IMES DISCUSSION PAPER SERIES

金融分野で活用される機械学習システムのセキュリティ分析

井上紫織（いのうえ しおり）・宇根正志（うね まさし）

Discussion Paper No. 2019-J-1

IMES

INSTITUTE FOR MONETARY AND ECONOMIC STUDIES

BANK OF JAPAN

日本銀行金融研究所

〒103-8660 東京都中央区日本橋本石町 2-1-1

IMES Discussion Paper Series 2019-J-1
2019 年 1 月

金融分野で活用される機械学習システムのセキュリティ分析

井上紫織（いのうえ しおり）*・宇根正志（うね まさし）**

要　　旨

近年、金融業界において、預金為替業務、融資業務、投資運用業務、保険業務をはじめとする、さまざまな領域で、人工知能、とりわけ機械学習システムの活用にかかる検討が進んでいる。機械学習システムには、情報システム一般に存在する脆弱性に加え、特有の脆弱性も存在する。機械学習システムを安全かつ安定的に利用していくためには、こうした脆弱性を悪用する攻撃について、予め対策を十分に検討しておくことが肝要である。本稿では、機械学習システムを金融分野で活用する際に想定される、各機能や役割の担い手の構成を分類、整理したうえで、各構成における脅威や対策等を分析するとともに、金融機関にとっての留意点や課題を明らかにする。

キーワード：機械学習、人工知能、脆弱性、セキュリティ
JEL classification: L86、L96、Z00

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

本稿の作成に当たっては、筑波大学の佐久間淳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、筆者たち個人に属し、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たち個人に属する。

目　次

１．はじめに

近年、人工知能（artificial intelligence：AI）の実社会での活用にかかる検討が急速に進んでいる。こうした動きは、金融業界も例外ではなく、預金為替業務、融資業務、投資運用業務、保険業務をはじめとする、さまざまな領域でみられるようになってきている。コールセンターの自動応答を実現するチャットボットや、株式運用におけるマーケット予測や融資業務における融資先の業績予測で知られるように、AI の活用は、事務の効率や精度の向上に資するほか、新たなサービス提供による収益向上、経営リスクの低減等に寄与することも期待される。もっとも、こうした新たな技術を導入する際には、そのメリットだけでなく、セキュリティ面のリスクに対しても十分に目を向ける必要がある。

AI は、一般に、推論や認識、判断等、人間と同様の知的な処理能力を持つコンピュータ・システムやその技術分野を指し、その機能を実現するツールとして用いられる技術が機械学習である。機械学習を実装したシステム（以下、機械学習システム）では、大量のサービス要求による機能低下等、情報システム一般に存在する脆弱性に加え、特有の脆弱性も存在する（宇根［2018］、吉岡［2018］）。こうした脆弱性が悪用されると、機械学習システムにおいて処理されるデータや学習モデル、判定・予測エンジンが、盗取されたり改変されたりする可能性がある。機械学習システムを安全かつ安定的に利用していくためには、これらの攻撃への対応策を予め十分に検討することが肝要である。

本稿では、機械学習システムの各機能や役割の担い手（エンティティ）の構成を 12 のタイプ（以下、構成タイプ）に分類し、それぞれにおいて想定される脆弱性と攻撃、その対応策を整理する。続いて、金融機関等で活用されている機械学習システムを例にとり、構成タイプ別に分類するとともに、想定される具体的な攻撃とその対応方針について考察したうえで、今後の課題を示す。

