機械学習システムのセキュリティに関する研究動向と課題

宇根正志（うねまさし）

要　旨

近年、金融を含む幅広い分野において、人工知能を活用した新しいシステムやサービスの開発・提供が進展している。そうしたサービスを安全に提供するためには、機械学習の機能を実装したシステム（機械学習システム）のセキュリティに配慮しておくことが重要である。本稿では、機械学習システムのモデルやセキュリティ対策の方針を示し、既知の主な脆弱性や攻撃手法に加え、攻撃への対策手法に関する最近の研究事例を紹介する。最後に、機械学習システムを安全に活用していくうえで留意すべき事項を示す。

キーワード：　機械学習、人工知能、脆弱性、セキュリティ

本稿の作成に当たっては、神戸大学の小澤誠一教授から有益なコメントを頂いた。ここに記して感謝したい。ただし、本稿に示されている意見は、筆者個人に属し、日本銀行の公式見解を示すものではない。また、ありうべき誤りはすべて筆者個人に属する。

宇根正志　日本銀行金融研究所企画役（E-mail: masashi.une@boj.or.jp）

1. はじめに

近年、金融を含む幅広い分野において、人工知能（artificial intelligence: AI）を活用した新しいシステムやサービスの開発・提供が注目を集めている（金融情報システムセンター［2017］、Financial Stability Board [2017]、中林［2018］）。AI は、一般に、推論、認識、判断等、人間と同様の知的な処理能力をもつコンピュータ・システムやその技術分野を指すことが多い（人工知能学会［2017］）。AI が人間と同様の知的な処理能力を実現・発揮するためには、画像や音声等を認識し、それに基づいて判定・予測等を行う必要があり、通常、そのためのツールとして機械学習（machine learning）が用いられる。現在、深層学習をはじめ、さまざまなタイプの機械学習の手法について実用化に向けた研究開発が活発となっており、技術面の検討のみならず、それらを活用したシステムの開発にかかるガイドラインの策定や、社会・経済に及ぼす影響に関する検討も盛んに行われている（Sze *et al.* [2017]、Brundage *et al.* [2018]、Chio and Freeman [2018]、AI ネットワーク社会推進会議［2017］）。

金融分野において新しい技術を導入し活用する際には、その技術やそれを実装したシステムのリスクに応じたセキュリティ対策を講じる必要がある（金融情報システムセンター［2018］、日本銀行金融機構局［2017］）。これは、機械学習の機能を実装したシステム（以下、機械学習システム）についても同様である。機械学習では、通常、学習モデルに訓練データを入力して（学習済みの）判定・予測エンジンを生成するとともに、そのエンジンを用いてデータの判定・分類や予測を実行する（本稿では教師あり学習のみを対象としている）。こうしたシステムで取り扱われるデータ、学習モデルや判定・予測エンジンの機密性や完全性等を分析・評価し、そのシステムに対して設定したビジネス要件が充足されているか確認しておく必要がある（Barreno *et al.* [2010]、Gardiner and Nagaraja [2016]、Brundage *et al.* [2018]）。そのためには、情報システム一般に存在する脆弱性やそれを悪用した攻撃に加え、機械学習に特有とみられる脆弱性等も把握しておくことが重要である。そうした脆弱性やとりうる対策に関しては、学界を中心に多くの研究蓄積が存在するものの、機械学習システムのユーザーを対象に、最近の動向をサーベイし網羅的に紹介した論考は、ほとんど見当たらない。

こうした状況を踏まえ、本稿では、機械学習システムの主な脆弱性と攻撃手法、攻撃への対策手法について、最近の研究成果を紹介する。まず、機械学習システムのモデルを設定し、システム・セキュリティの観点から、想定される脅威やセキュリティ対策の方針を整理する。次に、最近の研究成果を参照しつつ、機械学習に特有の脆弱性、それらを悪用した主な攻撃手法等を紹介する。さらに、主な対策手法とその有効性にかかる評価手法を紹介し、機械学習システムを安全に活用していく

うえでの留意事項を考察して本稿を締めくくる。

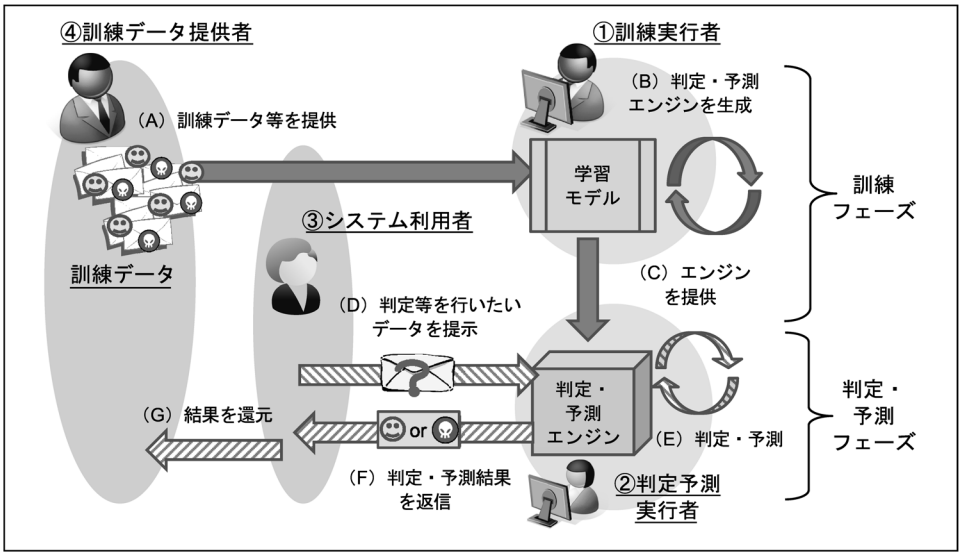
2. 機械学習システムと脅威

(1) システムの構成

機械学習システムは、一般に、次の 4 つのエンティティによって構成される。①訓練データと学習モデルを用いて判定・予測エンジンを生成する訓練実行者、②訓練実行者から判定・予測エンジンを受け取り、判定・予測を実行する判定予測実行者、③判定・予測エンジン生成やデータの判定・予測を依頼するシステム利用者、④訓練データを訓練実行者に提供する訓練データ提供者である[1]。判定・予測エンジンの生成や判定・予測における主な処理の流れは次のとおりである（図表 1 を参照）。

（A）訓練データ提供者は、訓練データの元になるデータを収集した後、システム利用者と協力しつつ、データを適宜加工するとともに、必要に応じてラベル

図表 1　機械学習システムとエンティティ（概念図）



1 訓練データ提供者とシステム利用者が同一の場合や、訓練実行者と判定予測実行者が同一の場合もありうる。

（訓練データにかかる判定結果等を表すデータ）を付加したうえで、訓練実行者に提供する。

（B）訓練実行者は、訓練データを学習モデルに適用し判定・予測エンジンを生成する。

（C）訓練実行者は、生成した判定・予測エンジンを判定予測実行者に提供する。

（D）システム利用者は、判定・予測を行いたいデータを判定予測実行者に提示する。

（E）判定予測実行者は、上記（D）でシステム利用者から受け取ったデータを判定・予測エンジンに適用し、判定・予測を行う。

（F）判定予測実行者は、上記（E）での判定・予測結果をシステム利用者に返信する。

（G）システム利用者は、上記（F）での判定・予測結果等を訓練データ提供者に還元する場合がある。例えば、訓練データ提供者は、判定・予測結果が誤っていた際に、それを修正し、正しいラベルを付加して訓練実行者に与え、判定・予測エンジンの改善を図るケースが考えられる。

上記（A）～（C）が訓練フェーズに対応し、上記（D）～（G）が判定・予測フェーズに対応する。なお、（A）における訓練データ等の提供に関しては、訓練データが機微な情報の場合、マスキング等の実施や暗号化などを行うケースが考えられるが、ここでは、分析を単純化するために、そうした処理が完了したデータが訓練実行者に提供されるものとする。

(2) セキュリティ目標

機械学習システムのセキュリティ目標として、そのシステムで取り扱われるデータや機能の機密性（confidentiality）・完全性（integrity）・可用性（availability）の達成が求められる（例えば、Barreno *et al.* [2010]、Papernot *et al.* [2016b]）[2]。本節（1）で示した機械学習システムの場合、保護対象となりうるデータや機能は、①訓練データ、②学習モデル、③判定・予測エンジン、④判定・予測を行う対象となるデータ（判定・予測用データ）、⑤判定・予測用データを用いた判定・予測エンジンの出力

2 ここでの機密性は「機械学習システムで取り扱われるデータや機能が無権限者に知られないこと」を、完全性は「データ等が不正に偽造・改変されないこと」を意味する。可用性は「機械学習システムが正常に稼働しサービスが滞りなく提供されること」を意味する。Papernot *et al.* [2016b] は、情報システム一般のセキュリティを論じる際に用いられるこれらのセキュリティ特性が機械学習システムにも有用であるとしている。また、Barreno *et al.* [2010] では、不正侵入検知システム等のセキュリティ対策に用いられる機械学習システムに焦点を当てて、完全性と可用性をセキュリティ目標として検討している。

