態（例：略語、絵文字など）の急速な進化などを考えると、この取り組みは決して簡単ではありません。

3

考慮すべき問題点

機械がどのように結論に達したかを人間は誰も説明できないのは、なぜか?

従来の統計モデルとは異なり、機械学習アルゴリズムで作成されるモデルは極めて複雑です。実際、突飛に思える動作の裏に緻密な計算があったとしても、それが人間にとって即座に理解できる明快なものとは限りません。例えば、ニューラル・ネットワークの厳密な処理経路を追跡するのは容易ではありません。モデルを定義するルールやパラメータの数は、数千（場合によっては数十億）に達する場合もあります。そのため、内部で具体的にどのような処理経路をたどるのかは、データ・サイエンティストにとってもブラックボックスなのです。

このため、より重要な疑問に突きあたります。つまり「目下の課題に対してアルゴリズムや手法が適切に適用されているのか?」ということです。

機械学習アルゴリズムがブラックボックスなら、何を根拠に信頼できるのか?

アナリティクスのメカニズム（より正確には論理的な処理経路またはルール）が明確ではなく、再現も容易でないとすれば、結果をどのように検証すればよいのでしょうか?

ブラックボックスの処理を受け入れることと妄信することを混同してはなりません。機械学習の場合、検証は拍子抜けするほど単純です。

新しいデータに対してテストを行う場合は、次の点を検証します。

- そのアルゴリズムは、将来の事象を正確に予測するでしょうか、あるいは、望ましい結果を導くでしょうか?
- その出力をアクションに反映させることは有益でしょうか?

これが全てであり、それ以上でも以下でもありません。

本当に **それほど** 単純なのか?

機械学習アルゴリズムの検証基準自体は単純です。ただし、妥当な結果を提供できるようにアルゴリズムを選択・監査・調整するプロセスは決して単純ではありません。

実に多くの要因を考慮しなければなりません。例えば、「課題やデータに最適なアルゴリズムは何か?」、「分析に含める必要があるデータ要素 (=「特徴」) は何か?」、「モデルに与える重要な要素の質を向上させるためにデータをクレンジング／変換／改良できるか? 」(=「特徴抽出」 および 「特徴エンジニアリング」)、「パフォーマンスを最適化するために、アルゴリズムのパラメータをどのように設定／チューニングするべきか?」など。

また、モデルのクロス・バリデーションと監査を行うことで、一見正確そうだが実際にはそうではない応答が人為的に導き出される状況 (=「過学習／過剰適合」) も回避しなければなりません。単純化した例 (極端すぎることは筆者も承知しています) を示すと、昨日の天候を高い精度で予測するモデルを作成することは比較的容易です。しかし、そのモデルは、明日の天候を同様の精度で予測するとは限りませんし、良好に機能する唯一のモデルであるとも限りません。

極言すれば、良好に機能する機械学習システムを開発する作業は、反復と集中を要するプロセスであり、その大部分はサイエンス (科学者の仕事) ですが、部分的にはアート (職人芸的な仕事) なのです。

複雑性やクリーン性は、常に優位性を意味するのか?

そうとは限りません。従来型のアナリティクスと同様、機械学習プロジェクトでも、時間の大半はデータの変換、検証、フォーマットに費やされます。しかし、データ品質は常に懸念事項ですが、機械学習の場合に重要なのは（人間の目線ではなく）アルゴリズムの目線で「必要十分」であるかどうかです。単純なアルゴリズムでデータが多い場合の方が、複雑なアルゴリズムでデータが少ない場合よりも、優れた結果につながるケースが少なくありません（データセットが大きくなるとダーティー率がいくぶん高まることを考慮したとしても、です）。

モデルの正確性に関しては、「高いほど良い」と考えがちであり、経験が浅い人々は特にそうです。しかし、現実の用途の多くでは、モデルの正確性をわずかに改善ところで、ビジネスにおける運用結果の改善に直結するわけではありません。データや特徴を増やしてもアルゴリズムが必要以上に複雑になるだけ、という場合もあります。複雑性と実用性の間でうまくバランスを取る必要があります。

注目情報：第一線の機械学習研究者であり、2017年3月まで百度（バイドゥ）社のチーフ・データ・サイエンティストを務めていたアンドリュー・ウ（Andrew Ng）氏は、「機械学習の将来の発展においては、新しいアルゴリズムの開発よりも、アルゴリズムがデータをよりスマートに処理できるようにする取り組みが重視されるだろう」という趣旨の見通しを述べています。

機械学習によって既存のアナリティクスは陳腐化する?