２．機械学習システムの構成とその分類

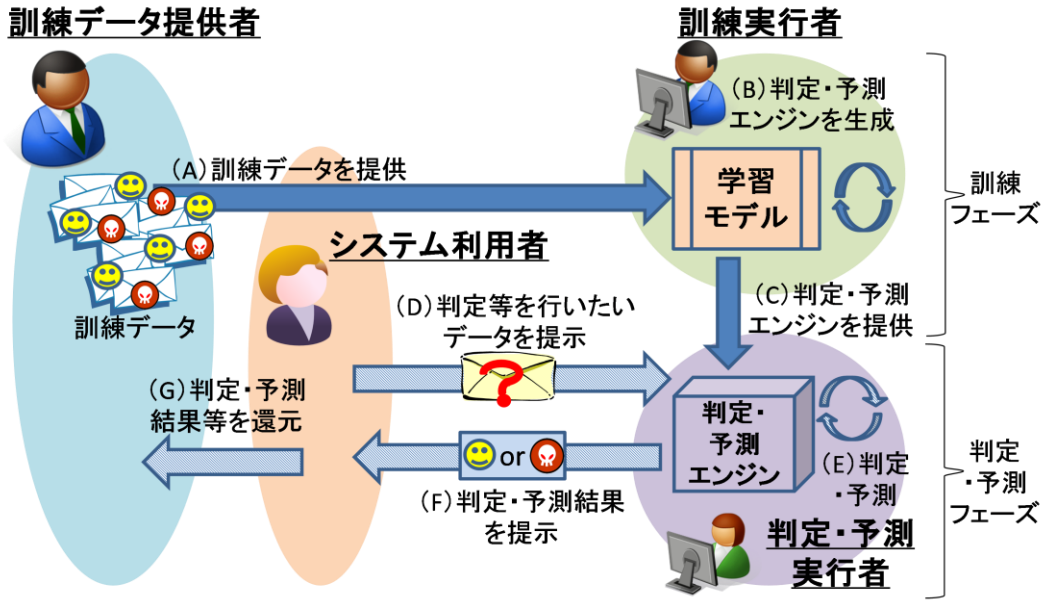
（１）エンティティ

宇根［2018］に基づき、次の 4 つのエンティティによって構成される機械学習システムを想定する[1]。すなわち、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・予測エンジンの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。判定・予測エンジンの生成と判定・予測における処理の流れは次のとおりである（図表 1 を参照）[2]。

[1] 本稿では、教師あり学習のシステムを対象とする。

[2] ここでの学習モデルや判定・予測エンジンは、学習アルゴリズム、判定・予測モデルとそ

図表 1. 想定する機械学習システムの構成（概念図）



【判定・予測エンジンの生成】

（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後、システム利用者と協力しつつ、これらのデータを適宜加工するとともに、ラベル（当該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[3]。

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

（C） 訓練実行者は判定・予測エンジンを判定・予測実行者に提供。

【判定・予測】

（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。

（E） 判定・予測実行者は、上記（D）でシステム利用者から受信したデータを判定・予測エンジンに適用し、判定・予測を実施。

（F） 判定・予測実行者は判定・予測結果をシステム利用者に提示。

（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[4]。

れぞれ呼ばれる場合もある。本稿では、宇根［2018］の用語を用いて議論することとする。

[3] 上記（A）における訓練データ等の提供では、訓練データが機微な情報を含む場合にはマスキング等を実施する必要があるほか、暗号化するケースが考えられる。ここでは、分析を単純化するために、こうした処理が完了したものが訓練実行者に送信されるものとする。

[4] 例えば、判定・予測結果が誤っていることが判明した場合、その判定・予測エンジンへの

（２）機械学習システムの構成の分類

機械学習システムでは、本節（1）における 4 つのエンティティの機能を単一あるいは複数の主体が担うことになる。例えば、4 つのエンティティをすべて 1 つの主体が担うケースとしては、金融機関が自社内のデータのみを訓練データとして採用し、それを自社で有する学習モデルに適用して判定・予測エンジンを生成し使用する場合が考えられる。一方、システム利用者と判定・予測実行者が異なるケースとしては、クラウド等の外部事業者が判定・予測エンジンを実行する場合が挙げられる[5]。さらに、訓練実行者と判定・予測実行者が異なるケースとしては、訓練データを受け取った訓練実行者が判定・予測エンジンを生成し、判定・予測のサービスを実施したい別の主体（判定・予測実行者）にそれを提供する場合がありうる。

上記のような検討を通じて、本節（1）における 4 つのエンティティの役割を担う主体の組合せに基づき、機械学習システムの構成のバリエーションを網羅的に整理すると 15 の構成タイプに分類することができる（各構成タイプの説明は補論を参照）[6]。ただし、実際に機械学習システムを活用する場面を考慮すると、システム利用者と訓練実行者を同一の主体が担い、その主体とは異なる主体が判定・予測実行者の役割を担う 3 つの構成タイプは想定しづらい。すなわち、これらの構成タイプでは、システム利用者（訓練実行者）は、判定・予測エンジンを生成するものの、判定・予測の実行は別の主体が担うため、判定・予測時にその主体と通信するなどの追加的な処理を要し、効率的な構成とはいえない。そこで、本稿では、上記の 3 つを除く 12 の構成タイプに焦点を当てて検討を進めることとする（図表 2 を参照）。

