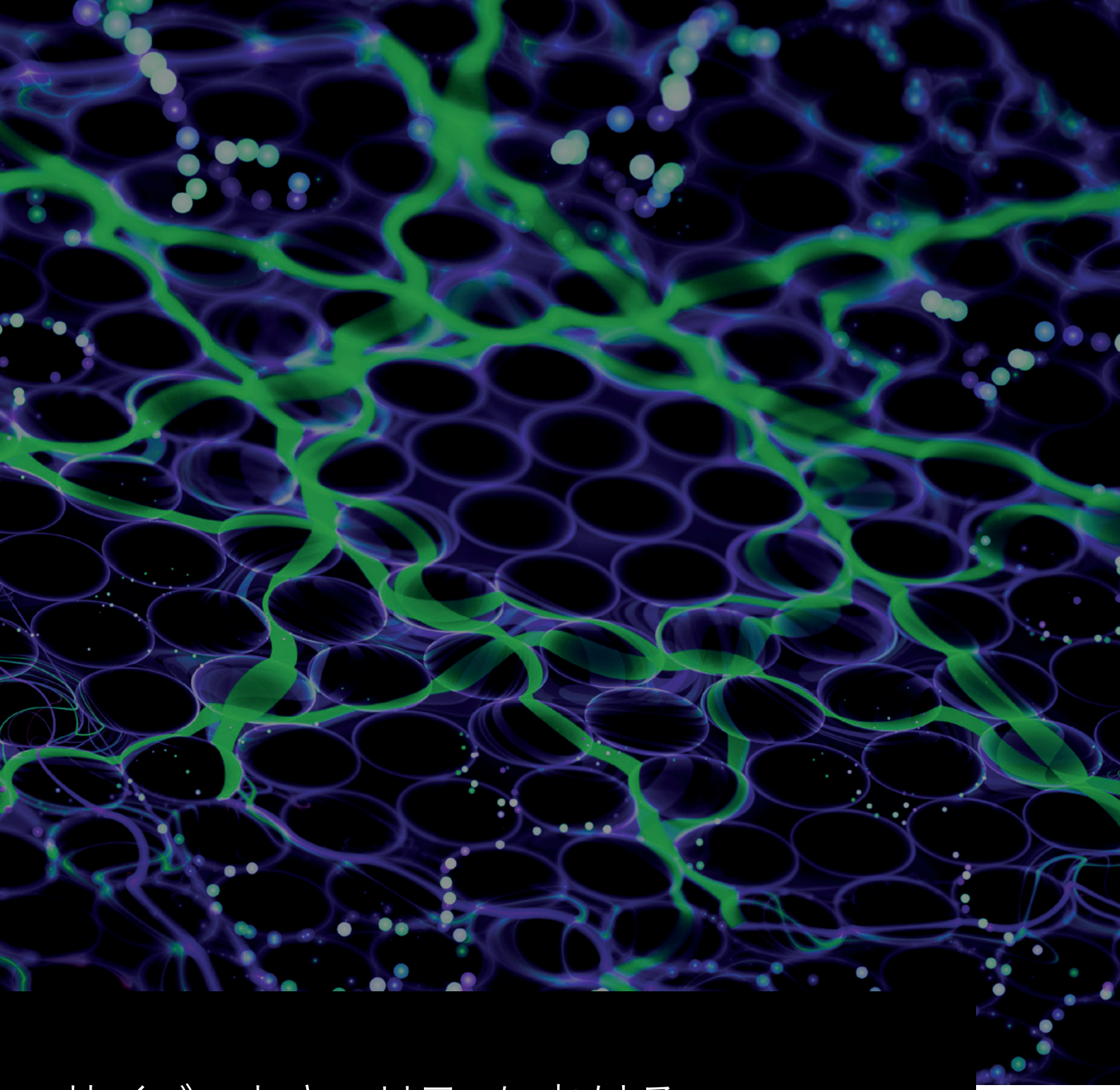




サイバーセキュリティにおける 機械学習の世代別解説（第 1- 第 5 世代）

AI/ 機械学習の真の活用とは？



CYLANCETM

概要

このホワイトペーパーの目的は、機械学習の各世代を定義し、今日のサイバーセキュリティに应用されている人工知能（AI）と機械学習（ML）の成熟度について説明することです。さらにこのペーパーでは、サイバーセキュリティ問題に対する機械学習の応用に大きな進歩があった一方で、この分野には限りなく大きなイノベーションと発展の機会がまだ残っていること、そして機械学習の応用の高度化が長期的に進展し続けると予想されることを明らかにしていきます。