機械学習はアナリティクスのツールボックスに入っているツールの１つにすぎません。どのようなツールでもそうですが、よく考えて使わないと、「ハンマーしか持たない人には、全てが釘に見える」という格言どおりの罠にはまってしまいます。機械学習は学術界から生まれてきたため、早期の導入者たちは、従来の統計アルゴリズムなら容易に解決できる課題に対し、膨大な時間と手間を投じる結果に甘んじることも珍しくありませんでした。

そのため、機械学習は従来の分析手法を完全に置き換えるのではなく、あくまでも補完するものだと考えるのがよいでしょう。

機械学習によって「人間」も陳腐化する?

機械学習を実践する際には、人間が科学的な手法とコミュニケーション・スキルの両方を適用・応用しなければなりません。

これは「データを入れてかき混ぜましょう」といった簡単なレシピではありません。

人間には、アルゴリズムをプログラミングするデータ・サイエンティスト以上に、次のような質問に答えを出すことが求められます。

- 何を予測しようとしているのか?
- 結果として得られる相関関係は、予測に役立つか?　因果関係を示すか?　データ自体にバイアス（偏り）は存在しないか?
- 結果は期待に添ったものか?　対処すべき例外はあるか?
- 予測値は何であり、一般化することは可能か?
- そのモデルと結果は、現実世界に適用できるか?
- 適切な反応は何か?

4

ベストプラクティス

機械学習は、人間と機械の相乗効果によって成り立つ取り組みです。

機械学習を実践する際には、人間が科学的な手法とコミュニケーション・スキルの両方を適用・応用しなければなりません。機械学習の活用に成功している組織は、分析のインフラや専門知識を保有していることに加え、アナリティクスとビジネス、それぞれのエキスパートの間で緊密なコラボレーションを実現しています。

1 原理ではなく概念に関して業務部門を啓発する

ニューラル・ネットワークや機械学習のアルゴリズムの内部動作は、どのような観点に立つかによって、胸が躍る人もいれば、ひどく複雑で退屈極まりないと感じる人もいるでしょう。実際問題、ほとんどの人は詳細を理解する必要がありません（そのつもりもないでしょう）。ただし、これは教育や啓発が不要という意味ではありません。必要な時間と資金に関して賛同を得るためには、機械学習で何が実現できるかについて、技術に詳しくない経営幹部や業務担当者に広く理解してもらう必要があります。

機械学習をビジネスに適用する方法を具体的に紹介する導入事例は、複雑なアルゴリズムにもとづくチャートや、p 値に関する議論よりも、賛同を集める効果があります。技術的な詳細を延々と熱く語るのではく、機械学習で解決できる課題の実例を紹介することに努め、他の業種や類似の企業に関するケーススタディも含めるようにします。手法自体を説明しなければならない場合も、工学的な図ではなく、例え話を考えてください。

2 機械学習を発見プロセスの一部にする

機械学習アルゴリズムは、行動や環境を観測し、パターンを検出し、一般化を行い、説明や理論を推定します。そこから得られる確率論的な相関関係は、結果を高い精度で予測できる可能性がありますが、その結果を生み出す要因を必ずしもピンポイントで特定できるわけではありません。

つまり、予測することと、因果関係を解明することは同じではありません。そのため、ビジネスやポリシーの意思決定プロセスにおいて、得られた洞察を行動に変える必要性とその方法を判断する際に、この点を必ず考慮しなければなりません。場合によっては、機械学習によって、さらなる調査や検討を要する領域が判明することもあります。あるいは、重要な意思決定ポイントや処理経路の判断をリアルタイムで自動化するために、機械学習のアルゴリズム自体を業務システムに組み込むことになる可能性もあります。

具体例で納得!

