［機械学習工学］

4 機械学習応用システムのセキュリティとプライバシ

吉岡信和 | 国立情報学研究所

セキュリティの重要性

機械学習は，医療や自動運転など人命や社会インフラに直結するシステムに組み込まれてきている．そのため，機械学習の判断を意図的に変更し，誤判断を起こさせることにより，社会，組織，個人に被害が及ぶ可能性が高まっている．たとえば，機械学習を使って自動で標識を認識する自動運転車を考えてみよう．道路標識をスプレー等で人が気づかないくらい軽微に書き換えて誤認識させることができれば，意図した事故や渋滞を引き起こすことができてしまい，人命の被害や社会的な混乱につながってしまう．

機械学習では，データから振舞いを決定し，精度の良い訓練済みモデルを得るためには，大量のデータを必要とする．さらに，そのデータは，しばしば不特定多数で共有される画像を用いたり，公道にある看板の情報など，誰でもアクセスできるデータが用いられる．そのため，比較的容易に訓練や推論のためのデータを書き換えることが可能となる．すなわち機械学習応用システムでは，従来のように直接アルゴリズムをプログラミングする場合よりデータの管理が重要になり，データの管理を怠るとセキュリティのリスクが高まる可能性がある．

機械学習の判断を意図的に変更する入力データの書き換えとして，敵対的サンプル（Adversarial Example）が知られている．これは，もとの入力データに対して，人の目にはノイズとしか認識できないような軽微な変更を施すことにより，訓練済みモデルの判断を変更する方法である．

本稿では，敵対的サンプルを始めとする機械学習応用システムに関するセキュリティの概要を述べる．

機械学習応用システムの特徴とセキュリティ脅威

機械学習応用システムは，データから振舞いを自動で生成するため，その特徴に起因する以下のような新たなセキュリティの脅威が懸念されている．

訓練や推論に使われるデータを変更することによりセキュリティの被害につながる意図的な誤動作を比較的容易に起こすことができる．特に，公道の標識や不特定多数が生成した画像データなど，誰でも入手できるデータを使う場合，データの信頼性の担保が難しく，データへの攻撃（書き換え）が容易になる．

訓練データに個人情報や企業の機密情報が含まれる場合，訓練済みモデルから訓練に使われたデータを推測することができれば，プライバシの侵害や機密情報の漏洩につながる．また，訓練データが，特定の組織から提供されているデータだと判明すれば企業の機密情報が漏れる可能性も出てくる．

たとえば，顔認識により特定のサービスを利用するようなシステムがあった場合，訓練に使われた顔画像が推測できれば，そのサービスの登録者リストを得ることができてしまう．

精度の良い訓練済みモデルを得るためには，大量のデータを必要とするため，データの一部が書き換わっ

ても，その発見が難しい．そのため，訓練データの一部を書き換えることにより，特定の入力の際に誤判断を起こす訓練済みモデルを作ることが可能となる．たとえば，不良品を診断する機械学習を行う際，特定の不良を故意に見落とすように，訓練データやテスト用のデータを書き換えられる可能性がある．

確率的に最適な振舞いを決定するため，すべての入力データにおいて100％確実な判断を保証することができない．そのため，信頼度が低い判断が必ず生じ，入力データを少し変更しただけで，その判断を変更することが容易になる．このため，誤判断を起こす入力の可能性を0％にできず，どのような訓練済みモデルでも必ず脆弱性が残る．