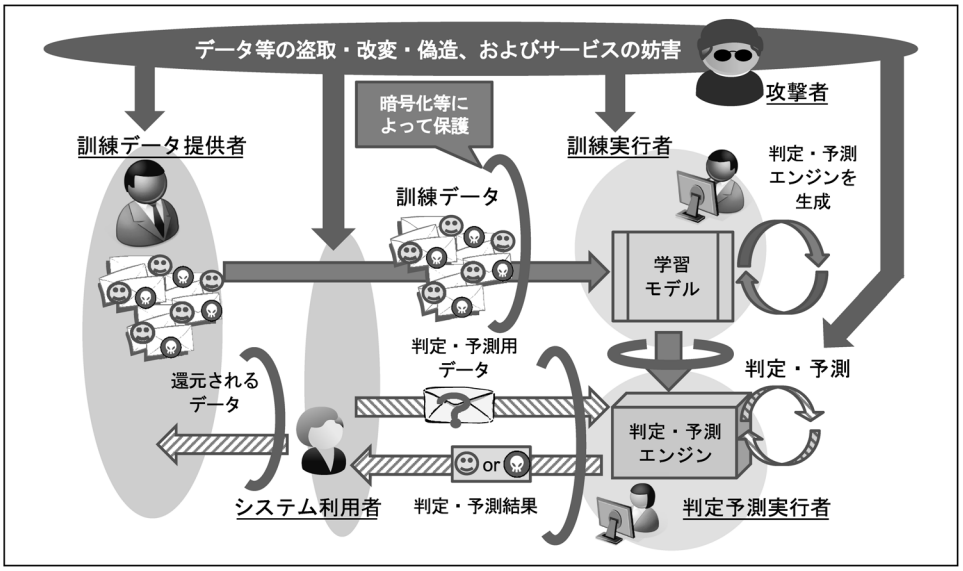
（判定・予測結果）、⑥システム利用者が訓練データ提供者に還元するデータ（還元データ）である。

例えば、訓練データに着目すると、機密性の観点からは、訓練データ提供者にかかる機微な情報（個人情報等）が含まれているなどの場合には、そうしたデータの盗取を防ぐ必要がある。完全性の観点からは、訓練データの改変や不当なモデルの生成（機能の改変）が判定・予測に大きな影響を与える可能性がある場合には、それらを防ぐ必要がある。可用性の観点からは、訓練データを訓練実行者に対して大量に送信する攻撃が行われ、訓練実行者の機能が低下する可能性がある場合には、そうした攻撃を防ぐ必要がある。各保護対象について、3 つのセキュリティ特性（機密性、完全性、可用性）の要否（およびその達成度合い）を検討したうえで、必要と判断した特性に関して、どのようなセキュリティ対策を講じるかを検討することが求められる。

(3)　攻撃と対策方針

セキュリティ対策の内容を検討するうえで、想定される攻撃を洗い出すことが必要である。本節（1）で示した機械学習システムを前提とすると、各エンティティと、エンティティ間の通信路が攻撃箇所となりうる（図表 2 を参照）。

図表 2　機械学習システムへの攻撃のポイント（概念図）



図表 3　各保護対象への主な攻撃と対策方針

<table>
<tr><th colspan="2" rowspan="2">攻撃箇所（どこで）</th><th rowspan="2">攻撃対象（なにを）</th><th colspan="2">攻撃（どうする）</th><th rowspan="2">対策方針</th></tr>
<tr><th>目的</th><th>手段</th></tr>
<tr><td rowspan="4">エンティティ</td><td>訓練データ提供者</td><td>・訓練データ
・判定・予測結果
・還元データ</td><td rowspan="4">盗取

改変・偽造

データの使用や機能の妨害</td><td rowspan="4">学習モデル等の脆弱性を悪用

各通信相手になりすましてアクセス

外部ネットワーク接続部分の脆弱性等を悪用してアクセス

大量のサービス要求を送信</td><td rowspan="4">学習モデル等の脆弱性の軽減・解消

通信相手の認証

端末やサーバ等へのアクセスの制御

保管データの暗号化やそれらの改変の検知

サーバ等への負荷軽減（CDN の利用等）</td></tr>
<tr><td>システム利用者</td><td>・判定・予測用データ
・判定・予測結果
・還元データ</td></tr>
<tr><td>訓練実行者</td><td>・訓練データ
・学習モデル
・判定・予測エンジン</td></tr>
<tr><td>判定予測実行者</td><td>・判定・予測用データ
・判定・予測エンジン
・判定・予測結果</td></tr>
<tr><td rowspan="4">通信路</td><td>訓練データ提供者と訓練実行者</td><td>・訓練データ</td><td rowspan="4">盗取

改変・偽造

通信の妨害</td><td rowspan="4">左記の各通信路にアクセス（中間侵入攻撃）</td><td rowspan="4">データの暗号化や認証（TLS 等のプロトコルの適用等）

通信路の負荷軽減（CDN の利用等）</td></tr>
<tr><td>訓練実行者と判定予測実行者</td><td>・判定・予測エンジン</td></tr>
<tr><td>判定予測実行者とシステム利用者</td><td>・判定・予測用データ
・判定・予測結果</td></tr>
<tr><td>システム利用者と訓練データ提供者</td><td>・判定・予測結果
・還元データ</td></tr>
</table>

イ．各エンティティへの攻撃

各エンティティへの攻撃として、それぞれが取り扱うデータ、学習モデル、判定・予測エンジンに関する情報の盗取・改変・偽造に加え、学習モデルや判定・予測エンジンを実行するサーバを停止させるなどのサービスの妨害行為が想定される（図表 3 を参照）。例えば、訓練実行者については、訓練データの盗取に加え、学習モデルや判定・予測エンジンにかかる情報の盗取が考えられる。また、訓練データ、学習モデル、判定・予測エンジンの改変・偽造のほか、訓練データの受信や学習モデルの実行等の妨害も考えられる。

攻撃の手段としては、学習モデル等の脆弱性を悪用することがまず想定される。また、各エンティティの通信相手になりすます、あるいは、外部ネットワークとの接続部分の脆弱性を悪用して、そのエンティティの端末やサーバ等に不正にアクセスすることが想定される[3]。また、機械学習システムに特有の事情として、訓練デー

3 通信相手へのなりすましに関しては、AI スピーカーや（AI 機能を実現する）スマートフォンに対し

タ提供者が（意図せずに）不正な訓練データを入手し訓練実行者に送信する（その結果、不正な判定・予測エンジンが生成される）可能性もある。さらに、可用性低下にかかる攻撃として、大量のサービス要求を各エンティティに送信してサーバをダウンさせることなどが考えられる。

対策の方向性としては、まず、機械学習システムに特有の脆弱性を軽減することが挙げられる。また、一般の情報システムにおいても想定されるものとして、主に機密性と完全性の観点から、①通信相手の認証、②各エンティティの端末やサーバ等（各種データ等を格納）へのアクセス制御、③保護対象の各種データの暗号化、④データベース上の各種データの改変（データベースへの入力前の改変は除く）の検知等が挙げられる[4]。さらに、可用性の観点からは、⑤コンテンツ配信ネットワーク（Contents Delivery Network: CDN）の利用等が考えられる。

通常、上記①～⑤の対策を十分に実施すれば、攻撃者は各エンティティが保有するデータ等にアクセスできず、それらを攻撃に利用できないと想定可能となるほか、可用性を維持することができる。この場合、セキュリティ対策の検討においては、機械学習システムに特有の脆弱性に焦点を当てることとなる。もっとも、各エンティティへの不正侵入、マルウェアによる攻撃、訓練実行者等へのソーシャル・エンジニアリング攻撃等、いわゆるサイバー攻撃が今後一層高度化する可能性は否定できない。そのため、サイバー攻撃の高度化のリスクにも配慮し、各エンティティが保有するデータ等に攻撃者がアクセスするケースも想定して検討する必要もある[5]。

ロ．エンティティ間の通信路での攻撃

エンティティ間の通信路では、両端のエンティティの通信を中継するように通信路にアクセスし（中間侵入攻撃）、通信データの盗取、改変・偽造を試行することが想定される。また、大量のサービス要求を各エンティティに送信することによって通信路の帯域を制限するなどの通信の妨害も考えられる。通信データの盗取や改変・偽造への対策方針として、TLS（Transport Layer Security）等の暗号プロトコルを活用することが考えられる。サービス妨害への対策方針としては、CDN の利用等が考えられる。これらは、いずれも情報システム一般において広く利用されてお

て偽の音声を提示し、正規のシステム利用者になりすますという攻撃が提案されており、通信相手の認証等による対策が必要とされている（飯島ほか［2018］）。その他の攻撃として、各エンティティの一部の内部者と結託する、マルウェアを用いて端末やサーバ等を遠隔操作するといったケースも想定される。