このホワイトペーパーは、以下のセクションに分かれています。

- サイバーセキュリティの中で機械学習の活用状況、サイバーセキュリティ機械学習モデルを世代別に分類
- 機械学習の世代を区別する際の主要な概念と考慮事項
- サイバーセキュリティ機械学習の5段階の世代に関する定義
- 機械学習の世代という枠組みが持つ意味
- まとめ

はじめに

Defense Advanced Research Projects Agency（DARPA）は、AIを3つの基本的な種類に分けて定義し、これらを [AIの3つの波](#)と呼んでいます。

第1の波は**手作りの知識**で、人間が特定の機能を実行するために使う規則を定義します。また、コンピューターはこの知識から学習してこれらの規則を自動的に適用し、論理的な推論を行うことができます。ただし、この第1の波では、より高いレベルの学習には適用できません。この第1の波に分類されるサイバーセキュリティの例は、[DARPA Cyber Grand Challenge](#)です。

第2の波は、**静的学習**です。この波に分類されるAIは、自動運転車、スマートフォン、顔認識などによく使われ、機械学習を使用して、すべきことやしてはならないことについて確率的意思決定を行います。この第2の波では、システムは学習能力に優れていますが、論理的な推論を行う能力に弱点があります。つまり、システムはデータを分類して予測するのですが、文脈を理解しません。

ここで**文脈適応**と呼ばれる第3の波の登場です。この波では、システムが実世界そのものの説明モデルを独自に構築します。第3の波では、システムは人間とまったく同じように、特徴付けが行われた理由を正確に説明できなければなりません。

機械学習は、攻撃の検出と防御を自動化する能力があることから、サイバーセキュリティにいち早く導入されてきました。特に、次世代アンチウイルス（NGAV）製品では一般的です。NGAVのMLモデルには、従来のAVと比較して基本的な利点があります。たとえば、新手法のゼロデイ攻撃やターゲットを絞ったマルウェアを特定できる可能性が高いこと、マルウェアの検出回避がより困難であること、長いオフライン期間中に効果が続くことなどです。

現在サイバーセキュリティにMLとAIを適用する試みのほとんどは、DARPAの第1の波に分類されます。つまり、人為的に定義された規則と、定義されたパターンを使用する、手作りの知識です。DARPAの第2の波である静的学習を取り入れていると言えるサイバーセキュリティテクノロジーはごくわずかです。その成熟度はこ

く低いというのが現状です。

市場で大多数を占める第1の波のMLモデルは、検知漏れが高くなることは避けられず、容易に回避されます。現在では、AVに対するMLの応用が進んでおり、AVの現行バージョンと次のリリースとの差分を語るだけでは説明は不十分です。近い将来、サイバーセキュリティに应用されるMLの進化レベルについて、網羅的に理解すべき時期が来たと言えるでしょう。

このペーパーでは、DARPAの第2の波、静的学習の中でサブカテゴリに分けた機械学習の世代について解説します。今日のサイバーセキュリティでの応用例に見られるAIの成熟度、および長期的にどのような進化が見込まれるかについて説明していきます。

概念と考慮事項

このセクションでは、本書で使用される、MLモデルの世代間の区別をつけるために役立つ用語と概念について説明し、さらにこれらの概念のセキュリティとの関連性を解説します。

ランタイム

機械学習アルゴリズムは、一般に以下の2つの基本的なステップから構成されます。

- トレーニング**：ここでは、モデルは既知のサンプルのデータセットから学習します。
- 予測**：ここでは、トレーニングされたモデルが新しい未知のサンプルに関して推論判定を行います。

トレーニングステップでは、非常に集中的な計算操作が行われます。現代のディープニューラルネットワークは、ハイパフォーマンスなクラウドサーバーの大規模クラスター上でも、トレーニングに数カ月かかることがあります。モデルのトレーニングが行われた後の予測は比較的簡単な処理ですが、やはり予測でもメモリとCPUを大量に必要とすることがよくあります。分類子のトレーニングを行うには、入力データセットからのサンプルにラベルが関連付けられている必要があります（たとえば、悪意があるものか、悪意のないものか等）。