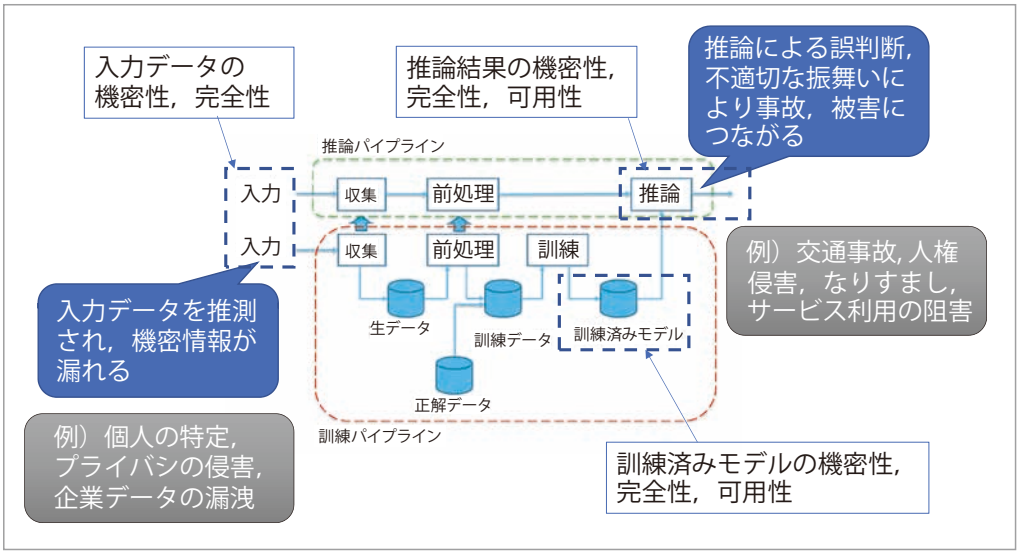
このように，機械学習応用システムでは，従来よりも多くのデータを扱い，データにより振舞いが決定されるため，それらのデータに関してセキュリティを考慮する必要がある．具体的には，図-1にあるとおり，訓練や推論に用いる入力データの機密性や完全性，そして推論結果の機密性，完全性，可用性を考える必要がある．さらに，訓練済みモデルに含まれる構成やパラメータを保護資産と捉えるならば，訓練済みモデルの機密性，完全性，可用性を考える必要がある．

表-1に機械学習応用システムのアプリケーション事例と，それに対するセキュリティの脅威の例を示す．

さらに，訓練データに意図的な操作をすることにより，人権を侵害したり，特定の組織の評判を落とす訓練済みモデルを構成することができる．これによりサービスの停止に追い込まれるならば，サービスの可用性が脅かされるため，広い意味でのセキュリティリスクと捉えることができる．その詳細を次章で説明する．

■表-1　機械学習応用システムのアプリケーションと脅威の例

アプリケーション事例	脅威例
自動運転車	故意による交通事故（物理セキュリティ）
チャットボット	不適切な発言によるサービス停止，人権侵害などの悪評判
製品・サービスの異常や故障の検知	異常・故障の見逃し
パーソナルアシスタント	誤認識によるなりすましと個人情報の流出
顔認証を使ったサービス	誤認識によりサービス利用を阻害する．なりすましによるサービス利用
メールフィルタリング	フィルタリングの誤判断によりメールの利便性が落ちる



■図-1
機械学習応用システムに関するセキュリティ（機密性，完全性，可用性）の考慮

プライバシや人権の侵害，悪評価の脅威

機械学習応用システムを悪意を持って変更，分析することにより，個人情報を特定するなどのプライバシ侵害や，人権を脅かす，もしくはサービスの評価が著しく悪くなり，サービスを停止しなくてはならない可能性がある．

機械学習アルゴリズムに入力する訓練データや予測データには，個人に関する情報が含まれることがあり，機械学習応用システムの出力からその入力を再現することができれば，プライバシを侵害する恐れがある．たとえば，病気を予測する機械学習システムは，訓練データとして実際の個人の病気の情報を使う必要があるが，誰の情報を使ったのかが予測できてしまうと情報提供者のプライバシを侵害してしまう．

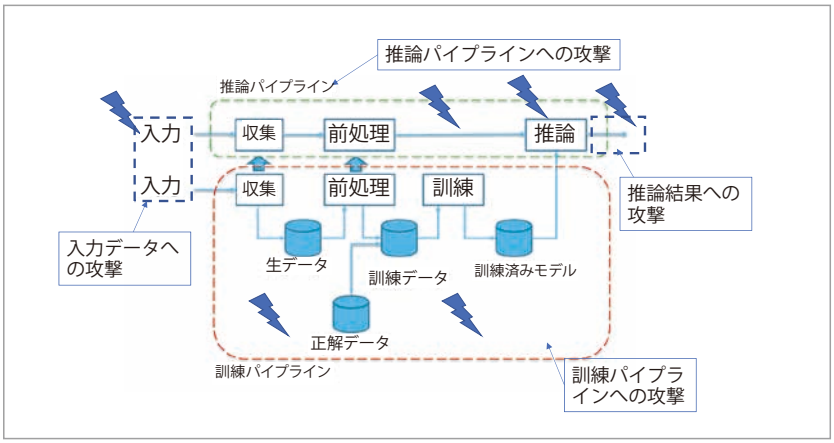
機械学習応用システムが，ユーザの人権を脅かす事態も実際に起こっている．具体的には，機械学習を使ってユーザから会話を学習するチャットボットが，公開後すぐに閉鎖に追い込まれた．ユーザが差別的な会話をチャットボットに教え込ませてしまい，ボットが差別的なジョークを発言するようになってしまったからである．そのような差別的なジョークを読んだユーザは，自分の人権を侵されたと感じるだけではなく，サービス提供者に発言に対する責任を追求する可能性もある．チャットボットに差別的な会話を教え込ませたユーザは軽い気持ちだったかもしれないが，それによりサービスの評判が悪くなっただけではなく，サービス提供者は，サービスの提供を中断せざるを得なくなってしまった．