4 データの暗号化については、訓練データを暗号化したまま学習や判定・予測を行う手法等が提案されている（例えば、Dowlin *et al.* [2016]、Phong [2017]、Phong *et al.* [2018]）。

5 具体的にどのような情報が悪用されることを想定するかについては、既存研究においていくつかのレベル分けが行われている。詳細は 3 節（1）で説明する。

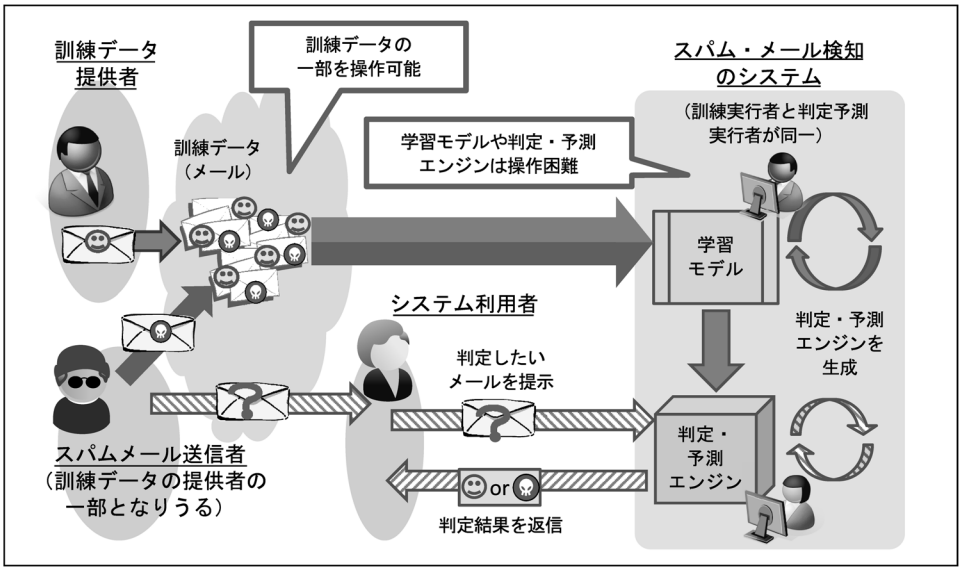
り、機械学習システム特有のものではない[6]。

ハ．考察例：スパム・メール対策の場合

実際のセキュリティ評価や対策は、個別のアプリケーションや実装環境等に応じて実施することになる。一例として、システム利用者が、自分宛のメールのなかからスパム・メール（spam mail）を検知・排除する目的で機械学習を用いたスパム・メール検知のシステムを利用する場合を考える（図表 4 を参照）。ここでは、想定される形態の 1 つとして、そのシステムが、訓練実行者と判定予測実行者の機能を有し、同一のエンティティによって運用されているとする。訓練データは、不特定多数の訓練データ提供者からインターネットを介して送信されるメール（スパムとそうでないものが混在）に対応する。したがって、スパム・メール送信者（攻撃者）が、訓練データや判定・予測エンジンへの入力（メール）を生成する可能性があり、訓練データ等にかかる情報を有し、その一部を操作することができるといえる。情報システム一般において広く利用されているセキュリティ対策が講じられていれば、学習モデルや判定・予測エンジンにかかる情報が秘匿されるとともに、システムへのアクセスも厳重に制御される状況を想定できる。

このような場合、攻撃者は、学習モデルや判定・予測エンジンへのアクセスが困難である一方、訓練データの一部や判定・予測エンジンへの入力データを操作する

図表 4　スパム・メール検知の機械学習システムへの攻撃（イメージ）



6 こうした対策については、金融情報システムセンターの安全対策基準においても規定されている（金融情報システムセンター［2018］）。

ことができる。したがって、一般的な情報システムで求められるセキュリティ対策を講じるとともに、訓練データや判定・予測エンジンへの入力データが不正に操作される可能性に配慮しつつ、どのような脆弱性が悪用されうるかを検討することが重要となる。

3. 学習モデルにかかる脆弱性と攻撃手法

本節では、学習モデルにかかる脆弱性やそれらを悪用した攻撃手法の研究事例を、最近の主な研究論文（Papernot *et al.* [2016b] 等）を引用しつつ紹介する。

(1) 攻撃者の能力にかかる前提

攻撃手法に関する個々の研究では、攻撃者の保有する情報や行動（攻撃者の能力）が異なっている。したがって、想定される攻撃者の能力をあらかじめ分類しておくことは、複数の攻撃手法のインパクトを横並びで比較するうえで有用である（Carlini and Wagner [2017a]）。

攻撃者の能力にかかる大きな分類として、ホワイト・ボックスとブラック・ボックスが広く知られている。ホワイト・ボックスは、攻撃者が、対象とする判定・予測エンジンの構造やパラメータ（損失関数や重み等）、エンジンの任意の入出力等、ほぼ完全な情報を得ることができる状況を意味する。一方、ブラック・ボックスは、こうした情報の入手や判定・予測エンジンへのアクセスに一定の制限が課せられている状況を意味する。こうしたホワイト・ボックスとブラック・ボックスの境界は研究論文によって異なっている。

2 節（3）で説明したように、サイバー攻撃を想定するとホワイト・ボックスの状況を想定した対策が必要となるが、学習モデルや判定・予測エンジンが企業秘密として厳重に管理されている場合等では、ホワイト・ボックスの状況が実現する可能性は相対的に低く、まずは、ブラック・ボックスでの攻撃が焦点となる（Suciu *et al.* [2018]）。攻撃者が利用する情報の種類として、これまでさまざまな分類が示されている。ここでは、最近の代表的な研究として先崎・大畑・松浦［2018］の分類を紹介する。この分類では、攻撃者が利用する情報の組合せを次の 6 つに分類している。

【分類1】判定・予測エンジンへのいくつかの入出力ペア（攻撃者が指定可能）

【分類2】判定・予測エンジンの任意の入力データに対する加工された出力[7]
【分類3】判定・予測エンジンの任意の入出力ペア
【分類4】判定・予測エンジンの任意の入出力ペア、訓練データ
【分類5】判定・予測エンジンの任意の入出力ペア、学習モデルのネットワーク構造
【分類6】判定・予測エンジンの任意の入出力ペア、学習モデルのネットワーク構造、訓練データ

比較的実現性の高い状況は分類 1〜3 であり、まずは、攻撃者がこれらの情報を利用可能と想定したうえで有効な対策を検討することが重要である。

(2) 学習モデル等にかかる脆弱性と攻撃手法

学習モデル等にかかる脆弱性のうち、セキュリティと密接に関係すると考えられるものを整理すると、①訓練データにかかる情報の漏洩、②判定・予測エンジンにかかる情報の漏洩、③訓練データの変化による判定・予測エンジンの精度低下、④入力の変化による判定・予測の精度低下が挙げられる。また、これらを利用した攻撃については、いずれも、攻撃者が機械学習システムにかかる何らかのデータを入手して実施するものである。

イ. 訓練データにかかる情報の漏洩

学習モデルの構造、判定・予測エンジン、そのエンジンの入出力から、特定のデータが訓練データの一部であったか否かの情報や、訓練データの特性にかかる情報が漏洩しうるほか、訓練データ自体も推定されうる（図表 5 を参照）[8]。

（イ） 特定のデータが訓練データの一部であったか否かの情報の漏洩

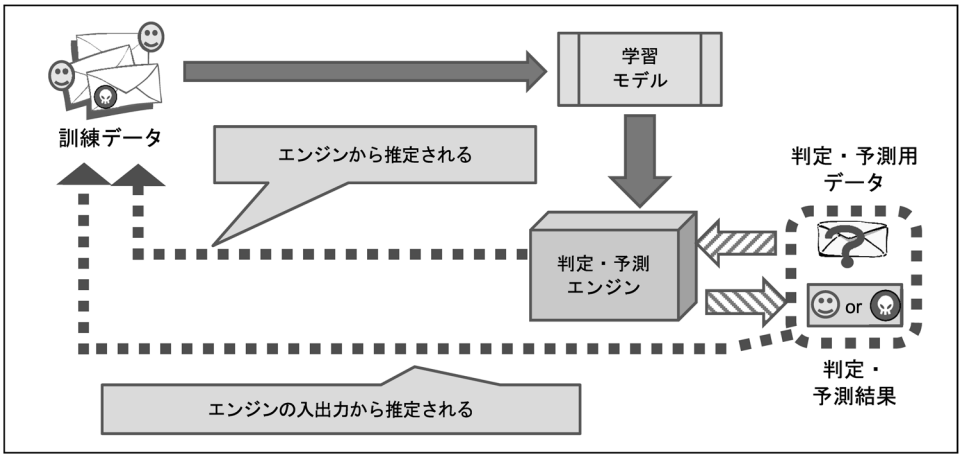
特定のデータが訓練データの一部であったか否かについては、例えば、画像認識、個人の購買履歴、医療機関受診履歴を利用する一部の機械学習システムにおいて、その情報が漏洩しうることを示す研究事例が知られている[9]。これらは、訓練

