@IT
atmarkIT

エンジニアが知っておくべき AI 倫理

保科 学世、右衛門佐 誠、張 瀚天、関 大吉／
アクセンチュア[著]

なぜ今 AI 倫理なのか？　AI がもたらす「意図しない結果」を防ぐために開発者ができること

正しく AI を作り、活用するために必要な「AI 倫理」について、エンジニアが知っておくべき事項を解説する連載。初回は、AI の普及により浮き彫りになった課題と、AI 開発プロセスに内在するリスクについて。

保科 学世 , 右衛門佐 誠, アクセンチュア（2022 年 02 月 16 日）

スマートフォン、スマートスピーカーに搭載されている音声アシスタントやお掃除ロボットなど、身の回りのさまざまなところに AI（人工知能）が導入され、私たちの生活は今まで以上に便利になっている。だが、新しい技術であるが故に、AI のブラックボックス化や AI による差別の助長など、今までなかったような新たな課題が見えつつある。本連載では正しく AI を活用するための手法である「AI 倫理」について解説する。

今回はまず、AI の普及により浮き彫りになった課題と、AI 開発プロセスに内在するリスクを開発フェーズごとに整理する。

AI がもたらした「意図しない結果」　どう防ぐ？

AI の普及が進むにつれ、さまざまな課題が顕在化してきている。中でも、「Black Lives Matter」運動に後押しされ、顔認証 AI への批判が集まったり、サービス停止に追いやられたりしたケースは記憶に新しい。以前から肌の色などで AI の認識精度に差があることや、黒人画像をゴリラと誤認識するなど顔認証 AI に関する課題は指摘されていた。それがプライバシーや AI の公平性、透明性の問題に関する昨今の世の中の関心と相まって大きな社会問題となったのだ。

IBM や Amazon.com などは顔認証ソフトウェアを捜査機関などに提供することを停止すると発表し、EU や米国では新たに法執行機関における顔認証 AI の利用禁止の規則が制定された。誤認識を引き起こす問題は差別以外にもさまざまなものがあり、AI を使ったサービスにおいて誤認識への対応は避けることのできない問題である。

もう 1 点、意図していなかった結果を AI が引き起こしてしまった例として触れておきたいのが、自動運転における「トロッコ問題」だ。「ブレーキが壊れたトロッコが暴走している。線路の先には 5 人の人間がいて、そのままでは間違いなく 5 人をひき殺してしまう。トロッコの進路を変えることができて、変えた場合の線路の先には 1 人の人間がおり、その人間をひき殺すことになる。トロッコの運転手は進路を変えるべきか否か」という、思考実験としても有名な問題である。

この問題が注目されるのは自動運転車（ここでは完全自動の自律運転車とする）が事故をどう回避するかの議論とも関係するからだ。「唯一無二の正しい答え」はなく、人間ですら国や地域、文化、宗教などによって回答の傾向が異なることが知られている。答えのない問題や国や地域によって答えの異なる問題に AI がどのような正解を出す必要があるかということも課題である。

このような状況下で、欧州を中心として AI の利活用に関する法律やガイドラインを整備する流れが加速している。詳しくは次回以降で解説するが、AI の利活用で生じるリスクに対する適切な対応などの安全性向上と、AI 利活用の促進を通じた競争力強化を両立させるものとして位置付けられている。

ではどのようにして AI の安全性と AI を用いた成長を両立していけばよいのか。単に精度だけを追い求めて AI を開発していくだけでなく、いかに社会に受け入れられる AI を開発していくか、その際の鍵となるのが AI 倫理である。

AI の開発、利用において、あらゆる場面での意思決定の中心は人間であるべきだ。そのため人に対しての行動指針、すなわち AI 倫理が必要といえる。本連載では、AI 倫理に基づいて、企業が顧客や社会に対して AI の公平性や透明性を担保する方法論を「責任ある AI」と位置付けて解説する。この方法論に基づいて AI を設計、構築、展開することで、真に人間中心の AI 活用を目指すことが可能となる。つまり AI 倫理によって企業は AI の持つリスクを正確に理解できるようになり、かつ AI が持つ潜在的リスクへの対策を行うことで AI への信用が生まれる。その信用が形成されて初めて、人間は AI を信頼できるようになるのである。この信用と信頼こそが、AI を自社のビジネスに応用し、拡大利用するための礎になるのだ。

アクセンチュアでは責任ある AI を形作るための行動原則（TRUST）として次の 5 つを提唱している。

T：信用できる（Trustworthy）

AI の設計・構築時、安全性を重視して物事に誠実に向き合い、多様で広い視点を持つ、という実績を一つ一つ積み上げる

R：信頼できる（Reliable）

積み上げられた信用から、将来の高度な判断とより良い意思決定への支持を集める

U：理解できる（Understandable）

信用を得るためには AI が透明性を持ち、人によって解釈可能にする

S：安全が保たれている（Secure）

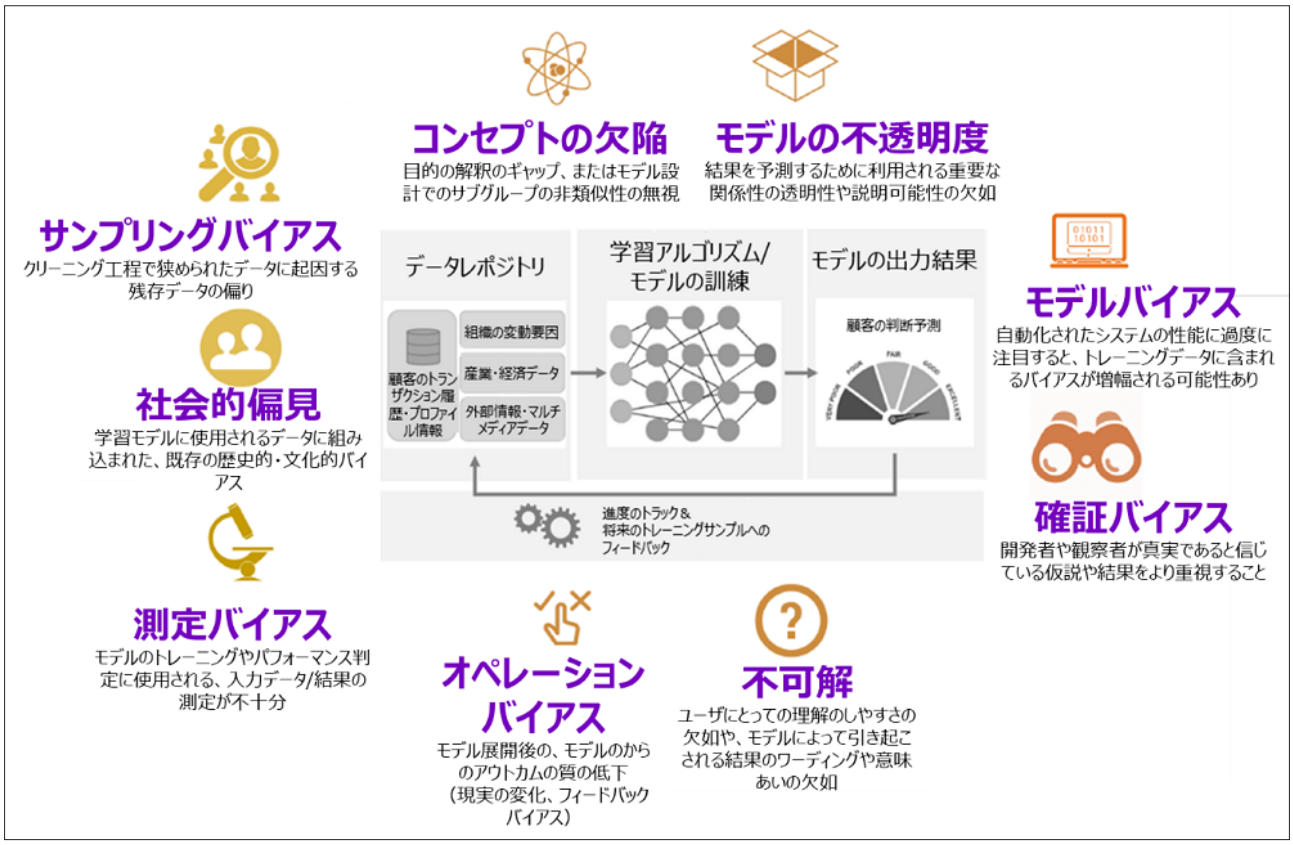
信用を得るためには企業や顧客の情報、データのプライバシーに配慮し、安全性を確保する

T：共に学びあう（Teachable）

信用、信頼を勝ち得たAIと人間とが情報交換し、共創し、相互教育をする世界を実現して人間中心のデザインを目指す

何をどう対処すべきか？　AI開発の段階ごとに解説

続いて、具体的に何をどう対処すればよいのか、AIの開発プロセスであるデータ収集、AI設計、AIモデル学習・評価、AIモデル運用について、ステップごとに解説していく。