機械学習応用システムの訓練データとしてユーザからの情報や公開情報を用いる際には，推論結果が不適切にならないようにシステムを構築する必要がある．たとえば，不適切な情報や不適切な判断につながる情報が含まれるかどうかを前処理の段階で確認し，そのような情報は訓練に用いないなどである．また，差別などの偏見がないように訓練させる方法が，機械学習における公平性として研究されている．

機械学習応用システムへの攻撃

機械学習モジュールは訓練パイプラインと推論パイプラインからなる．そのため，セキュリティの攻撃も訓練パイプラインに関連するデータへの攻撃と推論パイプラインに関するデータへの攻撃に分けられる．後者の攻撃には先に述べた誤判断を引き起こす敵対的サンプルのほか，スパムメールなど特定のデータであることを認識できなくする回避攻撃（evasion attack）がある．

図 -2 に機械学習応用システムに対する攻撃の可能性を示す．



■図 -2
機械学習応用システムへの攻撃は訓練パイプラインに対するものと推論パイプラインに対するものに分類される

機械学習応用システムに対する攻撃には，以下の２種類の可能性がある．

- 計算機上への攻撃：計算機上にあるデータ，通信，訓練済みモデルに対して，その書き換え，盗聴等を行う．特に，入力データを推論エンジンまでに送る途中経路に，不特定多数がアクセスできる通信がある場合，その通信路で攻撃される可能性がある．
- 物理的な攻撃：道路標識や物理的物体の画像，音声など，入力データとして物理的なデータを使う場合，看板の書き換えや落書きなど，物体そのものの書き換えを行うことができる．また，推論の高速化のためクライアント側に訓練済みモデルを置く場合，物理的な攻撃により訓練済みモデルの情報を盗まれる可能性がある．

さらに，攻撃者の知り得る情報によって，ブラックボックス攻撃とホワイトボックス攻撃の２種類に分けられる．ブラックボックス攻撃は，訓練データの情報，機械学習アルゴリズムや訓練済みモデルの詳細など，訓練パイプラインの情報を一切分からない場合の攻撃であり，ホワイトボックス攻撃は，これらの情報を利用した攻撃である．さらに，ホワイトボックス攻撃は訓練パイプラインにかかわるすべての情報を知っていると仮定した攻撃と，訓練済みモデルのアルゴリズムの種類だけ知っている場合の攻撃など特定の情報だけ利用する攻撃に分けられる．たとえば，ニューラル・ネットワークのアルゴリズムに特化した攻撃は，訓練済みモデルのアルゴリズムやそのネットワーク構造を知っている必要があり，ホワイトボックス攻撃となる．

機械学習応用システムの脆弱性

機械学習への脆弱性の１つとして，入力データを少しだけ変更することで予測や推論結果を変更する敵対的サンプルが知られている．ここでは，そのような脆弱性の概要を紹介する．

たとえば，図 -3 は道路標識に対して人の目にはスプレーやテープを使った落書きにしか見えない変更を施すことにより，機械学習応用システムが誤認識してしまう例である．このように人は判断できる看板を，機械学習応用システムに誤認識させることにより自動運転車に対して意図的に事故を起こさせることができる．