7 分類 2 の「判定・予測エンジンの任意の入力データに対する加工された出力」は、不正な入力データ等による攻撃への対策が実装されている場合を想定し、その対策によって影響を受けた出力を意味する。

8 機械学習システムの判定・予測エンジンを生成するベンダーは、エンジンの性能を向上させる訓練データの特性を企業秘密としているケースが考えられる。その場合、判定・予測エンジン等から訓練データにかかる情報が漏洩することを回避したいというニーズが存在する（Ateniese *et al.* [2015]）。

9 医療機関受診履歴から特定の疾病を判定する判定・予測エンジンにおいて、個人を特定可能なデータ（氏名等）が訓練データに含まれていた場合、ある個人にかかるデータが訓練データに含まれていたことが判明すると、その個人の健康状態が推定される可能性がある。

図表５　訓練データにかかる情報の漏洩（イメージ）



データにおける特定のデータの有無によって、生成される判定・予測エンジンの入出力関係が異なるという性質を利用している。

Shokri *et al.* [2017] では、複数の購買履歴等のデータを用いて複数の判定・予測エンジン（特定のデータが訓練データに含まれている場合のエンジンやそうでない場合のエンジン）を生成し、それらのエンジンの入出力を分析することができれば、特定のデータが訓練データに含まれていたか否かを高い確率で推定可能であることを示している[10]。具体的には、（推定対象以外の）訓練データの一部とそれらに対する判定・予測エンジンの出力等を利用し、70～90％の確率で入出力関係が再現可能なエンジンを生成するというものである。攻撃者は、生成した判定・予測エンジンを用いて、特定のデータが訓練データに含まれるか否かを判定する。この攻撃は、対象となる判定・予測エンジンや学習モデルが秘匿されていた場合でも有効であり、本節（1）分類４に相当する情報を用いた攻撃といえる。また、この研究は、クラウドが提供する一部の機械学習サービス（学習モデルや判定・予測エンジンの内容は秘匿）に適用可能であることを実証している[11]。

（ロ）　訓練データの特性にかかる情報の漏洩

訓練データの特性にかかる情報の漏洩に関しては、例えば、一部の音声認識のシ

10　こうした攻撃は、「membership inference attack」と呼ばれることがある。

11　近年、クラウド上で機械学習システムを実行するサービス（“ML-as-a-service” とも呼ばれる）が提供されるようになってきている。例えば、アマゾン社（Amazon Machine Learning）、グーグル社（Google Cloud Platform）、マイクロソフト社（Azure Machine Learning Studio）、BigML 社等が挙げられる。システム利用者は、クラウド上で判定・予測エンジンを生成したり、クラウド上のエンジンを用いて判定・予測結果を取得したりすることができる。このようなサービスでは、通常、判定・予測エンジンにかかる情報はシステム利用者から秘匿される。

ステムにおいて、訓練データの大半が特定の方言を含む音声データであったか否かが推定されうるという事例や、通信データから通信サービスの種類を判定するシステムにおいて、特定の大手インターネット・サービス・プロバイダーのサーバからの通信データが訓練データの大半を占めていたか否かが推定されうるという事例に関する研究が知られている（Ateniese *et al.* [2015]）[12]。この研究では、学習モデルや判定・予測エンジンにかかる情報を用いて、訓練データの特性を判定する判定・予測エンジンを生成する手法を示すとともに、隠れマルコフ・モデルに基づく一部の音声認識のモデルと、サポート・ベクトル・マシン（support vector machine）に基づく通信サービスを識別するモデルへの適用例（90％程度の確率で判定に成功）を報告している。この攻撃は、本節（1）分類 5 の情報を用いた攻撃に相当する。

（ハ） 訓練データ自体の推定

判定・予測結果の確からしさを示す確信度（confidence value）が判定・予測エンジンの出力に含まれている場合には、訓練データ自体が推定されうる。

例えば、一部の顔画像認識のシステム（確信度を出力するもの）において、訓練データとして個人の識別情報や顔画像等が使用されていた場合に、判定・予測エンジンの入出力等から顔画像を推定する研究が知られている（Fredrikson *et al.* [2014]、Fredrikson, Jha, and Ristenpart [2015]）[13]。訓練フェーズでは、個人の識別情報（例えば、氏名）と顔画像を訓練データとして使用し、判定・予測フェーズでは、顔画像を判定・予測エンジンに入力することで、対応する個人の認識情報と確信度を出力として得る。こうしたシステムを対象に、判定・予測エンジンの複数の入出力から、特定の個人の顔画像（あるいはその逆）を推定する。

これらの研究は、ソフトマックス関数を用いたニューラル・ネットワークやパーセプトロンに基づく一部の学習モデルに提案手法を適用している。ソフトマックス関数の場合、その入出力や内部の構造にかかる情報を利用可能な場合には、提案手法によって推定した特定の個人の顔画像が、80～90％の確率でその個人の（登録された）顔画像と一致すると判定された旨を報告している。この攻撃は、本節（1）分類 6 の情報を用いた攻撃といえる。

ロ． 判定・予測エンジンにかかる情報の漏洩

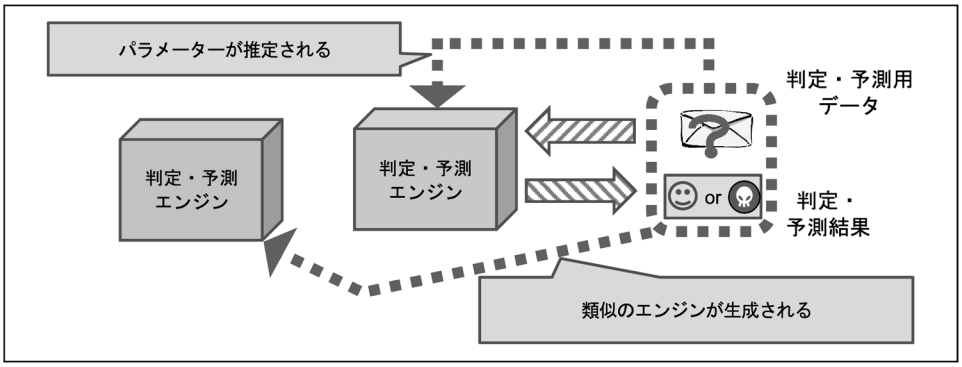
判定・予測エンジンの入出力から、そのエンジンのパラメータ等にかかる情報が推定される事態が生じうる（図表 6 を参照）。

例えば、Tramèr *et al.* [2016] では、判定・予測をクラウド上で提供するサービス（判定・予測エンジンへの入力はサービスの利用者がネットワーク経由で送信）の

12 音声認識システムが高い精度を達成している場合、訓練データとしての音声データの情報を推定できれば、このシステムの精度を向上させる要因の推定も可能となる。

13 こうした攻撃は、「model inversion attack」と呼ばれる。

図表 6　判定・予測エンジンにかかる情報の漏洩（イメージ）



うち、判定・予測結果の確信度をエンジンが出力するタイプについて、エンジンのパラメータを推定するとともに、ほぼ同一の入出力関係を実現する代替エンジンを生成する手法が提案されている[14]。

ロジスティック回帰や決定木の手法に基づくモデルを利用する実際のサービスに適用した場合、数百から数千の入出力ペアを用いて判定・予測エンジンのパラメータを推定することができれば、代替エンジンの生成に 90％以上の確率で成功する旨も報告されている[15]。この攻撃は、攻撃者がエンジンのパラメータや訓練データに関する知識を有していないことから、本節（1）分類 3 の情報を用いた攻撃に相当する。

ハ．訓練データの変化による判定・予測エンジンの精度低下

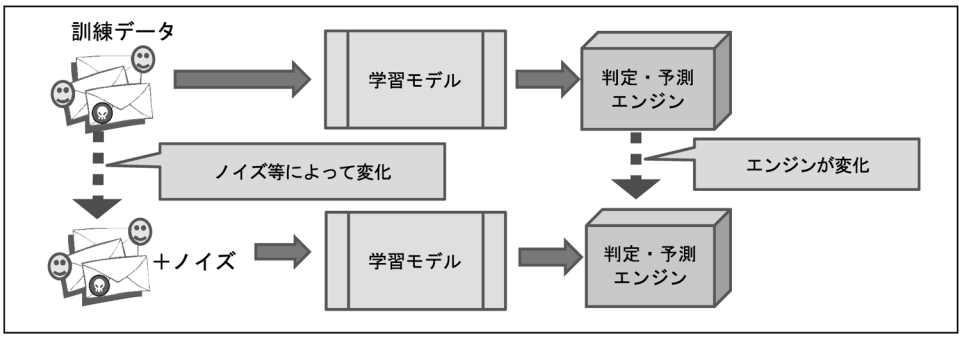
訓練データの分布がノイズ等によってわずかに変化した際、それらによって生成される判定・予測エンジンが有意に変化し、誤った判定・予測が出力される場合がある（Biggio, Nelson, and Laskov [2011, 2012]、Biggio *et al.* [2013]、Barreno *et al.* [2010]）。その結果、判定・予測エンジンの精度が低下することになる（図表 7 を参照）。こうした脆弱性を悪用する攻撃として、不正な訓練データ等を学習モデルに入力し、攻撃者にとって都合のよい判定・予測エンジンを生成させるという攻撃がよく知られている（Barreno *et al.* [2010]、Papernot *et al.* [2016a]、Chio and Freeman [2018]、Suciu *et al.* [2018]）[16]。こうした攻撃は、サポート・ベクトル・マシン、ロ