AI開発プロセスで発生するバイアスリスク（提供：アクセンチュア）

データ収集

社会的偏見とサンプリングバイアス

データ収集の段階では、AIモデルの学習のために集めたデータそのものにバイアスが含まれているという課題がある。具体的には「社会的偏見」や「サンプリングバイアス」だ。社会的偏見とは人種、性別、職業など社会的に存在しているギャップがデータの中にバイアスとして入り込んでしまうことである。

例えば、与信審査を担当するAIが、女性の貸出限度額を男性よりも少なく見積もってしまうという事象はこの社会的偏見によるデータバイアスである。特にAIモデルは後述するモデルバイアスによりデータの持つバイアスを増幅する傾向があるため、取得データに社会的偏見が入りこまないよう注意が必要である。

サンプリングバイアスは、データ取得時やデータ加工時に、取捨選択されたデータが入りこんでしまうことが原因で起こる。例えば、年齢による身体能力の衰えを予測するため体育館にはり紙をして多数の人から体力測定のデータを取得し、AI モデルを作ったとする。この場合、健康意識の高い属性の人々が体育館を利用していることが想像されるため、世の中の平均の身体能力よりも高い基準で予測する AI モデルができるだろう。

データ加工時やデータ変換の際はバイアスが入り込みやすい。分析者が先入観を持ったデータ分類を行ったり古いデータを使い続けたりすることによって発生するバイアスもある。

AI 設計

コンセプトの欠陥

本来の目的と AI 設計との間に生じたずれが主な原因で、コンセプトの欠陥が生じる。特に、データの中のグループ間に差異がある場合には AI を利用する中でそのバイアスを増幅することにもなりかねないので注意が必要である。

例えば、犯罪防止のために犯罪予測 AI を利用したと考えよう。地域による犯罪率に偏りがあると、特定の地域の警戒を強める。そのためその地域の犯罪検挙率が向上し、その地域の犯罪率が（実質変化していなくても、数値上は）さらに悪化してしまう、とった具合に結果的に負のフィードバックになってしまうことがある。

確証バイアス

確証バイアスとは、真実であると信じている仮説や結果をより重視することにより、紛れ込んでしまうバイアスのことである。人間は無意識に自分にとって都合の良い情報を優先し、そうでない情報を軽視する傾向があるため、先入観によってモデル設計の中でバイアスが発生するリスクがある。

有名な例としてよく取り上げられるのが、チョコレートの年間消費量が増えるとノーベル賞受賞者数が増えるという分析結果だ。冷静に考えると直接的な相関ではなく国の裕福度などによる疑似相関であると分かりそうだが、分析者がチョコレートに含まれるフラボノールで認知機能が向上するという仮説にとらわれるあまり、客観的なデータ分析ができなかったという事例だ。

このように分析、設計の段階で作業者の持つバイアスにより、誤った結果を導くリスクが大いにある。そのため分析者は論理的に分析仮説を構築するとともに、先入観なくデータに向き合うことが重要である。

学習、評価段階

モデルの不透明度

モデルの不透明度とは、モデルの入力と出力との関係性に対する透明性や説明可能性を指す。透明性の高いモデルにするためには、入力データの各項目が出力結果に及ぼす影響について理解しやすいことが重要だ。つまり、結果が悪かった際に入力データのどの項目が影響しているのか、解釈できるようにしなければならない。

一般にディープラーニングなどの複雑なモデルは予測精度が高い一方で透明性が低く、ブラックボックスになりやすい。こういったモデルに対しても近年は「SHAP」（SHapley Additive exPlanations）や「LIME」（Local Interpretable Model-agnostic Explanations）といった予測根拠を可視化する手法が開発されている。Google 傘下の DeepMind は眼底画像の AI 診断で、単に疾患を判別するだけでなく、予測に寄与した特徴量の可視化も含めたディープラーニングのモデルを開発している。今後、医療など複雑な判断を必要とする場面での AI 活用は一層進むだろう。モデルの精度と説明可能性の両立はますます求められていくことになる。

モデル自体のバイアス

測定バイアス

測定バイアスとは、モデルの学習や評価に用いるデータが不十分である際に紛れ込むバイアスである。データ数が少ない状態でモデルに学習させる過学習（特定のデータにモデルがフィットし過ぎて汎用《はんよう》性を失った状態）が起きる可能性が高く、モデルの精度検証時のデータ分布が実際の分布と異なるために正確な精度検証ができないリスクがある。

モデルバイアス

モデルバイアスとは、学習データに含まれるバイアスがモデル構築の過程で増幅されることである。構築した AI モデルの精度が高かったとしても、それがバイアス増幅によるものである場合には、特定グループの優遇や機会損失、差別の助長につながることがあるので注意が必要だ。

社会の中にある偏見を取り除くことは容易ではない。AI モデルの改良だけで対応するには限界があり、利用方法の工夫など別の対応方法も考える必要がある。過去に Google 検索のオートコンプリート機能で偏見を伴う予測が表示されることが問題視されたが、Google は議論を呼ぶ可能性のある単語をブロックした。このようにリスクへの対応は単にモデルを改善するだけでなく、運用も含めサービス全体で考えていく必要がある。

AI モデル運用

オペレーションバイアス

AI モデルの運用の段階では、運用中に徐々にモデルの質が低下していくオペレーションバイアスのリスクがある。モデルを学習した段階と現時点でのデータに差が出てきた場合にこのようなバイアスが生じるので、マーケティングなどデータの分布が変化しやすい領域では注意したい。

ユーザーや環境からのフィードバックを新たに学習データに追加して学習をし直したり（再学習）、既存モデルのチューニングを行ったり、フィードバックを使って AI の挙動を改善（強化学習）したりする取り組みも多いが、こういったフィードバックによって混入するフィードバックバイアスもオペレーションバイアスの一種だ。

ここまで具体的な例も取り上げつつ AI 開発の中に内在するリスクについて解説してきた。次回はこれらのリスクへの対処法を具体的な技術手法も交えつつ解説していく。

筆者紹介

保科 学世（ほしな がくせ）

アクセンチュア ビジネス コンサルティング本部 AI グループ日本統括 兼 アクセンチュア・イノベーション・ハブ東京共同統括 マネジング・ディレクター 理学博士。

アナリティクス、AI 部門の日本統括として、AI Hub プラットフォームや AI Powered サービスなどの各種開発を手掛けるとともに、アナリティクスや AI 技術を活用した業務改革を数多く実現。『責任ある AI』（東洋経済新報社）はじめ著書多数。

右衛門佐 誠（よもさ まこと）

アクセンチュア ビジネス コンサルティング本部 AI グループ プリンシパル 工学博士。大阪府立大学大学院博士後期課程修了。

製造業をはじめとする画像分析案件や AI やデータサイエンスに関する人材育成などの案件に従事。企業だけでなく大学との共同研究や講座提供なども担当。

AI がもたらす「不公平な結果」を防ぐために開発者は何を考慮すればよいのか

正しく AI を作り、活用するために必要な「AI 倫理」について、エンジニアが知っておくべき事項を解説する本連載。第 2 回は、データ収集の際に起きてしまうバイアスの発生を防ぐために、AI 構築に携わる技術者がとるべき対応について。

保科 学世 , 張 瀚天, アクセンチュア（2022 年 02 月 22 日）

前回の記事では AI 開発中に内在する代表的なバイアスやリスクを整理した。解説したリスクはどれも重大な損失につながる可能性があるため、AI がもたらす効果と AI 倫理を整理した上で、実現する AI の構築が必要だ。一方で「これさえ押さえれば倫理的な AI の出来上がり」という解決法はない。AI 技術者は、数ある手法群から状況に適した技術を選定し、活用しやすい AI の構築を目指していくことが求められる。