ランタイムは、トレーニングや予測が行われる環境、つまりローカル（例：エンドポイント上）、またはリモート（例：クラウド内）になります。それぞれの機械学習のランタイムは、新しいサンプルによってモデルを更新できる速度、意思決定の影響、およびCPU、メモリ、IOなどのリソースにも依存します。教師ありのモデルの場合は、ラベルが使用可能でなければトレーニングを実施できないので、トレーニング時にラベルが用意されている必要があります。現状では、トレーニングはクラウド内の分散サーバーのクラスターで行われるのが一般的です。予測もクラウドで行われるのが一般的ですが、ローカルで実行される例も増えています。ローカルユーザーまたは顧客のデバイス上での分散トレーニングは、新しいテクノロジーです。IOの削減や機密データの保護など、大きな利点が得られる可能性はありますが、異種混合のリソース、不安定な可用性、低速な実験反復など、多くの課題もあります。

特徴量

一連の特徴すなわち特徴量空間は、それぞれのデータのどの特性がモデルに取り込まれているかを正確に指定します。実行可能形式（PE）ファイルの場合、特徴セットにはファイルサイズやエントロピーなどの基本的な統計データ、および PE の解析されたセクション（たとえば、セクションテーブル内の各エントリの名前）に基づく特徴などがまず考えられます。別の派生的な特徴としては、例えばファイルサイズを底とする2の対数も挙げられるでしょう。特徴にはさまざまなものが考えられ、ほかの特徴に基づいて条件的に抽出される特徴や、組み合わせを表現する特徴などがあります。可能な特徴量空間は膨大です。これは、特徴に適用できる変換が無数にあることを考えれば当然のことです。

特徴は、学習データがどのように抽出されるかを決定するので、ML モデルには非常に重要な意味を持ちます。どの情報を含めるかという重要な問題のほかに、情報をエンコードする方法も大事です。モデルに合った特徴を作成するプロセスは、特徴量作成と呼ばれます。モデルによっては、特徴のエンコード方法に依存する度合いが他より高いものがあります。可能な限り多くの特徴を用意した方が良くと考えがちですが、使用する特徴が多すぎることには欠点があります。オーバーフィッティング（過学習）が生じるリスクが高くなったり、学習時のリソース消費が増大するほか、敵対的な攻撃に対する脆弱性が増すこともあります。モデルの能力、解釈可能性、頑健性はすべて、特徴によって決まります。

データセット

モデルのトレーニングと評価に使用されるデータは、モデルの性能にきわめて大きな影響を及ぼします。モデルのトレーニングに使用されるデータが実世界を表現していなければ、モデルは本番環境でうまく動作しません。それぞれのサンプルには、トレーニング用の分類子として、悪意がないかあるかといったラベルが必要になります。ラベルは慎重に吟味する必要があります。ラベルの間違ひはラベルノイズとも呼ばれ、モデルの性質を偏らせる可能性があるからです。データの収集を進めながら、一貫性と正確性を維持するために、ラベル付きデータセットのモニタリングを続ける必要があります。実際上、データはさまざまな種類のソースから収集されることが考えられます。それぞれのソースの信用度や信頼性はきちんと評価し、それを考慮に入れた上で活用すべきです。

多くのセキュリティアプリケーションに共通する課題としては、アンバランスなデータをどう処理するかということです。この状況は、1つのラベル（悪意がない）が他のラベル（悪意がある）より突出して多く存在するときに生じます。アンバランスなラベルが付いたデータの問題は、さまざまなモデリング戦略によって軽減はできますが、それぞれのラベルを表現する多数のサンプルが存在するのが一番理想的です。多くの特徴はトレーニングセットを使用して生成されるので、特徴セットとデータセットは密接に関連しています。データセットは、正規化などの重要な前処理や、単語出現頻度 - 逆文書頻度 (TF-IDF) などの重み付け方式にも影響を与えます。