次の2つの事例を考えてみてください。

2005年に発行された『Freakonomics』(邦訳:『ヤバい経済学(増補改訂版)』、東洋経済新報社、2007年)という書籍では、「家庭にある本の数は**全国統一テストの点数**が高いことと相関関係にある」というケーススタディが取り上げられています。この調査結果を受け、ある市長が、成績の芳しくない子供がいる家庭に無償で本を贈呈するという施策を実施しましたが、その結果は議論の余地もないほどに不調でした。本の数とテストの点数との間に相関関係はあっても、因果関係はなかったからです。

これと対照的なのが、新生児集中治療を受けている未熟児に装着したデバイスからのテレメトリー・データをリアルタイムで分析した、オンタリオ工科大学の研究です。この研究で開発されたシステムは、未熟児が**感染症にかかる**時期を高い精度で予測しました。しかも、その臨床症状が発現したのは、予測から48時間以上も経過した後のことでした。研究者と臨床医は現在でも、機械が感染症の発症を「どのように」特定したのかを解明できていません。しかし最終的に研究チームは、因果関係を完全には理解できないままで、この相関を踏まえて処置や対策を取ることにしました。因果関係の解明よりも予測できることの方がこのケースでは重要だったのです。

3 ブラックボックス化を避ける

確かに、機械学習という手法は捉えどころがなく、謎めいて見えることも多いでしょう。しかし、機械学習はブラックボックス型の取り組みではありません。人間には、次のような質問に答えを出すことが求められます。

何を予測しようとしているのか?

機械学習プロジェクトは、あらゆる分析的な取り組みと同様、探索の対象とする問題空間や仮説を明確に定義することから始める必要があります。

そのプロセスに対して最良と考えられる入力は何か?

データ・サイエンティストや各分野の専門家が連携して、どのようなデータソースを利用し、どのような特徴を機械に分析させるかを明らかにする必要があります。機械学習アルゴリズムへの入力となりうる特徴を特定および検証する作業では、データ・ビジュアライゼーション（視覚化）機能が重要な役割を果たします。

注目情報： 教師なし学習の場合でも、機械は自律的に動作するわけではありません。学習結果は、どのようなデータを与えるかに関する人間の意思決定の影響を受けます。Google が行ったマシンビジョン（機械視覚）の実験では、猫の画像を識別する方法を機械が自力で習得し、その結果から「典型的な猫」の画像を生成することにも成功しましたが、別の画像セットで学習したとしたら、別の「典型的な猫像」が特定される結果になったでしょう。

#3 ブラックボックス化を避ける（続き）

結果は期待に添ったものか？ 対処すべき例外はあるか? 誤った結果はどんな影響をおよぼすか?

スタンフォード大学とGoogleによるコンピューター・ビジョンの共同研究を考えてみてください。驚くほど優れたものでしたが、完璧ではありませんでした。ヤギをイヌ、一面のチューリップを無数の熱気球と誤認したこともありました。これらは確かに小さな失態かもしれませんが、人間が誤って分類されるような場合、それはどんな影響をおよぼすでしょうか?

結果はどのように応用できる（すべき）か?

機械学習は、何をすべきかを判断するのは得意ですが、どのように行うべきかを定義するのは必ずしも得意ではありません（ただし、この点も急速に変わりつつあります）。

適切な反応は何か?

例えば、世界規模の健康や政治に悪影響を及ぼすようなパターンが出現した場合、次に取るべき適切なステップ（群）は何でしょうか?