今回は、前回整理したデータ収集の際に起きてしまいがちなバイアスの発生を防ぐために、AI 構築に携わる技術者がどのように対応すべきか、具体的な方法を解説する。

データ収集段階でのバイアスへの対処法

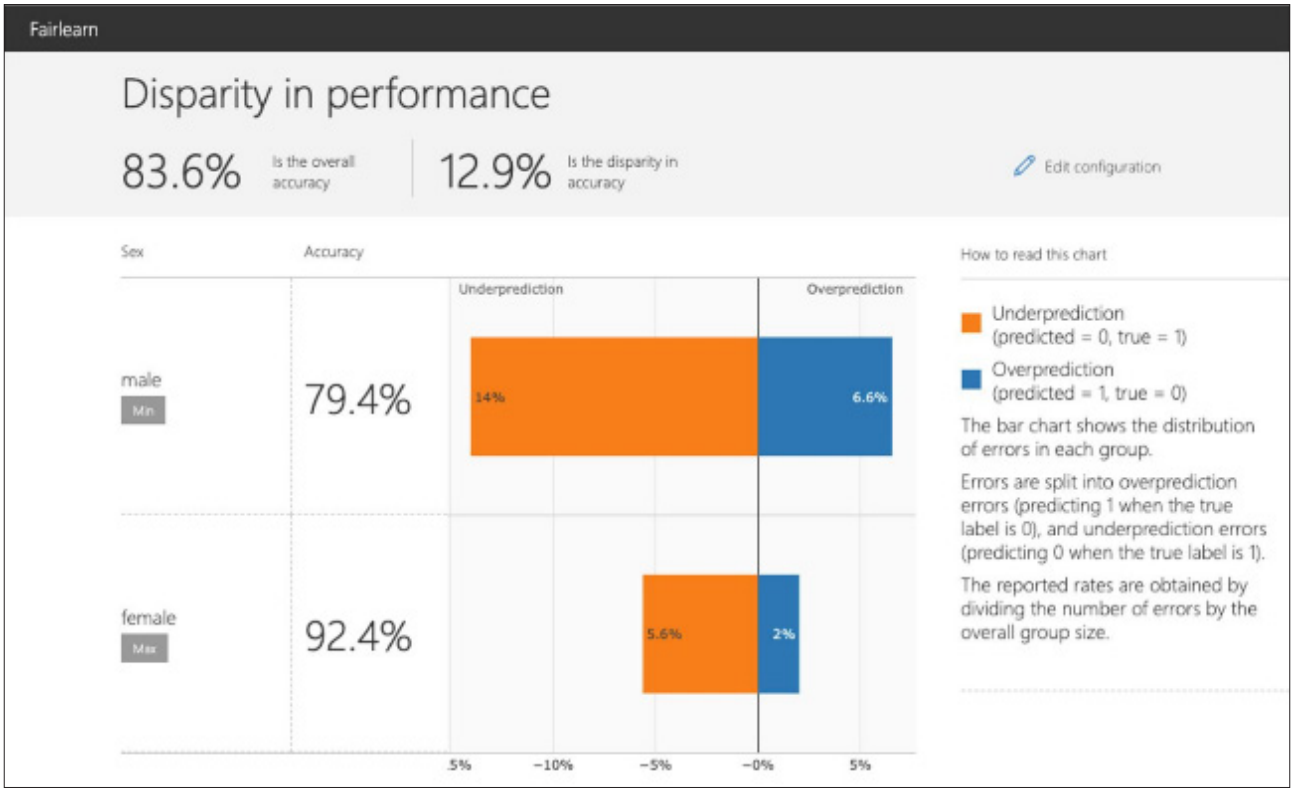
1. データが含む社会的偏見を公平性（Fairness）に配慮した AI で是正する

AI モデル学習に用いるデータセットが、人種や民族、性別などによる偏見やギャップの影響を受けている場合には、それに基づく AI モデルも容易に影響を受けてしまう。データが含む偏見やギャップを忠実に再現してしまうこともあるだろう。そのような課題を解決するには、単に、人種、民族、性別といったセンシティブな属性をモデルに学習させない、というだけでは不十分だ。

なぜなら、センシティブな属性と関連する特徴量を AI が学習することで、間接的にセンシティブ情報が AI モデルの予測結果に影響を与えてしまうからだ。この現象は「Red-Lining 効果」と呼ばれる。過去に米国で、低所得層の特定人種が居住する地域に対して、金融機関が融資を避けるという差別が発生した。その際、黒人居住区を地図上に赤線で囲っていたことがこの現象の名前の由来だ。この場合、適切な対策を講じないと人種に基づく偏見と居住地データが間接的にひも付いているため、居住地データを用いて AI モデルを構築してしまうと人種に基づく偏見を持った AI が構築されてしまう。

では、どのように対処すればよいだろうか。AI モデルにおいて差別や偏見を取り除くためには、まず差別や偏見がない公平な状況を定義することから始めることが必要だ。

以下の図は公平ではない AI モデルの予測精度を示している。センシティブ属性に依存して AI モデルの予測値が変わらないといったような状態は、公平な状況といえるだろう。しかし、性別によって正答率が 10%以上変わってしまっているため、正答率で比較すると AI モデルによる予測が性別によって不公平な状況となっている。



公平性ツールの一つである Microsoft「Fairlearn」の評価ダッシュボード例（出典：Fairlearn v0.5.0）

公平ではない AI モデルの分かりやすい例を紹介したところで、改めて AI における「公平性」とは何かについて掘り下げたい。対象とする問題や状況も考慮すべき要素ではあるが、ここでは代表的な集団公平性の定義である「民主的公平性（Demographic Parity）」と「均等オッズ（Equalized Odds）」について解説をする。

簡単に説明するため、センシティブな特徴を 2 値変数の $S \in \{0,1\}$ とし、$Y \in \{0,1\}$ を予測する 2 値の分類問題対象に説明する。

予測値を

$$\hat{Y}$$

とすると、民主的公平性は、以下のように表される。

$$Pr(\hat{Y} = 1 | S = 0) = Pr(\hat{Y} = 1 | S = 1)$$

これはセンシティブ属性にかかわらず、AI モデルにより推定された確率

$$Pr(\hat{Y} = 1)$$

→目次に戻る

が不変であるという定義となる。つまり、民主的公平性を保持できていたとすると、人種や性別といったセンシティブ属性によって予測スコアの分布が変わらないことになるので、その意味で公平な AI といえる。

一方、民主的公平性は本来の目的変数の分布を無視していることから逆差別につながりやすい指標ともいえる。それらの問題を解決する、AI における集団的公平性定義が、均等オッズだ。均等オッズは、以下のように表される。

$$Pr(\hat{Y} = 1 | S = 0, Y = y) = Pr(\hat{Y} = 1 | S = 1, Y = y), y \in \{0,1\}$$

一見難解であるが、これは y=1 の場合、左辺と右辺はともに真陽性率（True Positive Rate）を表しており、センシティブ属性ごとに真陽性率が変わらないことを示している。また、y=0 の場合には偽陽性率（False Positive Rate）がセンシティブ変数によらず均等であるという定義となる。つまり、センシティブ属性によって、真陽性率、偽陽性率といった精度の違いが存在しないことが定義となっている。

その他にも均等オッズを y=1 のみの場合に限定した「Equal Opportunity」という公平性の定義も存在する。機械学習において、y=1 の場合に治療や採用など、機会が与えられることが多いことに由来しており、特定クラスに対する公平性を担保したい場合に有用な定義である。こうした公平性の定義に関しても AI を構築する際には、適切な公平性の定義を選択していく必要がある。

ここまで公平な AI を実現するための指針となる定義を解説した。次にその定義を直接または間接的に実現するアプローチについて紹介する。

選択した指標に沿って公平な AI を実現するためには、a. 学習前にデータセットをより公平にする「Pre-Processing」、b. 公平な学習を AI モデル内部の学習アルゴリズムで行う「In-Processing」、c. 学習した AI モデルの予測値やモデルアウトプットを補正する「Post-Processing」の 3 つのアプローチが存在する。

a. 学習前に公平な状態を目指す Pre-Processing

Pre-Processing は、モデル学習前に、より公平な状態を目指すアプローチだ。例えば、センシティブ属性と相関性が高い特徴量を除くのも 1 つの手法だ。センシティブ属性と関係が強い特徴量を除去することで、センシティブ属性と目的変数の暗黙的な相関を緩和することができる。

例えば、Pre-Processing の 1 つ、「Reweighing」という手法では、それぞれのセンシティブ属性とクラスの組み合わせにおいて、公平な状態での出現頻度と現在の出現頻度の比から「公平な状態からどれほどずれているか」を算出する。その比を AI モデルの損失関数に重みとして利用することで、本来公平な状態での出現頻度を加味したうえでモデルを学習できる。データを改変せずに、公平に学習させられるため、解釈性を損なわない手法だ。

b. 学習時に公平な状態を目指す In-Processing

モデルが出した予測値と、実測値のズレを表現する関数を損失関数と呼ぶ。In-Processing では、この AI モデル内部の損失関数に損失項を加えることで、公平な状態を目指すアプローチだ。公平に近い場合には小さく、公平が保たれていない場合には大きく損失項を加えることで、実現したい公平性を直接扱うことが特徴だ。例えば、In-Processing のアプローチの一つである「Reduction Method」では、実現したい公平性基準（Demographic Parity など）を明示的に指定し、それが損なわれている程度を定量化し、制約として表現したうえで、学習を実行する手法である。

また、「Adversarial Debiasing」とよばれる手法も紹介したい。この手法は、生成モデルの「Generative Adversarial Networks」などで用いられる「Adversarial Training」を用いる。Adversarial Debiasing では、まず片方のニューラルネット（モデル A）で、通常の目的変数を予測するモデルを構築した上で、もう片方のニューラルネット（モデル B）で、モデル A の予測値を入力として、センシティブ属性を予測する。モデル A は、目的変数への予測誤差を最小化するように学習すると同時に、モデル B の損失が最大化されるように学習を進める。

つまり、学習が進んだ理想的な状態としては、モデル A は高精度予測が可能である一方で、その予測値は非線形変換を加えたとしてもセンシティブ属性を予測し得ない状態となり、実現したい公平性基準（Demographic Parity）が実現されることに相当する。この枠組みを用いればさまざまなタスクに応用できる上、実現したい定義についても、例えば均等オッズを満たしたい場合には、モデル B に対して目的変数を実数値 y と入力とすることで実現できるため、自由度高く公平性を扱うことができる。

c. モデル構築後に公平な状態を目指す Post-Processing

Post-Processing は、AI モデルが構築された後にモデル出力をより公平に補正することを目指す。

例えば、2 値分類で予測されたスコアを基に、実現したい公平性（Demographic Parity や Equalized Odds）を最大限実現するような予測確率の閾値を探す手法（Threshold Optimization Approach）などが挙げられる。この手法では、公平性を実現する最適な値を線形計画問題により探索する。