14 こうした攻撃は、「model extraction attack」と呼ばれる。

15 攻撃者が判定・予測エンジンの入出力を取得する際に、過去に取得した入出力を分析して次の入力を選択するタイプ（adaptive attack や hill-climbing attack と呼ばれる）と、過去に取得した入出力と独立に選択するタイプ（non-adaptive attack と呼ばれる）が存在する。Tramèr *et al.* [2016] では、判定・予測エンジンの種類に応じて両方のタイプを使い分けている。

16 こうした攻撃は、訓練データを不正に操作する点に主眼を置く場合には、「poisoning attack」または「training-set attack」と呼ばれるほか、異常検知等のアプリケーションにおいて、（本来検知すべき）異常な事象を検知できないように判定・予測エンジンを不正に操作するという点に主眼を置く場合に

図表 7　訓練データの変化による判定・予測エンジンの精度低下（イメージ）



ジスティック回帰、ニューラル・ネットワークに基づく一部の学習モデル等に対して適用可能であることが示されている[17]。各研究成果では、攻撃者が学習モデル等について事前に知識を有している状況のもとで、高い成功率を達成しうる不正な訓練データを探索する手法の検討に主眼が置かれており、本節（1）分類 5 あるいは分類 6 の情報を用いた攻撃といえる。

例えば、Kloft and Laskov [2010] は、（攻撃と疑われる）不正な通信か否かの判定を実施したうえで、判定対象となった通信データを訓練データとして使用して判定・予測エンジンを順次更新するタイプの不正通信検知のモデル（重心モデル）を攻撃対象とする研究である[18]。この研究では、与えられた訓練データに対して不正なデータを追加することにより、不正な通信か否かを判定する境界を徐々に移動させることができることを示したうえで、その移動が最大となるようなデータを探索する問題を定式化し、その解法を提案している。また、訓練データ全体のうち、どの程度のデータを改変すれば、（訓練データによって生成された）判定・予測エンジンにおいてどの程度の確率で誤判定が発生するかについても関係性を示している。提案手法は、攻撃者が学習モデルと訓練データを知っているという状況を前提としたものであり、本節（1）分類 6 の情報を用いた攻撃に相当する。

Mei and Zhu [2015] は、攻撃用の訓練データと本来の訓練データの差分を一定以下にするという制約のもとで、攻撃者が目標とする判定・予測エンジンとの差分を最小化するエンジンを生成するように、攻撃用の訓練データを探索する問題とその効率的な解法を提案し、サポート・ベクトル・マシン、ロジスティック回帰、線形回帰に基づく一部の学習モデルへの適用事例を示している。Kloft and Laskov [2010]

は、「evasion attack」と呼ばれることが多い。

17 訓練データを操作する方法として、教師あり学習の場合、ラベルのみを不正に操作するという攻撃も知られている（Biggio *et al.* [2013]）。もっとも、そうした方法は、現時点では計算量が大きくなる傾向にあり、攻撃の効果が低いとの見方もある（Papernot *et al.* [2016b]）。

18 判定・予測エンジンを順次更新するという機械学習システムの性質は「アクティブ・ラーニング」と呼ばれる。

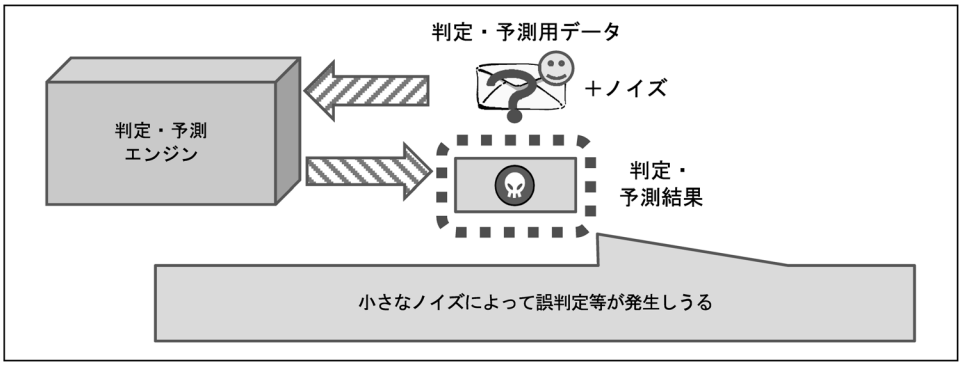
と同様に、攻撃者が学習モデルと訓練データを知っていることが前提とされており、本節（1）分類 6 の情報を用いた攻撃といえる。

ニ．入力の変化による判定・予測エンジンの精度低下

判定・予測エンジンへの入力がノイズ等の影響によってわずかに変化した際、誤った判定・予測結果が出力される場合がある（図表 8 を参照）。こうした脆弱性を悪用して判定・予測エンジンに誤判定等を引き起こす攻撃手法が数多く提案されている（Szegedy *et al.* [2014]、Nguyen, Yosinski, and Clune [2015]、Sinha, Kar, and Tambe [2016]、Kenway [2018]、Papernot *et al.* [2016a, 2017]、Carlini and Wagner [2017b]、小澤［2018］等）[19、20]。

最近では、深層学習に基づく機械学習のモデルを対象とした研究成果の発表が目立つ。Szegedy *et al.* [2014] では、深層学習に基づく画像認識や手書き文字認識の一部のモデルを対象に、誤った判定結果が出力される入力データを探索する手法を提案している。こうした入力データは、例えば、訓練データ（画像）に一定のノイズが付加されたデータとして表現されたりする。提案手法は、目標とする（誤った）判定結果の出力を実現しつつ、その判定・予測結果の誤差を最小化する入力の近似値を探索するという問題を定式化するとともに、その近似解を効率的に求めるものである。深層学習等に基づく 4 種類の学習モデルに提案手法を適用したところ、訓練データに微小なノイズを付加した（誤判定を引き起こす）入力データを探索することができた[21]。攻撃者は、学習モデルや判定・予測エンジンにかかる情報を入手

図表 8　入力の変化による判定・予測エンジンの精度低下（イメージ）



19 このような脆弱性は強化学習の場合でも存在する（Huang *et al.* [2017]）。
20 攻撃に用いられる（判定・予測エンジンへの）入力は「adversarial example」と呼ばれる。
21 攻撃用の入力データ（画像）の探索には、非線形連立方程式の近似解を効率的に解く手法の 1 つである「L-BFGS 法」が利用されている。また、探索結果の入力データは、グレースケールの変化量が 1 ピクセルあたり平均 1％未満のノイズを付加した訓練データであった（人間の肉眼では、ノイズ付加前の訓練データとの差分を検知することは困難）。

することが前提とされており、本節（1）分類 6 の攻撃に相当する。

Papernot *et al.* [2016a] では、深層学習を用いた手書き文字認識の一部の学習モデルを対象に、攻撃用の入力データを探索する手法を提案している。提案手法では、攻撃者が訓練データ、学習モデル、判定・予測エンジンを事前に知っている状況を想定しており、本節（1）分類 6 の情報を用いた攻撃に相当する。そのうえで、入力データに付加されるノイズや改変が判定・予測エンジンの出力に及ぼす影響を示す関係式を構成し、意図した誤判定を引き起こす攻撃用の入力データを探索している。提案手法の有効性を実験で確認したところ、正規の入力データ（文字画像）を構成するピクセルのうち、平均で約 4％のピクセルに一定の改変を加えると、約 97％の確率で、攻撃者が意図したクラスに誤判定させることができたとしている。

このほか、ある判定・予測エンジンにおいて誤判定等を引き起こしやすい入力は、同一の学習モデルによって生成された別のエンジンにおいても誤判定等を引き起こしうることも知られている[22]。例えば、画像データの判定を行う一部の機械学習システムにおいて、判定・予測エンジンを複数の手法によりそれぞれ生成したうえで、訓練データとして使用した画像データに一定の処理を施して入力すると、複数のエンジンにおいて有意な確率で誤判定が発生した事例がある（Szegedy *et al.* [2014]、Goodfellow, Shlens, and Szegedy [2015]）[23]。

Szegedy *et al.* [2014] では、ある特定の判定・予測エンジンにおいて誤判定等を引き起こす入力が、訓練データ、レイヤー数、重み減衰のパラメータ等を変更して生成した他のエンジンにおいても比較的高い確率で誤判定等を発生させる場合があることを示している。例えば、ソフトマックス関数を用いたニューラル・ネットワークにおいて、重み減衰のパラメータを変化させつつ複数の判定・予測エンジンを生成したうえで、あるエンジンにおいて誤判定等を引き起こす（攻撃用の）入力データを探索してそれを他のエンジンに適用したところ、10～80％の確率で誤判定が発生した旨を報告している。