そのような敵対的サンプルを作成する方法として，訓練データや学習に使ったアルゴリズムの情報を用いて行うホワイトボックス攻撃が知られている．その方法は，もし，特定の目標に誤認識させたい場合，その目標に向かって損失を最小にする（その目標と判断させる可能性が高くなるようにする）と同時に，もとの入力データとの違いが最小になるような書き換え（ノイズ）を探し出すことになる．たとえば，“8” と書いてある文字を “9” に誤判断させたい場合，“9” と判断したときの損失を最小にするノイズを見つけることとなる．

道路標識の停止標識を認識させなくするだけなら，高速に敵対的サンプルを探す方法が提案されている．この場合，誤認識させたい目標と判断した場合の損失を増加させるような最小のノイズを見つけることになる．たとえば，停止標識の入力に対して，停止標識と判断した場合の損失を増加させるノイズを見つけるなどである．

さらに，訓練データや機械学習に用いたアルゴリズムについての知識を用いないブラックボックス攻

■図 -3　攻撃者が停止標識にスプレーやテープを貼ることにより，標識を認識させなくすることができる（文献 1）から引用）

撃も提案されている．これは，攻撃する機械学習のアルゴリズムとは無関係に，特定のアルゴリズムで生成した敵対的サンプルを用いる方法である．こういった攻撃が有効なのは，1つのモデルで見つけた攻撃は，他のモデルでも有効である（攻撃の再利用性がある）ためである．

入力データを書き換える攻撃のほか，訓練データのほうを軽微に書き換えることで訓練済みモデルに攻撃者にとって都合の良い脆弱性を埋め込む攻撃（中毒攻撃）も知られている．これは，ある特定の条件下で発生するバグをプログラムに埋め込む方法に似ており，たとえば，特定条件下で，攻撃者の侵入を発見させなくする（侵入していないと誤認識させる）攻撃である．

機械学習の他の脆弱性として，訓練データを推測できてしまう場合がある．図 -4 が顔認識の訓練済みモデルから訓練に使われた顔写真を推測する例である．図の左が推測した画像で，右が訓練に使われた画像である．このように訓練データの一部を推測できてしまう場合，機密情報にしたい特定サービスに参加しているメンバの情報などが漏洩してしまう．

機械学習応用システムのための セキュリティ対策

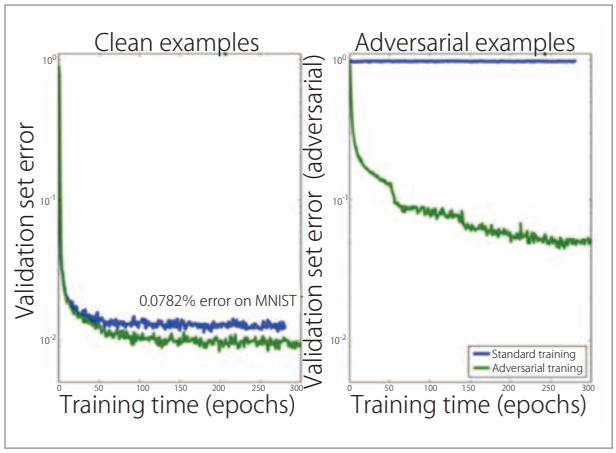
機械学習応用システムのためのセキュリティ対策としては，(1) 訓練の方法を改良することにより訓練済みモデルの品質を向上させる方法，(2) 暗号化された訓練データから訓練済みモデルを生成する方法，(3) システムレベルの対策などが考えられる．

訓練を改良する方法として，敵対的サンプルを用いた訓練（Adversarial Training）が提案されている．この方法で訓練した場合，図 -5 のように敵対的サンプルが含まれない入力データに対しても精度が若干向上するという効果も含まれている．図の左のグラフは，通常の入力に対する訓練済みモデルの精度を，敵対的サンプルを用いずに訓練した場合と敵対的サンプルを用いて訓練した場合を比較している．どちらの場合も精度は高いが，敵対的サンプルを用いて訓練した場合のほうが若干精度が高くなっているのが興味深い．図の右のグラフは，敵対的サンプルを訓練データに加えた場合の精度の比較である．敵対的サンプルを用いて訓練したモデルは，敵対的サンプルを入力した場合でも精度の低下が抑えられているのが分かる．

しかしながら，敵対的サンプルを用いた訓練を行ったとしても，敵対的サンプルによる誤認識のリスクは残る．この誤認識率をシステムの仕様として受け入れられない場合，システムレベルのリスク軽減が必要になるであろう．