上記の 3 つのアプローチはそれぞれ適する場面が異なるため、AI モデルを構築する際に、適切なアプローチを選択することが求められる。Pre-Processing はモデル学習前に実行するため、その後のモデル選択や精度評価に対して比較的自由度が高い。また、Post-Processing についても既存のモデルが使えるため自由度が高い。

一方で、これら 2 つは AI モデルの解釈性を損なってしまう可能性がある。In-Processing に関しては、公平性の基準を損失関数に加えるため直接公平性を扱えるメリットがあるものの、使える AI モデルやアルゴリズムが限られるためデメリットが存在する。

特に分類問題に焦点を当て紹介したが、その他にも公平な回帰モデル、生成モデルなどのさまざまな領域に対して研究が盛んにされている。詳細なアルゴリズムや、興味がある方は後述の参考文献を参考にしていただきたい。上記のアルゴリズムについて OSS（オープンソースソフトウェア）の Python ライブラリである「Fairlearn」や「aif360」により、機械学習のライブラリ「scikit-learn」に準拠して実装されているので、scikit-learn に慣れていればすぐに活用できるだろう。

2. サンプリングバイアスを緩和する

サンプリングバイアスとは、取得できているデータや対象に偏りがあるため、母集団に代表性がないことを指す。社会的偏見と同様にデータに混在しているバイアスといえる。サンプリングバイアスがある状況下で、AI モデル構築を行うとデータが取得できている集団の影響を強く受けやすくなってしまう。

サンプリングバイアスに対処するとき、実現したい AI 像を明確にする必要がある。もし特定のセンシティブ属性に対してのみデータが取得しやすく、それによって精度が変わることを避けたいのであれば、公平性の指標として均等オッズなどを基準にセンシティブ要因に対して精度が変わらないように AI モデルを構築することが有効だ。

一方で、母集団全体に対して代表性を持つモデルを構築する文脈では、観測のしやすさをモデル化することが有効だ。例えば、あるサービスの契約者に Web 上でアンケートを送付し、そのアンケート結果を基に AI モデルを構築する場面を想定しよう。Web に慣れている若年層やサービスへのロイヤリティーが高いユーザーからの回答が多くなり、AI モデルは偏ったデータを元に学習してしまう。

この場合、母集団と比べて観測が用意なユーザーのデータが偏ってしまう。対処法として、AI モデルで最適化させるべき損失関数を「観測のされやすさ」（上記の例では、アンケートの回答されやすさ）の逆数で補正をかけることで、サンプリングバイアスの影響を緩和することができる。本来推定したい母集団を推定する AI モデル構築の方法は、以下の通りだ。

1. 観測のされやすさを分類するような 2 値分類モデルを構築する
2. 構築した 2 値分類モデルによる、観測される予測確率を各サンプルに対して算出する
3. 各サンプルに対して、算出した予測確率の逆数を損失関数に対して重み付けたモデルを構築する

Python での実装時は、観測のされやすさを scikit-learn のロジスティック回帰実装である「LogisticRegression」などの 2 値分類モデルにて構築し、その予測確率の逆数を算出する。そして算出した重みを、scikit-learn の AI モデル実装における「fit」関数の引数である「sample_weight」に指定する。そうすることで、損失関数に対して重み付けることが可能になる。

scikit-learn 以外にも、AI モデル構築で頻出となる勾配ブースティング木モデルライブラリの「XGBoost」や「LightGBM」などにおいても「sample_weight」により損失関数の重みを指定することが可能であるため、容易に利用可能だろう。

次回は AI モデル学習、評価段階や運用時のバイアスへの対処法について解説する。

筆者紹介

保科 学世（ほしな がくせ）

アクセンチュア ビジネス コンサルティング本部 AI グループ日本統括 兼 アクセンチュア・イノベーション・ハブ東京共同統括 マネジング・ディレクター 理学博士。

アナリティクス、AI 部門の日本統括として、AI Hub プラットフォームや AI Powered サービスなどの各種開発を手掛けるとともに、アナリティクスや AI 技術を活用した業務改革を数多く実現。『責任ある AI』（東洋経済新報社）はじめ著書多数。

張 瀚天（ちょう かんてん）

アクセンチュア ビジネス コンサルティング本部 コンサルタント

筑波大学大学院卒業後に、アクセンチュア入社。ヘルスケアや通信など複数業界でのアナリティクス適用、AI モデル構築や運用による業務改革を支援。

AI モデル学習の評価時／オペレーション時に発生するバイアスリスク、どう対処する？

正しく AI を作り、活用するために必要な「AI 倫理」について、エンジニアが知っておくべき事項を解説する本連載。第 3 回は、AI モデル学習の評価時、オペレーション時のバイアスリスクへの対処法について。

保科 学世 , 張 瀚天, アクセンチュア（2022 年 03 月 15 日）

前回はデータのバイアスリスクへの対処法を、具体例を交えて解説した。今回は、AI モデル学習の評価時、オペレーション時のバイアスリスクへの対処法について解説する。考え方や手法、ツールが AI 開発の一助になれば幸いだ。

AI モデルの不透明さ、不可解さをどう解決するか

AI モデルにおいて、「精度は高いがなぜその出力がなされているか理解しがたい」といった問題が発生し得る。特にディープラーニングなどの AI モデルにおいて、モデルの入出力関係が解釈しづらく、ブラックボックスとなり、AI が誤った基準で判断をしているかを解釈できないケースが存在する。例えば、既存の知見と異なる判断基準が学習されているケースや、人種、民族、性別、年齢、信仰などの公平性観点から差を生むべきではない、いわゆる「センシティブ属性」が AI の判断基準に使われてしまっているというケースである。そうした問題を検知し解消するため、モデルの解釈性を向上させることが重要だ。