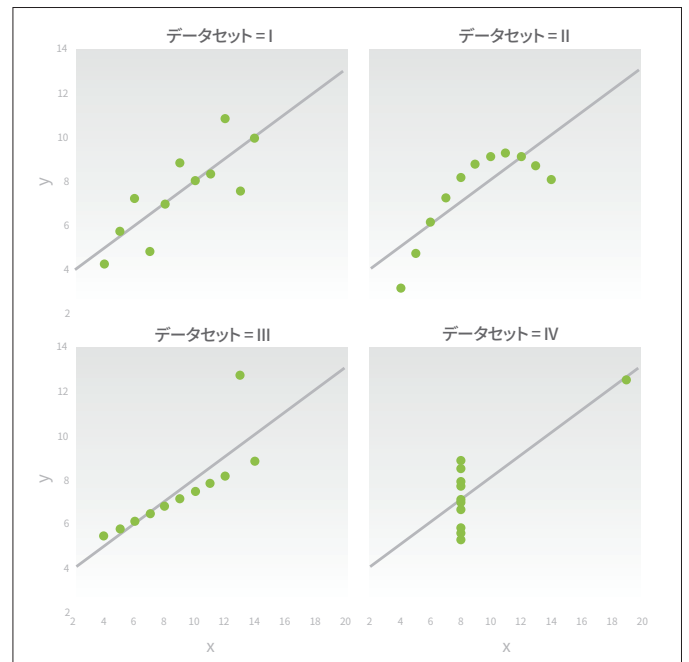


図 1—4 つの大きく異なるデータセット（点で示されている）から、同じ適合予測モデル（それぞれの線で表される）が得られる。この有名な一連のデータセットは、アンスコムスの例と呼ばれる。

高度なモデルでは、非常に大規模なデータセットが必要です。一方で、十分に大きなデータセットがあれば性能が良くなる、と思ひ込むのは危険です。一般に、データセットが大きければ、より高度なモデルのトレーニングが可能になりますが、巨大なデータセットがあれば性能が無条件に良くなるという訳ではありません。適切なデータセットには幅広い多様性が必要で、モデルが実装されたときに会おうと考えられる実サンプルを十分に表現しているべきです。望ましい多様性は、ラベル付きのサンプルデータにおける特徴量のバランスとして定量的に表すことができます。

人間との対話

モデルはブラックボックスと考えられることがよくありますが、必ずしもそうではありません。人間との対話のモードをサポートできるモデルには、いくつかの利点があります。モデルはエキスパートからのフィードバックをただちに受け取ることができ、このことはラベルと特徴の両方の改良に役立つため、他の方法では難しいモデルの改良が可能になります。モデルの意思決定がどのように行われているか理解する手段があれば、モデルに対する人間の信頼や信用をより早く確立できます。

モデルを改良する方法があることは、基礎のデータを検証するためにも役立ちます。図 1 はアンスコムスの例を示しています。この例では、4 つの大きく異なる入力データセットから、まったく同じ線形回帰モデルが得られます。要約統計量とモデルパラメータを根拠とすれば、4 つのケースは実質的に区別ができません。プロットすれば、左上の象限のモデルのみがデータセットに適合していることがただちに明らかになります。他のモデルでは、次元とパラメータがさらに多いので、検証したり理解することがはるかに困難です。しかし、人間による何らかの検証がなければ、モデルの品質やデータの問題が見過ごされ、効果の低下や脆弱性が生じる可能性があります。

モデルが間違いを犯した場合にも、人間との対話のモードをサポートしていることは重要です。モデルがブラックボックスならば、体系的なモデリングにエラーが生じた原因を特定することは

難しくなるでしょう。モデルを検査して理解するためのツールを使用すれば、トラブルシューティングと診断が可能になります。こうしたツールは慎重に管理する必要があり、最終的な製品に組み込んではいけません。これらのツールは知的財産の漏洩を起こし、攻撃者に対して脆弱性を露呈する可能性があるからです。

適合度

モデルによって、実世界をよく表現しているものと、そうでないものがあります。モデルが簡素化されすぎていると、効果は低くなりますが、新しいデータへの汎化能力は高くなります。このようなモデルは「アンダーフィット（適合不足）」と呼ばれます。つまり、モデルに提供される情報の量がモデルの適合能力を超えていて、情報がモデルに完全に取り入れられないという状態です。逆に、モデルが暗記をする、つまり「オーバーフィット（過剰適合）」になることがあります。過剰適合の場合、モデルはトレーニング対象である特定のサンプルについて学習しすぎて、実世界の新しいサンプルにうまく適応することができなくなります。

図2では、緑とグレーの点に対する、過剰適合の分類子の決定境界を点線で表しています。緑の線は、適切な決定境界を表します。示された点の分類は完璧ではないものの、新しい点に対する性能は良くなります。