#4 適切なレベルの科学的厳密性を適用する

機械学習は、揺るぎない統計的推論や科学的なデータ分析の必要性を否定するものではありません。機械学習は非常に強力ですが、自己修正しながら何でも魔法のように解決してくれる分析の「万能薬」ではありません（少なくとも今のところは）。

機械学習が最も効果を発揮するのは、データ・サイエンティストがシステムの構築方法を的確に理解している場合です。この場合は、対象領域の知識に加えて、データの属性がどのように応答するかも考慮した上で適切なアルゴリズムを特定可能となります。このプロセスには、異なるアルゴリズムの特性に関する知識と、ある程度の直感力も必要です。

要注意事項：機械学習アプリケーションの開発は実験と反復を重ねるプロセスであり、この点は経験豊富な開発チームが定番のアルゴリズムを使用する場合も変わりません。どのようなケースでも、ビジネス・コンテキストと利用可能なデータにもとづき、アルゴリズムのトレーニングとチューニングを行わなければなりません。

また、モデルを検証する際も、開発チームは「自分たちの考えは全て正しいと思い込む」罠に陥ることのないように、健全な（しかし過剰ではない）レベルの「懐疑主義」と厳密性を適用しなければなりません。

具体例で納得!

分析モデルの開発に失敗した事例は多々あります。

一例を挙げると、ゲノミクス研究者のチームが、化学療法に対する患者の反応を予測するアルゴリズムを開発しました。残念ながら、彼らは最初の学習用データセットに関して、データの変動性や整合性の問題を考慮しませんでした。そして、そのことが原因となり、臨床試験のキャンセルをはじめとして、いくかの残念な結果につながりました。

別の事例では、エコノミストのグループが発表した論文で、GDP の成長と政府債務残高の多さとの関連性が逆になっていました。発表後、回帰モデルの中で使われている要因の重み付けに関して疑問が呈され、それを修正したところ、全く異なる結論が導かれてしまいました。

#5 必要十分な複雑さに留める

一般論としては、データが多いほど分析精度は向上します。しかし、特徴（属性やデータポイントとも呼ばれます）も多ければ多いほど、より優れた結果が得られるのかと言えば、そうとは限りません。ここでの問題は、データセットが大きくなるほど、データの変動幅が大きくなり、予測エラーの可能性も高まることです。

多くのケースで、比較的単純なモデルで（たとえデータのダーティー度が高いとしても）多くのデータを分析する方が、複雑なモデルで少ないデータを分析する場合よりも優れた結果が得られます。これは、「集積のパワー」を利用して結果を予測するアンサンブル・モデリング手法の基礎となる考え方です。

機械学習をうまく展開するためにはバランスが重要であり、複雑化に伴って徐々に高まる予測価値と、解釈のしやすさ／使いやすさ／適用性との兼ね合いを勘案しなければなりません。

もちろん、シンプルさが唯一の美徳ではありません。最終的に現実の運用条件の下で良好に機能することも、モデルの必須要件です。

具体例で納得!

最良のモデルでも、現場の業務に展開して活用できなければ意味がありません。Netflix Prize というコンテストの例を考えてみましょう。映像ストリーミング配信会社の Netflix 社がデータ・サイエンス・コミュニティに出した課題は、「次にどの作品を視聴すべきか?」を予測するハイパフォーマンスなモデルを作成することでした。優勝したモデルは、あらゆる期待を上回っており、ユーザーの好みをかつてないレベルで予測することができました。

しかし問題がありました。このモデルのデータ要件と処理要件では、リアルタイムまたはニア・リアルタイムでの実行が不可能だったのです。今すぐ観たい作品を探しているユーザーの関心を引くという目的を考えれば、リアルタイム性は最重要の要件のはずでした。結果的に、夢のように正確でありながら、何のビジネス価値も生まないモデルとなってしまったのです。

#6 ビジネスユーザーを検証プロセスに積極的に巻き込む

データ・サイエンティストはモデルを作成・調整するにあたり、デュー・デリジェンス（≒適正な注意義務と努力が担保されていること）を徹底しなければなりません。そのためには、データやビジネス領域の専門家とのコミュニケーションやコラボレーションを効果的に行いながら、モデルの検証と精査を進める必要があります。この点は、分析チームが特に以下の事項を確実に行うために重要です。

全ての選択肢が検討されたことを確認する

著しく正確なモデルがあるからといって、同程度にうまく当てはまる他のモデルの存在価値がなくなるわけではありません。それらのモデルが別の結論を示唆してくれる可能性もあります。