モデルの解釈性を向上する方法は大きく分けて 2 種類ある。「AI モデル自体の解釈性を向上させる方法」と「学習したモデルを解釈する方法」だ。

1. AI モデル自体の解釈性を向上させる方法

A）条件分岐構造を持つモデル

モデル自体の解釈性を向上させるには、条件分岐構造を持つモデルが有効である場合が多い。条件分岐構造を持つモデルの代表例は決定木である。

前述の通り、解釈性の高いモデルを構築することで AI モデルが誤った判断をしているかを解釈できる。ここで、決定木の簡単な活用例を紹介する。

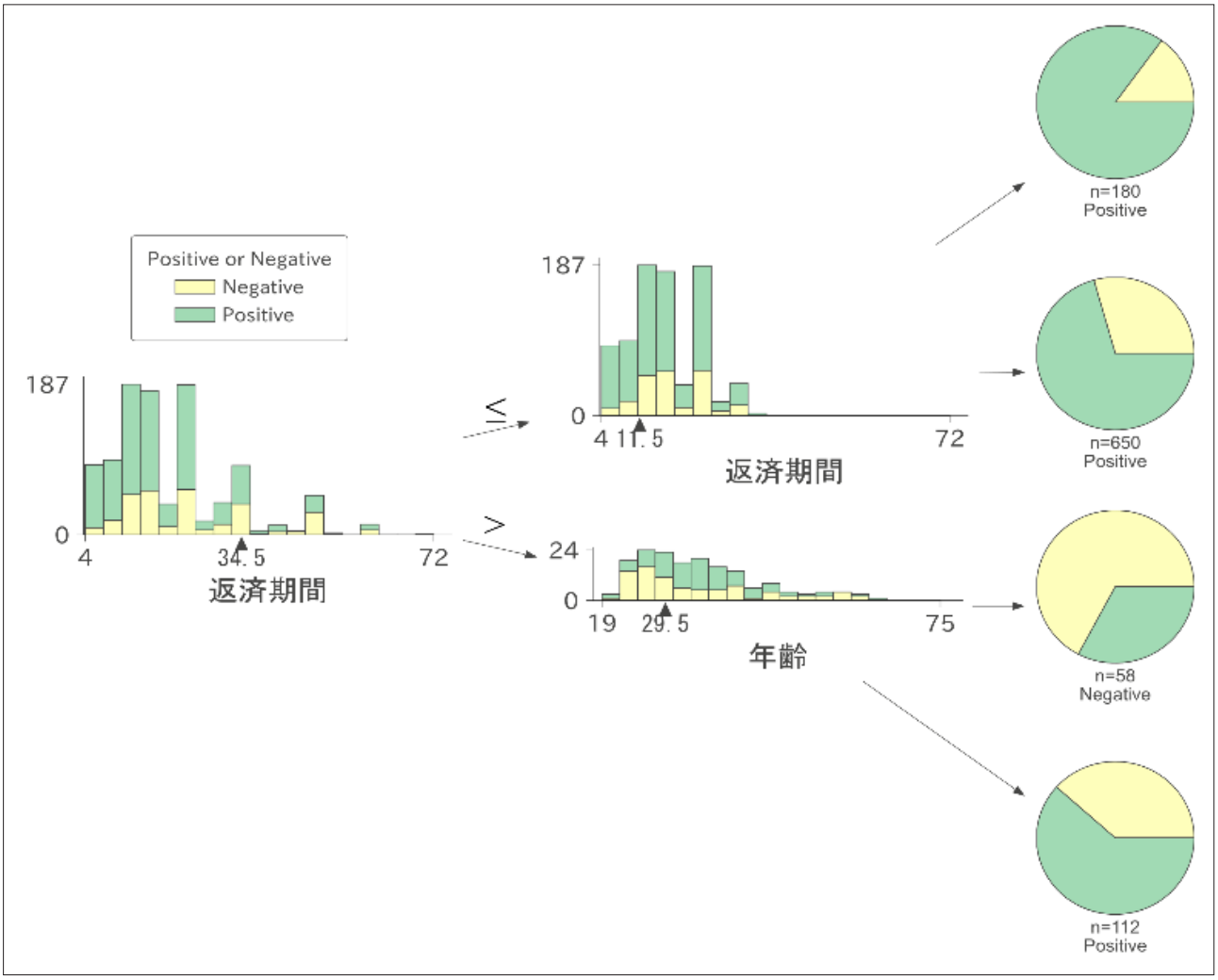


図 1：決定木による Statlog（German Credit Data）学習例（dtreeviz による可視化）（作成：アクセンチュア）

図 1 は 1970 年代ドイツの金融機関における融資リスクを、融資返済期間や年齢といった特徴（今回は例示のため、あえて年齢などセンシティブ属性をそのまま変数に加えている）から予測する決定木を構築した例である。

決定木は分岐ごとに条件が AND で足される構造となっている。上図を読み解くと、4 つのパターンが学習されている。返済期間が 11.5 カ月よりも短い場合に最も低リスク（Positive）であり、11.5 カ月超 34.5 カ月以下の場合も低リスクと予測されている。また、返済期間が 34.5 カ月以上と長期で、年齢が若い場合（29.5 歳以下）に特に高リスク（Negative）となっており、年齢がそれよりも高い場合には低リスクとなる。このように、決定木はデータから解釈可能な条件分岐構造を可視化できる。

一方で、モデル解釈した後に、次のステップとして考慮すべきことがある。それは、「年齢が若いと融資リスクが高い」とするモデル判断が妥当か否かである。

米国の金融サービスの公平性を担保するために作られた消費者信用機会均等法（ECOA）においては、「年齢」はセンシティブ属性と定められているため、年齢が判断基準に含まれるモデルは公平性を欠いたモデルといえる。

そのため年齢の影響を除いたモデルを構築する必要がある。しかし、ここで年齢を除いてモデルを構築するだけでは、年齢に関した暗黙的な関連がモデル内部に含まれ、公平でないモデルが構築される可能性がある。そのため、前回記事で解説した公平性を配慮した AI モデル構築技術の適用が必要になるのである。

上記例のように、決定木など解釈性が高いモデルは、モデル内部でどのように判断されているかを可視化でき、モデル判断が妥当かを考察することが可能になる。そして、その結果によってはバイアスを取り除く、といった改善プロセスを取ることができるため、よりバイアスリスクを抑えた AI モデルが構築可能になるのである。

また、今回の例においては、そもそも融資リスクを予測するようなモデルを構築すべきか、という根本的な問題がある。第 1 回記事で解説した、AI 設計時のバイアスリスクである「コンセプトの欠陥」や「確証バイアス」などを慎重に考慮しなければならない。

その他にも、条件分岐構造を持つモデルとして「RuleFit」がある。RuleFit では、勾配ブースティングなど、複数の木で構成されるモデルで学習した条件分岐構造を用いる。学習した条件分岐構造を特徴量として表現し、線形モデルとして学習することで、各条件分岐の重要度を示すことができる手法である。

この節で紹介した決定木は機械学習ライブラリ「scikit-learn」で公開されている。また RuleFit や可視化ツールである「dtreeviz」についてもオープンソース実装が公開されている。

B）加法的な構造を持つモデル

過度に複雑なモデルでない場合には、加法的なモデルでも解釈性を向上させることができる。加法的なモデルは、

$$g\big(E(y|X)\big) = f_1(x_1) + f_2(x_2) + \cdots + \beta$$

と表される。特徴量ごとに目的変数 (y) へ関係が表現されているため、各特徴量の目的変数への影響を理解しやすいモデルである。

加法的なモデルの代表例として、「線形回帰モデル」が挙げられる。線形回帰モデルは、目的変数を特徴量の重み付け線形和（f が重み係数）として表現するため、各変数の重みにより目的変数への影響度を把握できる。また、目的変数に対して適切な非線形変換（リンク関数 , g）をする一般化線形モデル（Generalized Linear Models）が存在する。

その他にも、より柔軟なモデルとして、各特徴量の関係 (f) についても、スプライン曲線など非線形関数を用いる「一般化加法モデル」（GAM、Generalized Additive Model）がある。

図 2 が GAM の学習例だ。これは、カリフォルニアにおける各地区の家の値段の中央値を予測するモデルを、平均住宅占有率（AveOccup）や築年数の中央値（HouseAge）、平均部屋数（AveRooms）などの特徴量から GAM により構築し、それぞれの影響度 (f) を可視化したものだ。青字が予測値であり、赤字は 95%信頼区間である。

結果を見ると、AveOccup が高くなるほど目的変数の値が下がり、HouseAge、AveRooms が高くなるほど信頼区間は広いもののおおむね目的変数が上昇する傾向が学習されている。

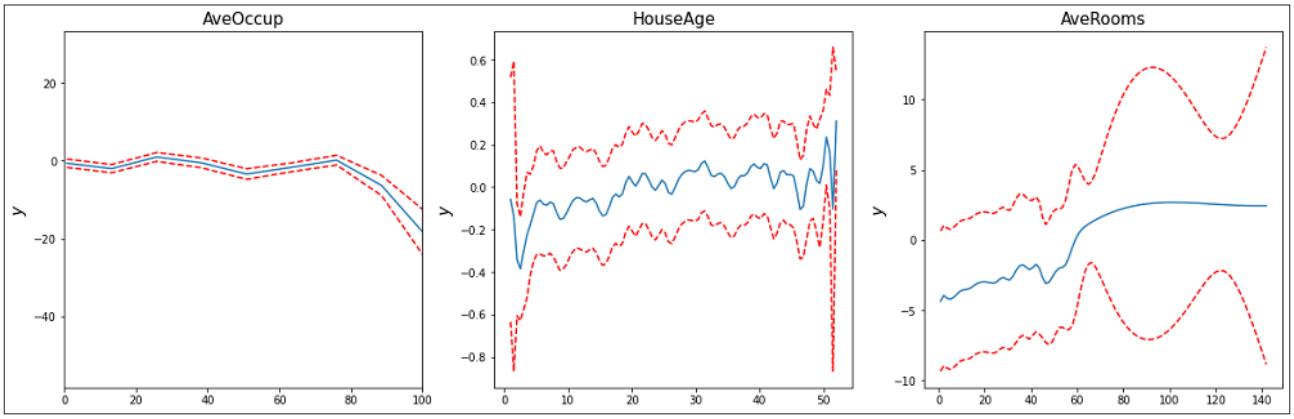


図 2：GAM による変数影響の可視化例（pyGAM）（作成：アクセンチュア）

このように、加法的モデルでは各項の影響を可視化できる。GAM や GLM に関して、pyGAM にて実装が公開されている。

C）関係性制約付きモデル

特徴量の目的変数へ与える影響が分かっている場合には、その知見を反映させたモデルを構築することで解釈性を向上させることができる。

例えば、販売価格が下がったら需要が増えるといった「単調性」や、年齢が上がるほど収入は上がるが一定の年齢を過ぎると低減するといった「凹凸性」などの関係だ。

単調性制約に関しては「XGBoost」や「LightGBM」などの代表的な機械学習ライブラリでもサポートしている。例えば、LightGBM については引数「monotone_constraints」に対して、各変数に単調性制約を課すかどうかを指定できる。前節で紹介した pyGAM も、単調性や凸（convex）と凹（concave）を制約として関係性 (f) を学習できる。

関係性制約を課して学習させた例を下図に示す。図 3 の左は log 関数で表現される目的変数に対して、凹性制約付き GAM、単調性制約付き LightGBM の学習例だ。右の図は目的変数に対して 2 変数がそれぞれ凹凸関係の場合に制約付き GAM で学習した例である。