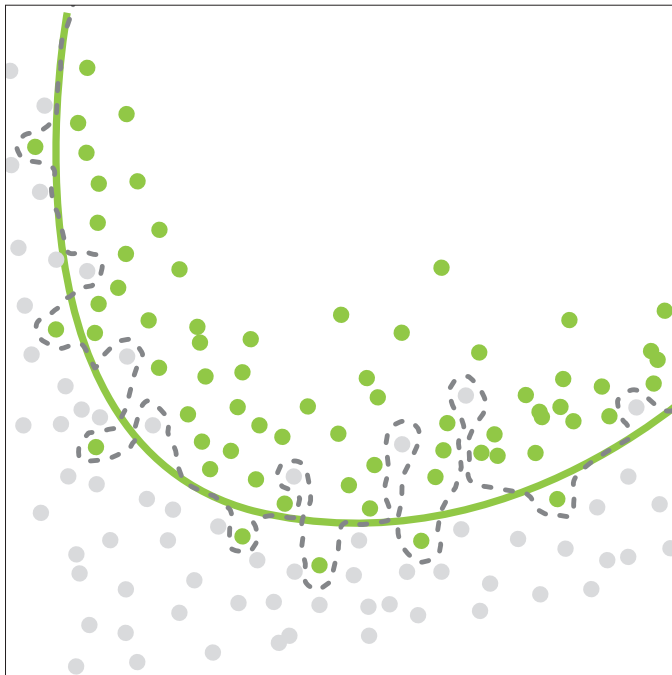


図2—2つのクラスからのデータ点。それぞれのクラスは色分けして示されている。2本の線は、仮定した分類子から得られる2種類の決定境界を示している。

適合性の良いモデルは、製品に実装された後もその性能を維持します。関連した概念の1つ、コンセプトドリフトは、データに長期的な非定常の変化があるときに生じます。たとえば、エンドポイント上でPEファイルのセットは毎年変化していきます。サンプルPEファイルの母集団が変化するので、モデルはターゲットとする母集団の変化に適応できるように備えておく必要があります。

世代はどのように定義されるか

サイバーセキュリティの機械学習の世代は、5つの主要要素に従って互いに区別されます。これらは、データサイエンスとサイバーセキュリティの両プラットフォームの共通部分を反映しています。

- **ランタイム**：MLのトレーニングと予測はどこで行われるか（たとえば、クラウド内、またはローカルエンドポイント）。
- **特徴**：いくつの特徴が生成されているか。特徴の前処理と評価はどのように行われるか。
- **データセット**：どれだけのデータがどのようなプロセスで選別処理されているか。ラベルはどのように生成、提供、検証されているか。
- **人間との対話**：人間はモデルの意思決定をどのように理解し、フィードバックを提供するか。モデルの継続的なモニタリングはどのように行われるか。
- **適合度**：モデルはデータセットをどれだけよく反映しているか。どれほどの頻度で更新する必要があるか。

これらの要素によって、サイバーセキュリティのテクノロジーを機械学習における5段階の世代に分けることができます。それぞれの世代は、各カテゴリでの進歩の割合によって定義されます。通常、製品化されたモデルが次の世代に進むまでに2～3年かかり、機械学習を組み込んだテクノロジーの大部分はまだ第1または第2世代から先には進んでいません。第3世代に到達している製品はまだかなり少数で、その進化は市場で多くの検証や評価を経てしか獲得されないものになります。第4世代、第5世代への進化にはさらに膨大な研究と開発が必要です。サイバーセキュリティ分野への応用に必要とされる条件は、特に敵対的学習、能動学習、フェデレーションラーニング、モデルの解釈可能性の領域であり、機械学習研究の最先端に関わってきます。

以下の表は、前述の各要素での達成度に応じた各世代別の特性を要約したものです。

世代	ランタイム トレーニングと予測 が行われる場所	特徴 特性、要素の数	データセット サイズとラベルの 提供元	対話 人間との対話	適合度 モデルが実世界に どれだけよく適合 するか
第 1 世代	- クラウドでの トレーニング - クラウドでの 予測	1,000 を超える 特徴	100 万以上 - データサンプル 人間による ラベル付け	人間が意思決定 を理解する	過小適合、高い 誤検知率
第 2 世代	- 第 1 世代 - ローカル予測	10 万を超える 特徴	1 億以上 - データサンプル 人間による ラベル付け、 一部はヒュー リスティックに よるラベル付け	モデルが意思 決定を説明する ことは困難	過剰適合、高い 誤検知率
第 3 世代	- 第 2 世代 - クラウド拡張 モデル	100 ～ 300 万の 特徴	10 億以上 - データサンプル 大部分がヒュー リスティックに よるラベル付け	モデルは理解 可能な説明を 提供する	正しい適合、 精度メトリック による汎化
第 4 世代	- 第 3 世代 - ローカル トレーニング	300 万を超える 特徴	オンライン学習	モデルは戦略を 説明し、 ハイレベルな フィードバック を受け取る	モデルは現在の 入力にも将来の 入力にも適合
第 5 世代	- 第 4 世代 - 教師なし ローカル トレーニング	無制限、半教師 あり学習で発見	能動学習	人間による入力 はオプション、 解釈可能な知見	モデルが コンセプト ドリフトを 認識し、適応