4. 攻撃への対策手法

本節では、対策の有効性を評価するための主な尺度を説明した後、3 節で紹介した攻撃への主な対策手法とその有効性評価にかかる代表的な研究成果を紹介する。

22 ある判定・予測エンジンにおいて誤判定等を引き起こす入力が他の類似のエンジンでも誤判定等を生じさせるという性質は、「cross model generalization」、あるいは、「transferability」と呼ばれる（Papernot *et al.* [2016a, b]）。

23 判定・予測エンジンの入力（画像データ）として、人間が認識することが困難な（微小な）変更を訓練データに加えたものが準備された。

(1) 評価尺度

想定すべき攻撃者の能力のもとで、対策手法により攻撃がどの程度軽減されうるかを定量的に評価しようとする場合、評価尺度が重要となる。既存の研究論文では、判定・予測エンジンの出力の正確性にかかる評価尺度として、①「『不正』と判定された入力データのうち、正しく判定したものの割合」を示す適合率（precision）、②「不正な入力データ全体のうち、正しく『不正』と判定したものの割合」を示す再現率（recall）、③「入力データ全体のうち、正しく判定したものの割合」を示す正解率（accuracy）が用いられるケースが多い。

判定・予測エンジンへの入力データの総数を N、それらのうち、不正な入力データの総数を A（$< N$）とする。そのうえで、N 個の入力データのうち T 個を「不正」と判定したとする。このとき、T 個のうち P 個の入力データが実際に不正なものであったほか、（「不正」と判定されなかった）$T - P$ 個の入力のうち、U 個が実際に正規の入力データであったとすると、適合率は P / T と表され、再現率は P / A と表される。正解率は、$(P + U) / N$ と表される。

また、ノイズや改変を加えた入力データによって誤判定等を引き起こす攻撃の場合には、ノイズや改変の度合いを評価の尺度とするケースもある。例えば、ノイズ等を加えた入力データと元の入力データとの距離（例えば、両データ間のユークリッド距離の平均値）を尺度とすることがある。この距離が小さいほど、不正な入力データとしての検知が困難になると考えられることから、攻撃としての有効性がより高いと考えることができる。逆に、対策を講じる側からみると、対策によって距離がより拡大するほど、攻撃を成功させるためにより多くのノイズや改変を入力データに加える必要が生じるという意味で、対策の効果が相対的に大きいといえる。

もっとも、研究論文によっては、これらの指標がすべて記載されているとは限らず、攻撃手法による判定・予測エンジンへの影響度合いを横並びで比較することが困難な場合が少なくない。さらに、対策手法を講じた結果、判定・予測の精度が有意に低下してしまうと、対策実施の意味が失われることとなる。

Carlini and Wagner [2017a] は、こうした点を指摘したうえで、攻撃手法や対策手法を提案・評価する論文においては、少なくとも、適合率に加えて、実用性とのトレードオフを評価する観点から、「判定・予測エンジンに入力されたデータのうち、正規の入力データを不正と誤って判定したものの割合」を示す偽陽性率（false positive rate）も研究論文に明記すべきであると提案している。また、判定のしきい値を変化させたときに適合率と偽陽性率がどう変化するかを表す ROC 曲線（receiver operating characteristics curve）を示すことによって、対策手法の有効性を示すことができればより望ましいとしている。

(2) 対策手法

対策としては、各攻撃手法を実行するうえで必要とされる情報を攻撃者に入手させないようにする、あるいは、そうした情報が攻撃者に入手されたとしても攻撃が成功しないように学習モデルや判定・予測エンジンを改良することが考えられる。前者は、攻撃者が利用できる情報を制限し、ブラック・ボックスの状況を実現するという対応である。後者は、ホワイト・ボックスの状況を前提として、学習モデル等のセキュリティを向上させるという対応である。

イ. 判定・予測エンジンや訓練データにかかる情報の盗取への対策

判定・予測エンジンのパラメータ等の情報の盗取・推定に対しては、攻撃に必要な（判定・予測エンジンの）出力や確信度等を攻撃者が入手できないようにすることが考えられる。

そうした手法の 1 つとして、確信度の値を丸めたうえで出力する、あるいは、確信度を出力しないようにすることが挙げられる（Fredrikson, Jha, and Ristenpart [2015]）。もっとも、確信度を出力せず、例えば、判定結果として最も確からしいクラスのみを出力するように構成したとしても、より多くの入出力を攻撃者が入手することができる状況であれば、判定・予測エンジンのパラメータ等を推定することができる場合があるという分析結果もある（Tramèr *et al.* [2016]）。

また、暗号化したデータを入力として学習モデルや判定・予測エンジンに適用し、その出力（判定・予測結果）も暗号化したまま得られるようにするという手法が提案されている。こうした手法における暗号として、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ている[24]。例えば、Dowlin *et al.* [2016] では、暗号化したデータのまま訓練や判定・予測を実行可能なニューラル・ネットワークのアルゴリズム（CryptoNets と呼称）が提案されている。また、Phong *et al.* [2018] では、準同型暗号によって、暗号化したデータのまま深層学習を実現する手法が提案されている。これらの研究では、画像データ等を用いた実験により、一定の処理性能と判定・予測の精度が実現可能である旨が示されている。

訓練データを復元・推定する攻撃に関しては、既存の攻撃を実施するうえで確信度が必要となることから、上記と同様に、確信度が出力されないように判定・予測エンジンを構成することが考えられる。

ロ. 訓練データや入力データの操作への対策

訓練データや入力データを操作する攻撃への対策について、これまでに数多くの研究成果が報告されている。主な対策手法は、学習モデルや判定・予測エンジンに

24 準同型暗号については、四方［2019］を参照されたい。

入力される不正なデータを検知・排除するものと、不正なデータによる判定・予測エンジンへの影響を軽減・解消するものに大別することができる。これらの対策は、概ね、訓練データと入力データの両方に共通している。

不正な入力データによる攻撃への対策手法については、Carlini and Wagner [2017a] において網羅的に検討されている。検討の対象として、ニューラル・ネットワークを用いた画像認識のモデルへの適用が想定される代表的な対策手法が挙げられている。これらのうち、不正な入力データを検知・排除する手法が 8 件となっており、①ニューラル・ネットワークを利用するもの（3 件）、②主成分分析を利用するもの（3 件）、③入力データの分布の差異を利用するもの（2 件）に分類される（図表 9 を参照）。

各手法の有効性の評価においては、攻撃者の能力として、次の 3 種類が想定されている。すなわち、①攻撃者が対策手法にかかる情報を一切有していない（ゼロ知識攻撃〈zero knowledge adversary〉）、②攻撃者が、対策の存在を知っているものの、そのパラメータや対策手法のモデルの入出力を入手することができない（限定知識攻撃〈restricted knowledge adversary〉）、③攻撃者が対策手法のパラメータやその入出力を入手することができる（完全知識攻撃〈perfect knowledge adversary〉）場合である。完全知識攻撃は、いわゆるホワイト・ボックスの攻撃に類似したものといえる。

各対策手法の有効性は、最新の攻撃手法（Carlini and Wagner [2017b] で提案されているもの）を、各攻撃者の能力に応じて、各対策手法が実装された判定・予測エンジンに適用することで評価されている。評価結果は、攻撃がどの程度軽減されるかによって示されている。こうした評価のもとでは、偽陽性率を小さく抑えて実用性を確保すると同時に、適合率の向上と、入力データ間のユークリッド距離の拡張を実現する対策手法が高い評価を得ることになる。

評価の結果、多くの対策手法が既存研究で示されている攻撃手法に対して十分に効果を発揮しているとは言い難い状況であることが判明している[25]。完全知識攻撃の場合、いずれの対策手法も不正な入力データを十分に検知することができなかったほか、ゼロ知識攻撃の場合も、一部の手法（入力データの正規化）を除き、高い適合率と低い偽陽性率の両立が困難であるという結果が示された。

25 有効性が低い複数の対策を組み合わせるという方法（アンサンブル）も考えられるが、そうした場合でも有効性は高まらないとする研究報告が知られている（He *et al.* [2017]）。