また，データが多少異なっただけで，判断結果が大きく変わらないような訓練済みモデルを生成する

■図 -4　訓練済みモデルから訓練に使われた写真を推測した例：左が推測した画像で，右が訓練に使った画像（文献 2）から引用）



■図 -5　敵対的データを訓練に使うことで，学習の精度を上げることができる（文献 3）から引用）

方法（Distillation）が提案されている．このような訓練済みモデルは，頑健なモデルと呼ばれ，学習の信頼性を測定する基準の１つにもなっている．また，暗号化された訓練データを用いて訓練することにより，訓練データの情報を秘匿にすることができる．これにより，機密情報や個人情報を使った訓練でも，その情報が漏れるリスクを減らすことができる．このための学習アルゴリズムには，データを暗号化したまま加算・乗算が可能な準同型暗号が利用される．

さらに，システムレベルのセキュリティ対策として，入力データ等が攻撃により書き換えられないような対策のほか，機械学習に使うアルゴリズムや構造，出力に付随した確信度などの情報を不用意にユーザに公開しないなどの対策が考えられる．特に，学習アルゴリズムに関する情報や出力の信頼度は，機械学習のパラメータやアルゴリズム等の情報を利用した攻撃（ホワイトボックス攻撃）に利用されることにつながるため，システムの提供に必要な情報以外は公開しないほうがよい．

敵対的サンプルを用いた攻撃の範囲については，研究段階であるが，画像のスケールや解像度が変わると攻撃の成功率が著しく下がることも報告されている．そのため，推論に使う入力データの画像に対して，目標物を切り取ったり，解像度を揃えたりせず，スケールや解像度が異なる複数のデータを入力に使うなどのシステムレベルの対応も考えられる．

今後の展望

敵対的サンプルに関する研究は，まだ始まったばかりであり，どのような原理で脆弱性が発生するのかは不明な点も多く，脆弱性に関する理論的な解明が求められる．特に，その事例は画像認識に関して多く，多次元の画像以外のデータに対する訓練済みモデルの脆弱性を明らかにする必要がある．

また，機械学習アルゴリズムに関するセキュリティの研究は，認識精度だけで論じられることが多いが，実際には，システムレベルのセキュリティは，リスクの大きさを把握することが重要になる．セキュリティリスクの大きいものから優先的に対策を施す必要があるからである．そのため，どのような判断を間違うのか，間違いを起こすのがどの程度難しいのか，間違いが検出できるのかなど，間違いの内容に踏み込んでセキュリティを分析する必要がある．今後は，システムレベルのセキュリティリスクと機械学習の脆弱性との関係を明確にしていく必要があろう．そして，セキュリティリスクが大きくなる間違いを減らすための訓練アルゴリズムや，推論結果を間違えた場合のリスクを軽減するようにシステムを設計する方法を確立する必要がある．

さらに，近年，機械学習のアルゴリズムの発展は日進月歩である．新たなアルゴリズムに関するセキュリティの観点の評価とその脆弱性の情報共有が重要になり，脆弱性が発見されたアルゴリズムを用いた訓練済みモデルを効率良く更新する仕組みが求められる．機械学習システムの脆弱性の情報共有方法と，セキュリティアップデートのために訓練済みモデルを効率良く更新する方法は，今後検討すべき課題である．

参考文献

1) Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Xiao, C., … and Song, D. : Robust Physical-World Attackson Deep Learning Models (2017). Retrieved from http://arxiv.org/abs/1707.08945
2) Fredrikson, M., Jha, S. and Ristenpart, T. : Model Inversion Attacks that Exploit Condence Information and Basic Countermeasures, Proceedings of the 22nd ACMSIGSAC Conference on Computer and Com-munications Security - CCS '15, pp.1322-1333 (2015).
3) Goodfellow, I. : Adversarial Examples, Deep Learning Summer School Montreal (2015), Retrieved from http://www.iangoodfellow.com/slides/2015-08-09-adv.pdf

(2018 年 9 月 3 日受付)

■吉岡信和（正会員） nobukazu@nii.ac.jp

2002 年より国立情報学研究所に勤務，現在，同研究所准教授，2007 年より総合研究大学院大学准教授を兼務，セキュリティ・プライバシソフトウェア工学，ソフトウェア工学，学術クラウドの研究・開発に従事．