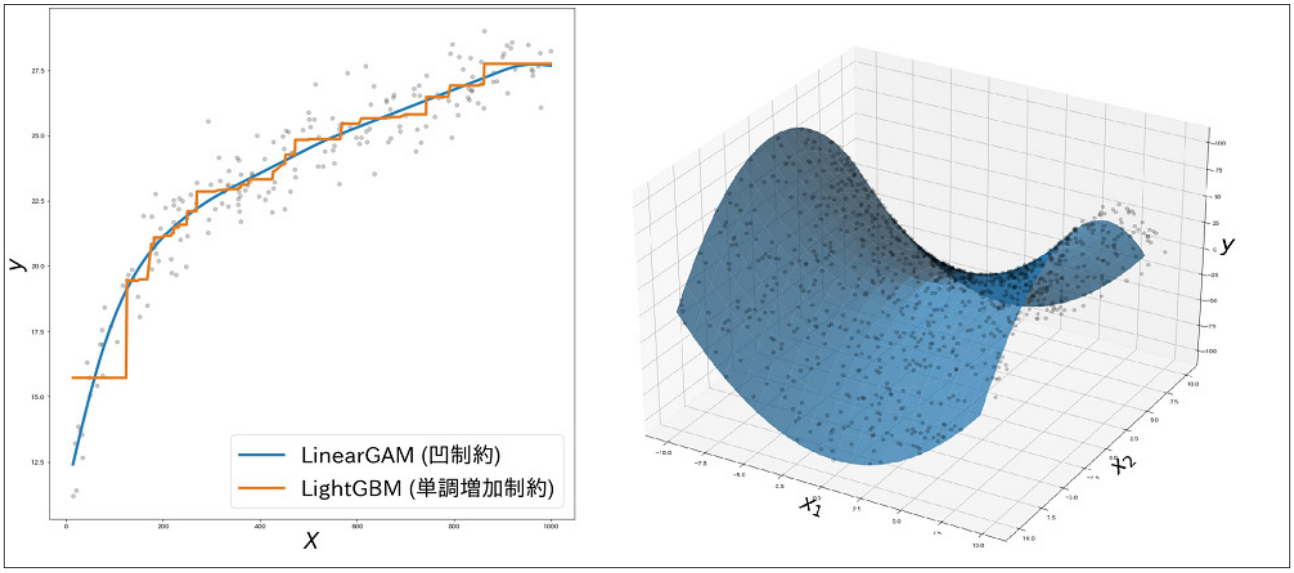


図 3：（左）凹性、単調性制約付きの学習例（右）凹凸制約を加えた GAM の学習例作成（作成：アクセンチュア）

このように制約を加えることで、データに対する知見に合うモデルを構築可能だ。また、凸性制約付きモデルは、凸最適化によって取り扱いやすい関数が学習されるため、最適化問題との親和性も高い。

ただし知見を入れることができるということの裏返しとして、その知見を重視してしまうような「確証バイアス」のリスクを意識しなければならない。

2. 学習したモデルを解釈する方法

A）モデル内部で寄与する特徴量の理解

予測モデルを解釈するには、AI モデル内でどの特徴量がどの程度寄与しているかを読み解く必要がある。変数の重要度を算出する手法として、「Permutation Feature Importance」が挙げられる。

これは、AI モデル学習後に各特徴量のデータをシャッフルして予測を実施し、「予測誤差が増加する度合い」をその特徴量の重要度とするシンプルな手法である。シャッフルしているのに予測誤差が増えない変数は重要度が低く、誤差が増える特徴量は重要度が高いと考えられる。

また、特徴量値とモデル予測値の関係を可視化する手法として、「Partial Dependence Plot」（PDP）が挙げられる。PDP は各特徴量値を単一の値で固定した場合の予測値の平均値をプロットしたものであり、各特徴量の値の大小で、どのように予測値が変化しているかを可視化できる（図 4）。

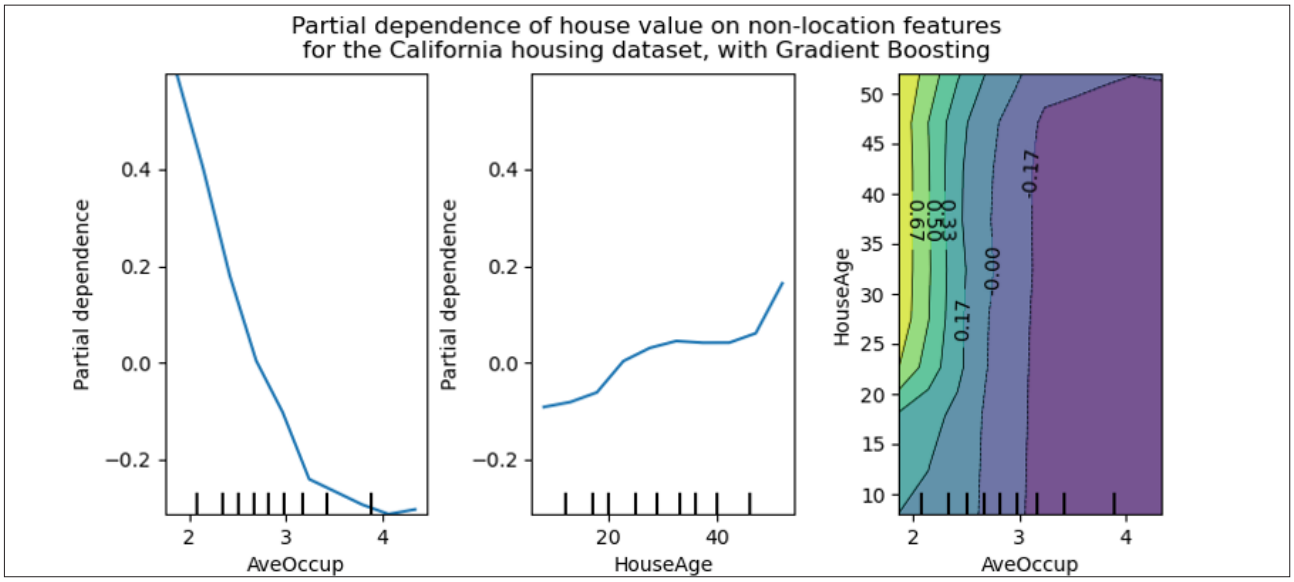


図 4：PDP の例（出典：https://scikit-learn.org/stable/_images/sphx_glr_plot_partial_dependence_003.png）

例えば、図 4 の左にある AveOccup については、値が上昇するにつれて予測値が減少し、一方で HouseAge が上昇すると予測値が増加する傾向が見られる。

PDP によって可視化するプロットは、GAM による可視化（図 2）と近い。PDP はモデル非依存という利点を持つが、特徴量の分布を無視して値を固定化させた場合の予測値を算出するため、現実にはありえないデータ分布となる可能性があり注意が必要だ。

上記で紹介した、Permutation Feature Importance や PDP は、scikit-learn に実装が公開されている。

B）SHAP

モデル解釈において代表的な手法が「SHAP」だ。勾配ブースティング木やニューラルネットワーク（の一部）など複雑なモデルに対して、「各サンプルにおいての各特徴量の寄与度」を算出できる。

図 5 は 1 サンプルに対する SHAP の例で、縦軸は特徴量、横軸は寄与度を表している。例えば、このサンプルに対しては特徴量 13 の値が 4.98 であることが予測を +5.79 と押し上げているが、一方で特徴量 6 の値が 6.575 ということが予測値を -2.17 に押し下げていると解釈する。

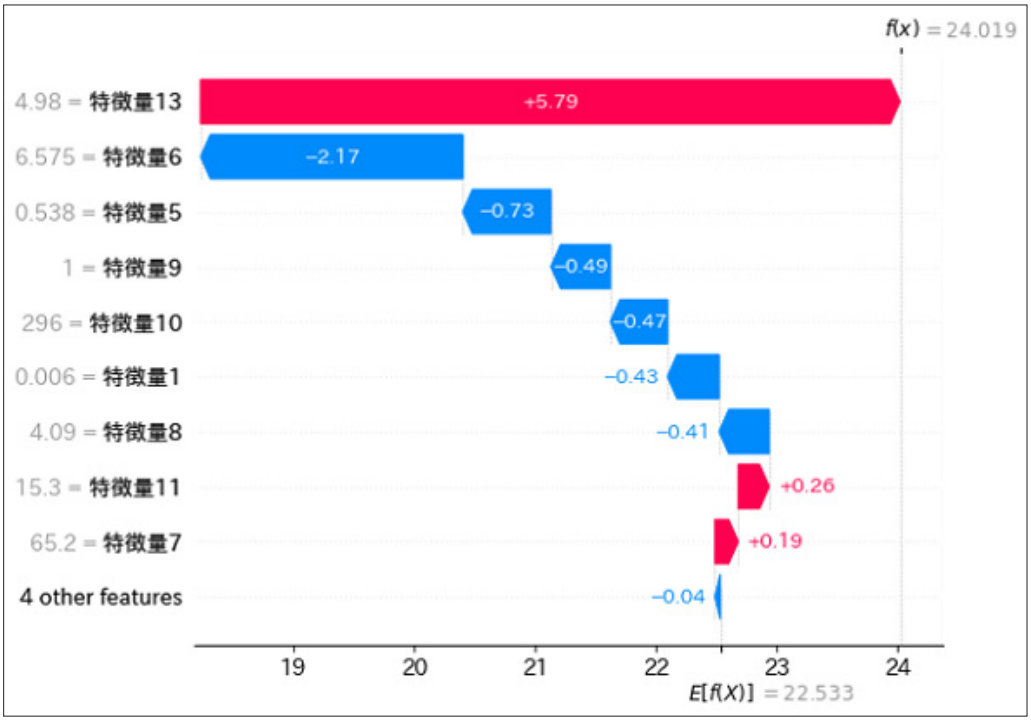


図 5：SHAP の推定例（作成：アクセンチュア）

このように各サンプルに対して各特徴量の寄与度を算出できる。

また、SHAP による寄与度は数多くの望ましい性質を持つ。例えば、寄与度（上例では +5.79 や -2.17 など）の合計は、そのサンプルの AI モデル予測値と一致する。つまり、SHAP による寄与度は予測値の分解として解釈できる。