図表 9　不正な入力データを利用する攻撃への主な対策と評価結果の概要

対策方針（画像認証のモデルの場合）		対策手法の概要	各対策手法の有効性評価（概要）
不正な入力データの検知・排除	ニューラル・ネットワークの利用	不正な入力データを生成し、それを検知するための判定・予測エンジンを別途生成（Grosse *et al.* [2017]、Gong, Wang, and Ku [2017]）。	対策手法にかかる情報を用いることなく訓練データから不正な入力データを生成（ゼロ知識攻撃）。適合率が約 70%、偽陽性率が約 40%（Grosse *et al.* [2017]）。
		学習途中の処理データから、不正な入力データを検知する判定・予測エンジンを別途生成（Metzen *et al.* [2017]）。	ゼロ知識攻撃による入力データによって、適合率が約 80%、偽陽性率が約 30%。
	主成分分析の利用	入力データや学習途中の処理データから主成分を抽出。各成分の重み等を検知に利用（Hendrycks and Gimpel [2017]、Li and Li [2017]）。	ゼロ知識攻撃によって、適合率が約 60%、偽陽性率が約 40%（Li and Li [2017]）。
		画像データの主成分を抽出し、不正な入力データを検知する判定エンジンを生成（Bhagoji, Cullina, and Mittal [2017]）。	完全知識攻撃による入力データでは、入力データ間のユークリッド距離は拡張されず。
	入力データの分布の差異の利用	最大平均差異（maximum mean discrepancy）を利用（Grosse *et al.* [2017]）。	ゼロ知識攻撃による入力データと正規の入力データの間に、分布の有意な差異はみられず（検知困難）。
		隠れ層の出力の尤度を算出。しきい値以下の場合、不正な入力データと判定（Feinman *et al.* [2017]）。	一部のデータセットによる評価では、ゼロ知識攻撃によって、適合率が 20%以下。
判定・予測エンジン等への影響を軽減	入力データの正規化	ドロップアウトを適用。判定・予測結果の分散の和がしきい値以上の場合、不正な入力データと判定（Feinman *et al.* [2017]）。	ゼロ知識攻撃によって、適合率が 75%以上。限定知識攻撃と完全知識攻撃の場合、適合率がそれぞれ約 10%、約 2%。
		画像データに平均値フィルターを適用し入力データとする（Li and Li [2017]）。	ゼロ知識攻撃によって、適合率が約 80%。完全知識攻撃では、入力データのユークリッド距離は拡張されず。

備考：Carlini and Wagner [2017a] の内容を基に作成。

5.　結びに代えて：機械学習システムを活用する際の留意点

4 節で示したとおり、機械学習システムに対する主な攻撃手法への対策としてさまざまなものが提案されているが、現時点では、十分な有効性が確認された対策手法はほとんど存在していない。また、有効性を評価するにあたり、いくつかの定量的な評価尺度が提案・使用されているものの、どの尺度を使用するかは研究論文に

より異なっているなど、複数の対策手法の評価結果を横並びで比較することも容易ではない。これらの点については、既にいくつかの研究論文で指摘され課題として認識されており、今後の研究の進展が期待される。

こうした状況を踏まえると、機械学習システムを今後活用するうえでユーザーが留意すべき事項として、以下の 3 つが挙げられる。

第 1 に、機械学習システムの脆弱性や攻撃手法について、最新の研究動向を随時フォローし把握しておくことが必要である。最近の AI や機械学習への注目度の高まりを受けて、これらの分野の研究論文の発表数は増加傾向にある。そうしたなか、脆弱性や攻撃手法を指摘する研究論文も今後増えていく可能性が高い。機械学習システムを利用する側、あるいは、それを利用して自社の顧客にサービスを提供する側としては、最新の脅威や攻撃手法をフォローし、機械学習システムの利用や顧客へのサービス提供にどのような影響が及ぶ可能性があるかを確認していくことが求められる。

第 2 に、既存研究で提案されている機械学習システムの脆弱性や攻撃手法が、金融分野での機械学習システムの利用形態においてどの程度当てはまるかを明らかにしていくことが求められる。本稿では、最近の主な研究成果を紹介したが、それらで指摘されている攻撃手法は、画像・文字・音声認識に機械学習を適用する分野に焦点を当てたものが多かった。特に、大量のデータを利用する（深層学習等の）学習モデルを対象としつつ、「人間にとっては同一の画像のようにみえるが、判定・予測エンジンは異なる画像と判断する」といった人間の知覚感度の限界による脆弱性を利用したものが目立つ。こうした脆弱性が金融分野での機械学習システムのアプリケーションにどの程度当てはまるかは定かでない。既存研究における脆弱性が金融分野での機械学習システムにどの程度当てはまるか検討していくことが重要である。

また、既存研究における攻撃手法では、攻撃者が学習モデルや判定・予測エンジンの内容に関する情報を利用可能であるという状況を前提としているものが大半である。こうした前提条件が金融分野における機械学習システムの利用環境において成立するか否かを個別に評価し、何らかの対策が必要か否かを評価していくことが求められる。そのうえで、対策が必要であることが判明したものに関して、どのように対処すべきかを他のエンティティと協議しつつ決定していくことになる。

第 3 に、機械学習システムのセキュリティや対策手法の有効性にかかる評価手法を検討・確立していくことが重要である。機械学習システムにおけるセキュリティや対策手法の有効性にかかる評価手法は、研究途上の段階にあり、学界でも重要な課題として認識されている。今後、評価手法にかかる研究成果をフォローし、それらをどのように活用するか検討することが求められる。

クラウドのように、外部の機械学習システムをネットワーク経由で利用する形態

の場合も、こうした評価手法が重要となる。システムを運営する外部事業者が対策手法の実装やセキュリティ管理を適切に実施していることがセキュリティ確保の前提条件となることから、ユーザーとしては、外部事業者における対応の適切性をいかに確認・確保するかについても検討する必要があると考えられる。

4節（2）で説明したように、準同型暗号等を用いて、データを暗号化したまま秘密に学習や予測・判定を行う手法の研究が活発化している。こうした先端的な研究開発によって、訓練データ等を秘密にしたまま学習を行うことができるようになれば、クラウドを運営する外部事業者のセキュリティ管理への要求レベルを低下させたとしても、安全な訓練や判定・予測が実現する可能性がある。こうした研究開発の動向にも注目していく必要があろう。

参考文献

飯島　涼・南　翔汰・シュウインゴウ・及川靖広・森　達哉、「指向性スピーカを用いた音声認識装置への攻撃と評価」、『2018 年暗号と情報セキュリティシンポジウム予稿集』、電子情報通信学会、2018 年

小澤誠一、「機械学習によるサイバーセキュリティとプライバシー保護データマイニングへの取組み」、『NICT サイバーセキュリティシンポジウム 2018 講演資料集』、情報通信研究機構、2018 年

金融情報システムセンター、『平成 30 年版　金融情報システム白書』、財経詳報社、2017 年

――――、『金融機関等コンピュータシステムの安全対策基準・解説書（第 9 版）』、金融情報システムセンター、2018 年

四方順司、「量子コンピュータの脅威を考慮した高機能暗号：格子問題に基づく準同型暗号とその応用」、『金融研究』第 38 巻第 1 号、日本銀行金融研究所、2019 年、73～96 頁（本号所収）

人工知能学会、『人工知能学大辞典』、共立出版、2017 年

先崎佑弥・大畑幸矢・松浦幹太、「深層学習に対する効率的な Adversarial Examples 生成によるブラックボックス攻撃とその対策」、『2018 年暗号と情報セキュリティシンポジウム予稿集』、電子情報通信学会、2018 年

中林紀彦、「“AI を過大評価しない”導入成功の近道」、『日経 FinTech Monthly Newsletter』No. 22、日経 BP 社、2018 年

日本銀行金融機構局、「サイバーセキュリティに関する金融機関の取り組みと改善に向けたポイント―アンケート（2017 年 4 月）調査結果―」、『金融システムレポート別冊シリーズ』、日本銀行金融機構局、2017 年

AI ネットワーク社会推進会議、『報告書 2017―AI ネットワーク化に関する国際的な議論の推進に向けて―』、総務省、2017 年

Ateniese, Giuseppe, Luigi V. Mancini, Angelo Spognardi, Autonio Villani, Domenico Vitali, and Giovanni Felici, “Hacking Smart Machines with Smarter Ones: How to Extract Meaningful Data from Machine Learning Classifiers,” *International Journal of Security and Networks*, 10(3), Inderscience Publishers, 2015.

Barreno, Marco, Blaine Nelson, Anthony D. Joseph, and J. Doug Tygar, “The Security of Machine Learning,” *Machine Learning*, 81(2), Springer-Verlag, 2010, pp. 121–148.

Bhagoji, Arjun Nitin, Daniel Cullina, and Prateek Mittal, “Dimensionality Reduction as a Defense against Evasion Attacks on Machine Learning Classifiers,” arXiv: 1704.02654v2, Cornell University Library, 2017.

Biggio, Battista, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Šrndić, Pavel Laskov, Giorgio Giacinto, and Fabio Roli, “Evasion Attacks against Machine Learn-

ing at Test Time," *Proceedings of Joint European Conference on Machine Learning and Knowledge Discovery in Databases 2013 Part 3*, *Lecture Notes in Computer Science*, 8190, Springer-Verlag, 2013, pp. 387–402.

———, Blaine Nelson, and Pavel Laskov, "Support Vector Machines under Adversarial Label Noise," *Proceedings of Asian Conference on Machine Learning 2011*, *Proceedings of Machine Learning Research*, 20, Journal of Machine Learning Research, 2011, pp. 97–112.