潜在的なバイアスについて説明する

アルゴリズムは、本来的に間違いやバイアスとは無縁である、というわけではありません。気づかぬうちに混入する潜在意識レベルのバイアスを別にしても、モデルに提供されるデータには、そのデータを作成したときに下された意思決定に起因するバイアスが反映されている可能性があります。数学者のジェレミー・クーン（Jeremy Kun）氏がエレガントに説明しているように、「人間によって生成されたデータ（＝発見済みデータ）を用いてトレーニングする以上、個体群（マイノリティーまたはマジョリティー）や基底プロセスに本来的に備わっているバイアスを継承することになります」。

得られた洞察を行動に変えることの影響と意味を特定する

#7 データ・ストーリーテリングを活用する

他の手法の場合と同様、機械学習でも、データ・サイエンス・チームと最終利用者との間で、得られた洞察の“翻訳”がうまく行われない可能性があります。そのため、分析チームは上記の考慮事項に加え、得られた洞察がどのように提供／利用されるかについても慎重に検討しなければなりません。

まず最初に、受け入れやすく利用しやすい方法で洞察を提示する方法を判断する必要があります。このことは、得られた洞察が既存のパラダイムを覆すような画期的なものである場合や、標準の業務手順を改変する必要性を示すような場合には、特に重要です。対象者に数字や統計値の羅列を突きつけるのではなく、分析結果を視覚化して示すことや、説得力あるストーリーに織り込んで示すことを考えるべきです。

要注意事項：話をでっち上げてはなりません。そうではなく、得られた知見が業務改善の促進や革新的な新製品／サービスの実現に役立ちうることを、対象者も理解できる言葉と文脈で具体的に説明するような語り口を練り上げるのです。

#8 人間の認識力を過小評価しない

機械学習を利用すると、個人の将来の行動を高い信頼性で予測できます。その結果、ターゲッティングやマーケティングに関して、MailChimp 社のチーフ・データ・サイエンティストであるジョン・フォアマン（John Foreman）氏が言うところの、「レーザーのようにピンポイントで相手を誘導する不誠実な主張」が飛び交う可能性も生まれます。個人情報保護と提供価値をめぐる論争も含めた倫理面も重要であり、事前に対処しておかなければなりません。

もう1つの考慮事項は、機械が意思決定を行うようになったときに、「可謬性（誤りは避けられないという性質とその認識）」に対する人間の期待がどのように変化するか、という点です。

「機械」が目に見える形で意思決定の役割やエンゲージメント（顧客対応や従業員対応）を担う度合いが高まるにつれ、人間の反応にも適切に対処していなかければなりません。機械学習を最大限に導入・活用するためには、そうしたシステムへの信頼感を築くことが極めて重要です。

具体例で納得!

自動運転車が追突事故を起こした場合をご想像ください。私たち人間は、その事故を許せるでしょうか? また、その事故の原因が、人間のドライバーの場合はもっと高い頻度で発生している操作ミスだったと知らされた場合は、どうでしょうか?

マンモグラフィー画像による**乳がん診断**の精度は、機械学習を用いる場合は72%に達することが、複数の調査研究で示されています。それに対し、人間の医師の場合は65%です。しかし、例えば「機械が診断します。診断ミスはありえますが、誤診の確率は人間の方が高いですから、ご安心ください」などと告げられた場合、患者の期待はどのように変化するのでしょうか? あるいは変化しないのでしょうか?

#9 ビジネスプロセスをプロアクティブ（能動的）に適応させる

機械学習システムを現場業務に展開すると、意思決定やアクション実行がある程度まで自動化されることから、それまで人間が掌握していたコントロールを手放すことになる可能性があります。他方では、機械学習により、全く新しい製品、サービス、カスタマー・エンゲージメント・モデルの開発が可能になる可能性もあります。

- 機械学習がビジネスにもたらす意味を慎重に検討しなければなりません。
- 機械学習から得られた洞察を業務に組み込むという目標に向け、組織として行動を起こし、必要な変化に取り組む準備は整っているか? また、この取り組みに関する意欲は十分か?
- 既存のビジネスプロセスや役割のうち、変更が必要になるのはどの部分か?
- 全く新しいプロセスや役割の導入が必要になるか?
- 機械学習モデルが自律性を備えている場合、それによって実現する自動システムは、人間のワークフローの中でどのように機能することになるのか? それは人間側の担当者や同僚にとって意味のあることか?