また、全サンプルに対して変数値と寄与度をプロットすることで PDP のように特徴量値と予測値の関係を把握でき、各特徴量に対しての寄与度の絶対値の平均などを変数重要度として解釈できる。実装は「shap」ライブラリとして公開されている。

ここまでで、モデルの解釈性向上の方法を解説した。取り上げたもの以外にも、既知の特徴量間の関係を反映した予測モデルを構築できるベイジアンネットや、「各サンプル」の AI モデルにおける重要度を算出する影響関数（influence function）などが存在する。より詳細には後述の参考文献を見ていただきたい。

モデルバイアスへの対処

データに潜む社会的偏見などのバイアスがモデルによって増幅されてしまうことがあるため、前回紹介したように公平性に配慮した AI によりリスクに対処する必要がある。

さらに、目的変数のデータの質が担保できず、モデルの学習や評価が難しい場面がある。特に大規模データセットで、複数の担当者がラベルを付与する場合などは、ラベル（目的変数）の質を担保することが難しい。その場合に、目的変数にラベルの振り間違いなどのノイズが存在する前提で学習する手法がある。

その手法の一つが「Label Smoothing」だ。通常、クラス分類問題の目的変数は、各クラスに所属する場合 1 として、所属しない場合 0 の 2 値で表現するが、Label Smoothing ではその値をいわば「にじませる」のだ。0 か 1 か、と明確に 2 値に分けるのではなく、例えば、1 とラベル付けされているものを 0.9 とし、0 を 0.1 とするといったように、多少の振れ幅をもたせることで与えられた信頼できないラベルを過度に学習しないような頑健なモデルを構築できる。

その他に、「Confident Learning」という手法もある。Confident Learning では学習データの中で信頼できないラベルのサンプルを判別、除去した上で、再学習することで、より信用できるモデルを構築できる。

Confident Learning を MNIST（手書き数字分類）データセットに対して実施した例が図 6 である。赤枠で囲われているのがラベルの振り間違いが起きていると提示されたデータであり、与えられたラベルが given、予測されているラベルが guess である。

例えば左上の画像の場合、ラベルは 4 であるが実際の値は 7 であり、ラベルのミスが発生している。

図 6：MNIST において Confident Learning を実行した例（出典：https://github.com/cleanlab/cleanlab）

Confident Learning はモデル非依存手法だ。任意の分類モデルで利用でき、実装が「cleanlab」というライブラリとして公開されている。

AI モデル運用でのリスク、オペレーションバイアスとは

せっかく訓練させた AI モデルの精度や品質も、運用の段階で劣化してしまうリスクがある（オペレーションバイアス）モデルが学習したデータ分布や目的変数への関係式が、推論時、運用時に変化することが主要要因だ。例えば、気候変動で地球の気候そのもののパターンが変化していて、気温データに関する前提条件が変わるケースなど、数多くの事象が考えられる。

その中でも、特徴量データの分布が変わってしまう場合は共変量シフト、目的変数の分布が変わってしまう場合はターゲットシフト、特徴量データと目的変数の条件付確率が変わってしまう場合はデータセットシフト（概念ドリフト）などがある。これらのシフトに対応するには、大きく分けて 2 つの方法がある。

1 つ目の対処法は、再学習だ。定期的に、推論したときと異なる現象が発生していると考えられる過去データを捨てて再学習する、もしくは新しいデータに対して大きな重みを付けて再学習するのだ。

あるいは、シフトを検知したタイミングで再学習を実施する方法もある。シフト検知は、例えば特徴量と目的変数の基礎統計量をロギングし、その標準偏差などでシフト発生を判断する。その他に、学習データと推論（に近い）データ分布について違いがないかを仮説検定する方法がある。1 変数に対しては「コルモゴロフ - スミルノフ検定」、多次元の場合は「Maximum Mean Discrepancy」（MMD）などにより仮説検定し、分布の違い、つまりシフトを検知する。

2 つ目の対処法としては、シフトによる分布の違いなどを考慮した新たな AI モデルの構築だ。例として、共変量シフトに対しての対処法を解説する。

共変量シフトが発生するときは特徴量の分布が変化しているため、「学習用」の特徴量データと、「推論用」の特徴量データを分類する AI モデルを構築できる。

そうして構築したモデルを用いた各サンプルに対して、推論用データである確率と学習用データである確率の比を算出する。その確率比を損失関数に対して重み付けて予測モデルを学習することで、「推論用データの特徴量分布」に対して推定したモデル構築ができる。

これはサンプリングバイアスの緩和策として紹介した手法と近い考えであり、実装自体も重みを変えるだけで同様に可能である。

また、増分学習（Incremental Learning）やオンライン学習（Online Learning）モデルを用いて最新データの分布に適用しながらモデルを更新していく方法も有用だ。

MMD などの仮説検定は「Torchdrift」、増分学習やオンライン学習のアルゴリズムについては scikit-learn や「river」「scikit-multiflow」などで Python 実装が公開されているため参考にしていただきたい。

今回までで、各 AI 開発プロセスにおけるバイアスリスクと、それを緩和する具体的な技術手法について解説した。AI の社会実装が進んでいく中で、今後はより倫理的側面への配慮が求められるであろう。次回は、日本を取り巻く世界の AI 倫理状況について解説する。

筆者紹介

保科 学世（ほしな がくせ）

アクセンチュア ビジネス コンサルティング本部 AI グループ日本統括 兼 アクセンチュア・イノベーション・ハブ東京共同統括 マネジング・ディレクター 理学博士。

アナリティクス、AI 部門の日本統括として、AI Hub プラットフォームや AI Powered サービスなどの各種開発を手掛けるとともに、アナリティクスや AI 技術を活用した業務改革を数多く実現。『責任ある AI』（東洋経済新報社）はじめ著書多数。

張 瀚天（ちょう かんてん）

アクセンチュア ビジネス コンサルティング本部 コンサルタント

筑波大学大学院卒業後に、アクセンチュア入社。ヘルスケアや通信など複数業界でのアナリティクス適用、AI モデル構築や運用による業務改革を支援。

世界で進む AI 規制、開発者に求められる競争力とは

正しく AI を作り、活用するために必要な「AI 倫理」について、エンジニアが知っておくべき事項を解説する本連載。第 4 回は、AI 倫理に関する世界の主要な AI 法規制やガイドラインと、AI 開発者に求められることについて。

保科 学世 , 関 大吉, アクセンチュア（2022 年 03 月 30 日）

責任ある AI 連載最終回の今回は、世界における AI 利活用に関する法規制およびガイドラインについて、先行する欧米と日本における主要なものを概説し、AI 開発者が気に留めておくべき点について述べたい。

国内外で広がるガイドラインや法規制

世界中で、AI に関する法規制やガイドラインの整備が進みつつある。近年では 2019 年に OECD（経済協力開発機構）が「The OECD AI Principles」（OECD AI 原則）を、2021 年にユネスコ（国際連合教育科学文化機関）が「The Recommendations on the Ethics of Artificial Intelligence」（AI 倫理に関する勧告）といったガイドラインを相次いで発表している。日本では 2021 年に、経済産業省から「AI 原則実践のためのガバナンス・ガイドライン」が発表された。

法的拘束力のある規制も整備されつつある。例えば米国ニューヨーク市では、独立第三機関による AI に対するバイアス（偏見）の検査やその公表がない場合、採用候補者や社員に対する人事自動決定ツールの使用を法的に禁止した。

このようにガイドラインや法規制の整備が進む背景には、AI が社会に与える影響に対する人々の課題意識の高まりが挙げられる。米アルファベットの子会社は 2017 年 10 月に、カナダのトロント市をスマートシティーとして再開発することを発表したが、住民の間でプライバシー侵害への反発が高まった結果、カナダ自由人権協会によるプライバシー侵害訴訟が発生した。

現代社会において AI は多くの人から注目されている技術であるが故に、企業は AI を適切に活用することが求められる。このような事例が発生すると企業ブランドの悪化による機会損失や収益の悪化は免れないため、AI 開発者も注意が必要だ。

一方で、このようなガイドラインや法規制は、損失を招く負の側面のみを抱えているわけではない。新たなビジネスチャンスを創出するプラスの側面も持っている。例えば、「プライバシーテック」が近年大きく注目されるようになり、プライバシー管理プラットフォームを手掛ける米ワントラスト（One Trust）は、2020 年 12 月に 3 億ドル（約 330 億円）の資金調達に成功した。