機械学習の各世代の詳細な意味

上記の表は、それぞれの世代を区別する特性を示しています。それぞれの世代は、前の世代を基礎としています。データセットのサイズと特徴の数は、世代ごとに大幅に増加します。以下では、それぞれの世代について、前の世代との違いを中心に説明します。

- 第 1 世代の機械学習：**標準モデルを使用した、[scikit-learn](#) のような既製の ML ツールキットの応用です。善／悪のラベルはすべて人間の分析によって提供されます。つまり、特徴セットは小規模で、固定されており、人間によって選ばれたものです。これらのモデルはエンドポイントに実装できません。概して誤検知の発生する度合いが高く、効果は非常に限られます。また、回避も容易です。
- 第 2 世代の機械学習：**ほとんどのラベルはやはり手動で適用されますが、この段階ではヒューリスティックが人間によるラベルの補完に使用されます。この応用によりローカルモデルでの予測が可能になりますが、クラウドベースのトレーニングはまだ必要です。ローカルモデルは、クラウドモデルのクローンです。解釈は人間が記述子を指定する方式で行われます。これは事後

のもので、モデルの意思決定プロセスに真の意味では関与していません。この段階のモデルは、トレーニングデータに対して過剰適合になるのが通例です。このクラスのモデルには多少の予測能力がありますが、コンセプトドリフトの発生を防ぐために、依然として定期的な更新が必要です。

- 第 3 世代の機械学習：**クラウドモデルはより高度になり、ローカルモデルを補完、保護します。意思決定の説明が、意思決定プロセスを反映した方法でモデルによって行われます。モデルは評価され、攻撃に対抗して強化されるように設計されています。コンセプトドリフトは優れた汎化能力によって軽減されます。
- 第 4 世代の機械学習：**モデルはローカルデータから学習し、所見をアップロードする必要はありません。特徴は人間とモデルの間の戦略的な対話によって設計されます。新しい特徴とモデルは、継続的な実験によって常に評価されます。モデルの修正とガイドのために、人間からフィードバックを提供できます。ほとんどは、既知の攻撃に対して堅固です。
- 第 5 世代の機械学習：**教師あり学習はオプションになります。モデルは分散した半教師あり学習で生成されます。モデルが

提供する知見が、人間による分析のガイドになります。継続的なモニターと監査によってモデルの改ざんを見つけることができ、敵対的学習による攻撃を検出するための偽装機能をサポートします。

結論

サイバーセキュリティへの機械学習の応用は、2013 年頃から始まったばかりでこの業界ではまだまだ新しい取り組みです。このきわめて困難なサイバー攻撃からの保護に向けた長い道のりの中で、機械学習を応用する試みは始まったばかりです。機械学習が進化を遂げるたびに、将来の世代が5つの成熟度の領域（ランタイム、特徴、データセット、人間との対話、適合性）で提供する能力はますます高まっていくことになるでしょう。

サイバーセキュリティにおける機械学習にまつわる誇大広告の例として、レガシーなサイバーセキュリティベンダーが利用するきわめてシンプルな機械学習があげられます。

機械学習はある程度の適用性が幅広い分野で実証されてきましたが、サイバーセキュリティへの適応はまだ比較的新しい取り組みです。サイバーセキュリティの重要性により、サイバーセキュリティの未解決問題を対象とする新たな研究が期待されており、既存のデータセットで単純なモデルをトレーニングするだけでは不十分です。

それぞれの世代は、前の世代に対する質的な向上を実現します。成熟度は、機械学習が提供する価値に直接的な影響を及ぼします。これらの変化は効果のわずかな改善にとどまらず、機械学習がサイバー攻撃を検出して防止する基礎能力の飛躍的な向上を意味するからです。機械学習によるアプローチに価値があることは早々に証明されましたが、より高度な機械学習の世代によってはじめて、本当のセキュリティが実現されることでしょう。