具体例で納得!

製造業では、設備機器の故障を発生前にジャストインタイムで特定するために機械学習を活用している企業もあります。これを実現するには、設備機器にセンサーを装着し、分析型のセンシング・システムを組み込む必要があります。こうした取り組みには、顧客サービス、保守整備、保証のポリシーと手続きについて、根本的な再考を促進する効果もあります。

マーチャンダイジング担当者は、ネットショップの価格設定をリアルタイムで行うために機械学習を活用しています。この場合は、機械が最適価格を判断する一方で、総売上の閾値は人間が検証します。フィードバック・ループにより、アルゴリズムは観測された販売結果から学習することができます。機会逸失やエラーに関する仕入担当者のフィードバックを取り込むためのループもあります。こうした業務体制への移行は簡単に見えるかもしれませんが、その実現には、マーチャンダイジング担当者と仕入担当者の業績測定とインセンティブの在り方を抜本的に変革する必要があります。

Amazon の高度に自動化された物流センターでは、人間が商品のパッキングを行う一方で、要求された商品を棚から集めたり、適切な商品が箱に入っていることを検品したりするのはロボットシステムです。全てがうまく機能している理由は、人間とロボットの間の相互作用がうまく機能しているからです。こうした仕組みを成功させるためには、アルゴリズムだけでなく、人間と機械の継続的なエンゲージメントも十分に考慮した設計を重視し、それに徹することが重要です。

#10 継続的なケアとデータ供給を計画する

最先端の分析力を維持し、有益な結果を提供し続けるためには、最新状況を反映するデータにもとづき、継続的に機械学習アルゴリズムを更新・改良していかなければなりません。この点は、新たに分類した顧客マイクロセグメントが顧客維持率に及ぼす影響を評価することが目的の場合でも、エネルギー需要が想定外に急増したときに停電を回避するために電力網のバランスを再調整することが目的の場合でも変わりません。

機械学習の適用が成功し、用途を広げていくと（例：消費者の購買をさらに喚起する、別の商用チャネルも活用する、自動運転車に左折だけでなく右折もさせる）、モデルの作成に用いられるパターンも変化していきます。そのため、こうした新たな用途にも対応できるように、機械学習モデル自体も調整していく必要があります。ビジネスプロセスについても同様です。

この場合、モデルを現場業務に組み込むということは、単に業務活動のモニタリングを続けることではありません。モデルの実効性を維持管理することも、最初のモデル開発と同等あるいはそれ以上のデュー・デリジェンスが求められる、重要かつ継続的なプロセスです。

5

準備度の判断

チェックリスト

機械学習の活用に向けた準備状況は?

☐ 解決する必要のある課題を明確化する

機械学習が最も効果的に機能するのは、課題が明確に定義されている場合です。課題の定義には、実現すべきアクションや、達成すべき測定可能な成果を含める必要があります。さらに望ましいのは、課題が最優先の業務課題や戦略目標に対して明確に紐づけられていることです。

機械学習には多大な時間とデータが必要になるため、その課題に対して既存の分析モデル／アプローチや代替ソリューションを適用できないかどうかを厳格に評価することも理に適っています。これを済ませておけば、最終的に機械学習を選択した時点で、投入する時間と労力に見合った導入効果が得られる可能性を担保したことになります。

注目情報：定型業務における意思決定ポイントのうち、頻度が高く即座または至急の対応が求められるものや、変動性の高い入力に依存するもの、あるいはその両方が当てはまるものは、機械学習の有望な適用候補です。