———, ———, and ———, "Poisoning Attacks against Support Vector Machines," *Proceedings of International Conference on Machine Learning (ICML) 2012*, Omnipress, 2012, pp. 1467–1474.

Brundage, Miles, Shahar Avin, Jack Clark, Helen Toner, Peter Eckersley, Ben Garfinkel, Allan Dafoe, Paul Scharre, Thomas Zeitzoff, Bobby Filar, Hyrum Anderson, Heather Roff, Gregory C. Allen, Jacob Steinhardt, Carrick Flynn, Seán Ó hÉigeartaigh, Simon Beard, Haydn Belfield, Sebastian Farquhar, Clare Lyle, Rebecca Crootof, Owain Evans, Michael Page, Joanna Bryson, Roman Yampolskiy, and Dario Amodei, "The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation," arXiv: 1802.07228v1, Cornell University Library, 2018.

Carlini, Nicholas, and David Wagner, "Adversarial Examples Are Not Easily Detected: Bypassing Ten Detection Methods," *Proceedings of ACM Workshop on Artificial Intelligence and Security (AISec) 2017*, Association for Computing Machinery, 2017a, pp. 3–14.

———, and ———, "Towards Evaluating the Robustness of Neural Networks," *Proceedings of IEEE Symposium on Security and Privacy (SP) 2017*, IEEE, 2017b, pp. 39–57.

Chio, Clarence, and David Freeman, *Machine Learning and Security*, O'Reilly Media, 2018.

Dowlin, Nathan, Ran Gilad-Bachrach, Kim Laine, Kristin Lauter, Michael Naehrig, and John Wernsing, "CryptoNets: Applying Neural Networks to Encrypted Data with High Throughput and Accuracy," *Proceedings of International Conference on Machine Learning (ICML) 2016*, *Proceedings of Machine Learning Research*, 48, Journal of Machine Learning Research, 2016, pp. 201–210.

Feinman, Reuben, Ryan R. Curtin, Saurabh Shintre, and Andrew B. Gardner, "Detecting Adversarial Samples from Artifacts," arXiv: 1703.00410v3, Cornell University Library, 2017.

Financial Stability Board, "Artificial Intelligence and Machine Learning in Financial Services: Market Developments and Financial Stability Implications," Financial Stability

Board, 2017.

Fredrikson, Matthew, Somesh Jha, and Thomas Ristenpart, "Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures," *Proceedings of ACM SIGSAC Conference on Computer and Communications Security (CCS) 2015*, Association for Computing Machinery, 2015, pp. 1322–1333.

———, Eric Lantz, Somesh Jha, Simon Lin, David Page, and Thomas Ristenpart, "Privacy in Pharmacogenetics: An End-to-End Case Study of Personalized Warfarin Dosing," *Proceedings of USENIX Security Symposium 2014*, Advanced Computing Systems Association, 2014, pp. 17–32.

Gardiner, Joseph, and Shishir Nagaraja, "On the Security of Machine Learning in Malware C&C Detection: A Survey," *ACM Computing Surveys*, 49(3), Article No. 59, Association for Computing Machinery, 2016.

Gong, Zhitao, Wenlu Wang, and Wei-Shinn Ku, "Adversarial and Clean Data Are Not Twins," arXiv: 1704.04960v1, Cornell University Library, 2017.

Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy, "Explaining and Harnessing Adversarial Examples," Conference Paper at International Conference on Learning Representations (ICLR) 2015, arXiv: 1412.6572v3, Cornell University Library, 2015.

Grosse, Kathrin, Praveen Manoharan, Nicolas Papernot, Michael Backes, and Patrick McDaniel, "On the (Statistical) Detection of Adversarial Examples," arXiv: 1702.06280v2, Cornell University Library, 2017.

He, Warren, James Wei, Xinyun Chen, Nicolas Carlini, and Dawn Song, "Adversarial Example Defenses: Ensembles of Weak Defenses Are Not Strong," *Proceedings of USENIX Workshop on Offensive Technologies (WOOT) 2017*, Advanced Computing Systems Association, 2017.

Hendrycks, Dan, and Kevin Gimpel, "Early Methods for Detecting Adversarial Images," Workshop Contribution at International Conference on Learning Representation (ICLR) 2017, arXiv: 1608.00530v2, Cornell University Library, 2017.

Huang, Sandy, Nicolas Papernot, Ian Goodfellow, Yan Duan, and Pieter Abbeel, "Adversarial Attacks on Neural Network Policies," arXiv: 1702.02284v1, Cornell University Library, 2017.

Kenway, Richard, "Vulnerability of Deep Learning," arXiv: 1803.06111v1, Cornell University Library, 2018.

Kloft, Marius, and Pavel Laskov, "Online Anomaly Detection under Adversarial Impact," *Proceedings of International Conference on Artificial Intelligence and Statistics (AISTATS) 2010*, *Proceedings of Machine Learning Research*, 9, Journal of Machine Learning Research, 2010, pp. 405–412.

Li, Xin, and Fuxin Li, "Adversarial Examples Detection in Deep Networks with Convolutional Filter Statistics," *Proceedings of IEEE International Conference on Computer Vision (ICCV) 2017*, IEEE, 2017, pp. 5775–5783.

Mei, Shike, and Xiaojin Zhu, "Using Machine Teaching to Identify Optimal Training-Set Attacks on Machine Learners," *Proceedings of AAAI Conference on Artificial Intelligence 2015*, Association for the Advancement of Artificial Intelligence, 2015, pp. 2871–2877.

Metzen, Jan Hendrik, Tim Genewein, Volker Fischer, and Bastian Bischoff, "On Detecting Adversarial Perturbations," Conference Paper at International Conference on Learning Representations (ICLR) 2017, arXiv: 1702.04267, Cornell University Library, 2017.

Nguyen, Anh, Jason Yosinski, and Jeff Clune, "Deep Neural Networks Are Easily Fooled: High Confidence Predictions for Unrecognizable Images," *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2015*, IEEE, 2015, pp. 427–436.

Papernot, Nicolas, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z. Berkay Celik, and Ananthram Swami, "Practical Black-Box Attacks against Machine Learning," *Proceedings of ACM on Asia Conference on Computer and Communications Security (ASIA CCS) 2017*, Association for Computing Machinery, 2017, pp. 506–519.

———, ———, Somesh Jha, Matthew Fredrikson, Z. Berkay Celik, and Ananthram Swami, "The Limitations of Deep Learning in Adversarial Settings," *Proceedings of IEEE European Symposium on Security and Privacy (EURO S&P) 2016*, IEEE, 2016a, pp. 372–387.

———, ———, Arunesh Sinha, and Michael Wellman, "Towards the Science of Security and Privacy in Machine Learning," arXiv: 1611.03814v1, Cornell University Library, 2016b.

Phong, Le Trieu, "Privacy-Preserving Stochastic Gradient Descent with Multiple Distributed Trainers," *Proceedings of International Conference on Network and System Security (NSS) 2017*, *Lecture Notes in Computer Science*, 10394, Springer-Verlag, 2017, pp. 510–518.

———, Yoshinori Aono, Takuya Hayashi, Lihua Wang, and Shiho Moriai, "Privacy-Preserving Deep Learning via Additively Homomorphic Encryption," *IEEE Transactions on Information Forensics and Security*, 13(5), IEEE, 2018, pp. 1333–1345.

Shokri, Reza, Marco Stronati, Congzheng Song, and Vitaly Shmatikov, "Membership Inference Attacks against Machine Learning Models," *Proceedings of IEEE Symposium on Security and Privacy (SP) 2017*, IEEE, 2017, pp. 3–18.

Sinha, Arunesh, Debarun Kar, and Milind Tambe, "Learning Adversary Behavior in Se-

curity Games: a PAC Model Perspective," *Proceedings of International Conference on Autonomous Agents and Multiagent Systems (AAMAS) 2016*, International Foundation for Autonomous Agents and Multiagent Systems, 2016, pp. 214–222.

Suciu, Octavian, Radu Mărginean, Yiğitcan Kaya, Hal Daumé III, and Tudor Dumitraş, "When Does Machine Learning FAIL? Generalized Transferability for Evasion and Poisoning Attacks," *Proceedings of USENIX Security Symposium 2018*, Advanced Computing Systems Association, 2018, pp. 1299–1316.

Sze, Vivienne, Yu-Hsin Chen, Tien-Ju Yang, and Joel S. Emer, "Efficient Processing of Deep Neural Networks: A Tutorial and Survey," *Proceedings of the IEEE*, 105(12), IEEE, 2017, pp. 2295–2329.

Szegedy, Christian, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus, "Intriguing Properties of Neural Networks," Proceedings of International Conference on Learning Representations (ICLR) 2014, arXiv: 1312.6199v4, Cornell University Library, 2014.

Tramèr, Florian, Fan Zhang, Ari Juels, Michael K. Reiter, and Thomas Ristenpart, "Stealing Machine Learning Models via Prediction APIs," *Proceedings of USENIX Security Symposium*, Advanced Computing Systems Association, 2016, pp. 601–618.