顔認証 AI と人事支援 AI への法規制が進む米国

米国ではとりわけ、人種や社会的属性に基づく差別を助長しうる AI に対し懸念が高まっている。特に顔認証 AI と採用支援 AI については、実際に法的拘束力を持った規制が敷かれつつある。例えば、中央政府機関として摘発に動いている連邦取引委員会（FTC）は、2021 年 4 月に公式サイト内で、差別的な結果への注意、アルゴリズムの透明性と独立性の確保、データ使用や入手方法の透明性、説明責任の確保などを AI 開発企業に求めている。

しかし、米国は州によって法規制が異なるため、FTC の対応に加えて各州で法案がある。例えば、イリノイ州では企業が採用面接で AI のみを用いて採用可否を判断する場合に、面接を受ける人の人口分布情報を政府に提出し、そのデータが人種バイアスを含んでいるか否かを報告する必要がある。

もし違反が認められると莫大（ばくだい）な違反金につながる恐れがある。Facebook（改め Meta）は顔認証技術の違反により 2020 年、5.5 億米ドル（約 605 億円）の和解金支払いに応じた。FTC や各州の AI 法規制には注意が必要だ。

さて、AI 開発者はこうした米国の状況から何を学ぶべきだろうか。まずは、AI 開発に用いるデータの収集や選別に今後、倫理的配慮がますます求められる点が挙げられるだろう。社会的属性や顔などの生体情報を含むデータの使用を検討するときは、それが社会的差別を助長することにならないかを吟味しなければいけない。使用を決定する際は、上記のような制裁やブランド価値の低下といったリスクをはらむことを認識する必要がある。また、この種のデータには、データそのものに社会的偏見やサンプリングバイアスが入り込んでいる可能性があり、技術的観点からも注意が必要である（データ収集時のバイアスへの対処法については、本連載第 2 回で紹介している）。

その他に、アルゴリズムの透明性や説明責任の保証が挙げられる。この点について欧州では徹底した規制が敷かれつつあるため、次項で紹介する。

欧州では個人の権利保護を重視

EU（欧州連合）もまた、AI 倫理策定で先行する経済圏である。AI 構築に不可欠であるデータに関して、EU は 2018 年 5 月に、個人データの保護と処理に関する規制である一般データ保護規則（GDPR）の施行を開始した。だが本規制は「世界で最も厳しい」と評され、各国のあらゆる業種の企業に対し、無視できない影響を与えている。2022 年 2 月時点で GDPR の制裁は計 981 件に上り、制裁金の合計は 15 億 7000 万ユーロ（約 1950 億円）を超えた。これらの制裁には、大企業や B2C（Business to Consumer）サービス業だけでなく、中小企業や B2B（Business to Business）サービス業を対象とした案件も一定数含まれており、どの業種も注意が必要といえる。

なお、GDPR の内容や制裁事例については、以下に詳しく紹介されている。

- 株式会社 IT リサーチ・アート , EU 各国における個人情報保護制度に関する調査研究報告書
- CMS, GDPR Enforcement Tracker

EU で注目したい AI 法規制がもう 1 つある。2021 年 4 月発表された包括的 AI 規制法案だ。この法案では、AI システムの開発および使用を「禁止リスク、高リスク、限定的リスク、最小限のリスク」の 4 段階で規制しており、個人の基本的権利の保護を重視していることが分かる。「禁止リスク」には、基本的な権利を「侵害する」ような AI の利用、例えば、法執行を目的とした公共の場でのリアルタイムかつ遠隔での生体認証などが含まれる。万が一禁止リスクに抵触した場合、3000 万ユーロ（約 37 億円）もしくは全世界における年間取引額の 6%のうち、いずれか高額な方の罰金が科される。

一方、「高リスク」には市民の基本的権利や健康と安全を「脅かす恐れがある」AI の利活用が含まれる。これには保険などの金融サービス保険や公的サービス、健康、公益事業、運送、人事や採用、水、電気、ガスなどのインフラなど、非常に広範なサービスが対象となる。本規制案では、AI システムが満たすべき 7 つの要件が示されているが、これらの要件は定義などに曖昧さが残り、また開発者側の負担が大きいことから、世界各国の企業から修正が求められている。

このような欧州の規制状況から AI 開発者が留意すべきことは、徹底した透明性と安全性の確保だろう。上記 7 つの要件には、AI 開発におけるデータ収集や AI 設計、AI 学習・評価、AI 運用の各プロセスにおいて、それぞれ状況を開示する要件（データガバナンス、AI 技術仕様書の作成、精度、頑健性、セキュリティの確保、ログの自動記録）が含まれている。それに加え、リスク管理システムの導入やユーザーへの透明性の確保、人間による AI の監視が求められている。日本での状況も含め、今後アルゴリズムを説明する機会は世界的に増えていくと想定される。その際、確実にアルゴリズムを説明できる人は、当の AI 開発者を置いて他にはいない。AI の解釈性を上げ、入力データが出力結果へ及ぼす影響を、分かりやすく説明できる能力が求められるだろう。AI モデルの解釈性向上については、本連載第 3 回を参照してほしい。

包括的 AI 規制法案は 2021 年 12 月に改正案が提出されており、まだ提出段階であるため今後修正される可能性もある。駐日欧州連合代表部の見解では、早ければ 2022 年中には発行や移行期間が始まり、2024 年後半には基準の整備と最初の整合性評価の実施、そして事業者への規則の適用が始まるとみている。GDPR の制裁の厳格さから判断するに、本 AI 規制案の制裁も看過できない状況となることが想定されるため、引き続き動向には注目していきたい。

規制と促進の間で揺れる日本

日本では公正取引委員会が、AI は市場に及ぼす影響が独占禁止法に違反する恐れがあるとの見解を示してきた。某グルメサイトの飲食店評価の公平性を問う訴訟では公正取引委員会が、「同サービスの一方的なアルゴリズムの改変で、特定の店舗の評価が大きく下がることなどがあれば、独占禁止法に違反する恐れもある」と、異例の意見書を裁判所に提出し、被告側はアルゴリズムを一部開示する意向を示した。今後、公正取引委員会の動向、特に独占禁止法の適用範囲について注目したい。

データの利活用については、規制だけでなく利用促進に向けた動きもある。2022 年 4 月から改正個人情報保護法が施行される予定だが、変更点として保有個人データの適用範囲の拡大や罰金の大幅引き上げ（法人に対し、最大 1 億円）などが含まれる。一方、個人情報を仮名加工して個人を特定できないようにすれば、事業者の義務が緩和されるなどの規制緩和事項も含まれる。日本では「Data Free Flow with Trust」（DFFT）を掲げていることもあり、今後データの保護規制と自由な流通の間のバランスが焦点となりそうだ。

AI 開発者として注意すべきは、AI 倫理の議論で比較的遅れている日本でも、アルゴリズムの開示事案が現れつつある点だろう。今後日本は先行する欧米の規制の流れを踏襲する可能性がある。これまでに述べた注意点を意識しつつ、データやアルゴリズムの透明性確保に努めていきたい。

先に紹介した「AI 原則実践のためのガバナンス・ガイドライン」には法的拘束力はないものの、企業が責任ある AI の実装に向けて 7 つの AI 社会原則にのっとりつつ、実際にどう行動すべきかの目標が参考事例とともに提示されている。日本語で書かれ読みやすいため、興味のある方は確認してほしい。

今回紹介した日本を取り巻く世界の AI 規制の状況は、AI 開発者の目には負の影響が目立つものに映ったかもしれない。しかし、規制があるからこそのビジネスチャンスもある。Apple は 2017 年からプライバシー保護に注目し、同意なき個人情報の追跡を不可能にする機能や、個人データの端末上での匿名化処理機能などを追加した。こうしたさまざまな取り組みが奏功し、Apple の企業価値（EV：Enterprise Value）は 2022 年までの 5 年間でおよそ 4 倍に上昇している。

今後これらの規制をチャンスとして捉え、自社の強みを生かした責任ある AI を構築することが、AI 開発者の競争力となるだろう。

筆者紹介

保科 学世（ほしな がくせ）

アクセンチュア ビジネス コンサルティング本部 AI グループ日本統括 兼 アクセンチュア・イノベーション・ハブ東京共同統括 マネジング・ディレクター 理学博士。

アナリティクス、AI 部門の日本統括として、AI Hub プラットフォームや AI Powered サービスなどの各種開発を手掛けるとともに、アナリティクスや AI 技術を活用した業務改革を数多く実現。『責任ある AI』（東洋経済新報社）はじめ著書多数。

関 大吉（せき だいきち）

アクセンチュア ビジネスコンサルティング本部 AI グループ。博士（総合学術）。

日本学術振興会特別研究員（DC1）、ケンブリッジ大学応用数学理論物理学部客員研究員などを経て、現職。アクセンチュア AI センターでリサーチ部門を担当し、京都大学との社会課題解決に向けた AI 開発協業や AI 倫理の取り組みなどを手掛ける。

編集：@IT 編集部
発行：アイティメディア株式会社