チェックリスト

機械学習の活用に向けた準備状況は？（続き）

☐ 実験重視の考え方を組織全体に定着させる

機械学習は実験と反復を重ねるプロセスです。中核となるアルゴリズム自体のコモディティ化は進んでいますが、どのようなプロジェクトでも、ビジネス・コンテキストとデータにもとづくカスタマイズが必要不可欠です。優れた実験の例に漏れず、機械学習でも「仮説は正しくない」と証明されることがあります。その場合は、新しいデータを調達または作成する必要があるかもしれませんし、あるいは、得られた知見を踏まえて課題を定義し直すことが必要になるかもしれません。したがって、機械学習を成功させるためには、意思決定者か分析チームのメンバーかを問わず、関係者全員が「試行錯誤を厭わないメンタリティ」を身につけなければなりません。要所にチェックポイントを設けた反復型のプロセスを利用することで、進捗を速やかに評価できる柔軟性と俊敏性が実現し、ひいては、代替アプローチを担保できているかどうかや、十分な試行錯誤を尽くしたどうかの判断も可能になるのです。

☐ コラボレーション型のデータ・サイエンス・チームを編成する

機械学習の専門知識を確保することは、要件の１つにすぎません。それと同様に重要なのは、ビジネス／データ／技術の専門知識を備えた多様なエキスパートの関与を促すダイナミックなチーム編成モデルを導入することです。こうしたエキスパートには、必須のデータ資産の評価とオンボーディングを実行できる「データの専門家」、コンテキストを提供し、提案されたアクションや新しい製品／サービスの意味（ビジネス面／ソーシャル面／モラル面）を評価することができる「ビジネスの専門家」、 技術的なエコシステムを展開／保守することができる「IT 担当者」が含まれます。また、見落とされがちですが、「統計分析の専門家」（数学用語を話す人々）の言葉と「ビジネスの専門家」の言葉を相互に翻訳することができる人材も、忘れずに含める必要があります。

チェックリスト

機械学習の活用に向けた準備状況は？（続き2）

☐ 頑健なデータ戦略とエコシステムを開発する

機械学習の実行には大量のデータが必要です。そのため、高品質なデータと情報資産の特定、取得（または作成）、準備、アクセスを効果的に実行するためのデータプロセスを確立することが極めて重要です。これを達成するには、本番環境だけでなく、探索用環境（いわゆる「サンドボックス」）もサポートするような形で、ガバナンス・ポリシーとデータ・エコシステムを確立しなければなりません。そのためには、セキュリティ、個人情報保護、品質を犠牲にすることなく、アクセス性と俊敏性のバランスを調整することができる多層的なアプローチが必要です。また、非構造化テキスト、音声、画像をはじめとする非従来型の（ビッグ）データソースを導入する場合は、それに対応した新たなデータ管理機能が必要になる可能性もあります。

☐ 組織のリスク許容度を評価する

機械学習を活用しようとすると、何をもって「必要十分」とするかの判断基準に関する合意から、モデルの検証方法や開発方法の理解まで、品質保証（QA）やリスク管理の領域で実践してきた従来型のアプローチでは対応しきれない事態が多発します。その理由は、自転車の補助輪のような役割を果たすトレーニング・データと、ある時点で決別しなければならないからです。真の検証を行うためには、新しいデータを対象として機械のパフォーマンスをテストする必要があります。そして、ほとんどの場合、そのためにはシステムを本番環境に展開しなければなりません。この取り組みの内容は、顧客を望ましい行動に導くモデルとしての有効性を確認するために本番環境で「A/B テスト」を実行するようなケースから、いざという時にハンドルを操作できるように人間の監督者が乗車した状態で自動運転車の公道走行試験を行うケースまで、実に多岐にわたります。

チェックリスト

機械学習の活用に向けた準備状況は？（続き3）

☐ 確立済みのビジネスプロセスへの適用に備える

既存の意思決定ポイントを自動化する場合でも、全く新しい製品／サービスを実現する場合でも、機械学習の効果は“破壊的”です。そのため、既存のビジネスプロセス、機能、役割に及ぶ可能性のある影響やその意味合いを事前に評価することが重要です。これは、活用を開始する前の段階で、機械学習による変革の全体像を構想しておくべき、という意味ではありません。しかし、簡単な「現状チェック」を行っておけば、後からコストがかさむ可能性を軽減できます。計画を立てる際は、次のように自問することから始めます。

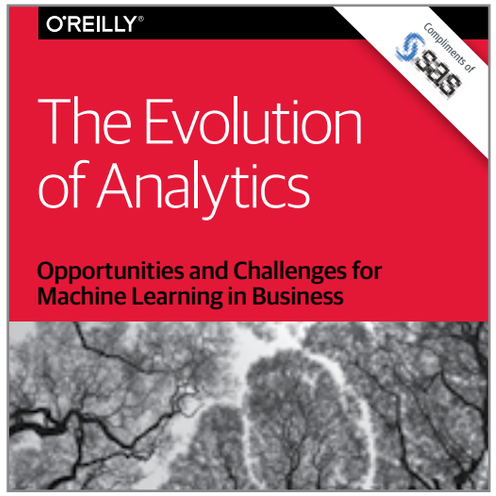
「この疑問への答えが得られる場合、あるいはこの仮説にもとづく予測が得られる場合、その情報で我々は何を実現できるだろうか?」、「これは既存のプロセスにどのような影響を及ぼすだろうか?」、「必要な変革を断行する意志と能力が我々にはあるだろうか?」

☐ 新たなITプラクティスの導入に備える

現場への展開が済んだ後も、機械学習モデルの反復的モデリングとチューニングを継続する必要があります。モデルの更新がどのような間隔で必要になるかは予測不能であり、従来のような日程計画ベースの展開管理手法には適合しません。そのため、機械学習モデルの展開と維持管理には、従来のDevOps型のITプラクティスとは根本的に異なるQAモデル、展開モデル、スキルセット、サービスレベルが必要です。

関連資料

学習をさらに進めたい場合は…



アナリティクスの進化：ビジネスにおける機械学習のチャンスと課題（英語版）

最先端の機械学習アプリケーションがどのようにビジネス価値を提供しているかを詳しく知ることができます。データに潜む洞察を発見するために機械学習を活用している組織がどのようなステップを実践しているかに注目した、2つのケーススタディが紹介されています。

コグニティブ・コンピューティング：経営幹部向けガイド（英語版）

SAS の EVP 兼 CTO であるオリバー・シャーベンバーガー（Oliver Schabenberger）が、コグニティブ・コンピューティングに関する疑問を解消します。この最先端の機能をいつ、どこに適用すれば最大限の効果が望めるのかについて、理解しやすい事例と重要な教訓が紹介されています。

sas best practices

THOUGHT PROVOKING BUSINESS

著者紹介

キンバリー・ネバラ（Kimberly Nevala）は、SASベストプラクティス部門のビジネス戦略担当ディレクター。現場に即した助言を世界中のクライアントに提供してきた19年間の経験を生かして、企業や組織がデータに潜む可能性を最大限に活用しようとする取り組みを支援しています。ビジネス・インテリジェンスおよびアナリティクス、データ・ガバナンス、データ管理の領域で、市場分析、業種別の教育、最新のベストプラクティスと戦略を統括しています。

講演者や執筆者としても活動しており、戦略イネーブルメントや組織の力学といったテーマで頻繁にコンサルティングを求められています。彼女の執筆物やインタビューは、CInformation Week、CIO Asia、Knowledge World、TDWIなどにも掲載されています。SAS Best Practicesペーパーでは「The Anatomy of an Analytic Enterprise」、「Sustainable Data Governance」、「Top 10 Mistakes to Avoid When Launching a Data Governance Program」も執筆しています。

SAS Institute Japan株式会社　www.sas.com/jp

jpnsasinfo@sas.com

本社　〒106-6111 東京都港区六本木6-10-1 六本木ヒルズ森タワー 11F
Tel: 03 6434 3000　Fax: 03 6434 3001

大阪支店　〒530-0004 大阪市北区堂島浜1-4-16 アクア堂島西館12F
Tel: 06 6345 5700　Fax: 06 6345 5655

109066_G61378.0917_JP2017BPeB_FK