入力データに正しい判定結果のラベルを付加し、それを訓練データとして用いて再度訓練を実行することで、判定・予測エンジンの精度改善を図ることが考えられる。

[5] クラウド上で機械学習システムを実行するサービス（Machine Learning-as-a-Service＜MLaaS＞と呼ばれる）が提供されている。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）等が挙げられる。また、複数の金融機関から訓練データ（顧客からの問合せやその回答）を収集し、それらに基づいて生成したチャットボットを用いて、顧客からの問合せに対する自動応答サービスを提供する事例も知られている（NTT データ［2017］）。

[6] なお、訓練データ提供者が訓練データを訓練実行者に提供しつつその処理の一部を自ら実行する場合等も想定される（Phong [2017]）が、このようなケースに関しては、学習モデルの出力となる判定・予測エンジンを最終的に生成する主体のみを訓練実行者とみなす。

図表 2. 機械学習システムの構成タイプの分類

構成タイプ	各エンティティを担う主体				主体数
	訓練データ提供者	システム利用者	訓練実行者	判定・予測実行者	
1	▲	○	■	◇	4
2	▲	▲	■	◇	3
3	▲	○	▲	◇	
4	▲	○	■	▲	
5	▲	○	■	○	
6	▲	○	■	■	
7	▲	▲	□	□	2
8	▲	○	▲	○	
9	▲	○	○	○	
10	▲	○	▲	▲	
11	▲	▲	■	▲	
12	▲	▲	▲	▲	1

備考：1. 各エンティティの役割を担う主体を▲、○、■、◇等の記号で表示。
2. 有色（白以外）かつ同色のセルのエンティティは同一の主体が担う。

３．機械学習システムの各構成タイプにおける攻撃と対応策

本節では、機械学習システムのセキュリティ目標、各エンティティの行動、攻撃者の能力について説明する。そのうえで、2 節で示した各構成タイプにおける攻撃と対応策を検討する。

（１）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、一般的な情報システムと同様、取り扱われるデータやシステムの機能の機密性（confidentiality）・完全性（integrity）・可用性（availability）の確保が求められる（Barreno *et al.* [2010]、Papernot *et al.* [2016a]）[7]。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、それらのデータやシステムの機能が不正に偽造・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動することを意味する。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・予測エンジンへの入力データ（判定・予測用データ）、⑤判定・予測用データに対応する判定・予測エンジンの出力データ、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。

[7] Papernot *et al.* [2016a]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。

例えば、機密性の観点からは、訓練データに訓練データ提供者にかかる機微な情報（個人情報等）が含まれている場合、そ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判定・予測エンジンの生成（機能の改変）が判定・予測に大きな影響を与える場合、それらを防ぐ必要がある。可用性の観点からは、大量の訓練データを訓練実行者に送信し、それを受信するシステムの機能度を低下させて業務を妨害するという攻撃を防ぐ必要がある。

（２）各エンティティの行動と攻撃者の能力

本稿における分析では、攻撃者は、機械学習システムの第三者であり、本節(1)で示した保護対象となりうるデータやシステムの機能に対して攻撃を試みるものとする。攻撃を受ける箇所としては、各エンティティとそれらの間の通信路が想定される。それぞれのセキュリティ対策の前提は以下のとおりとする。

まず、訓練データ提供者は、外部からの不正アクセス等による攻撃への対応策を講じているものの、高度なサイバー攻撃を受ける、あるいは、内部者の一部が不正行為を行うなどによって、攻撃者が訓練データ提供者のデータやシステムの機能を盗取・改変する場合があるとする[8]。システム利用者についても、訓練データ提供者と同様の想定を置く。

一方、訓練実行者と判定・予測実行者に関しては、それらが保有するデータやシステムの機能が盗取・改変されると機械学習システムとして使用できなくなるため、訓練データ提供者やシステム利用者に比べて、より高度なセキュリティ対策を講じているものとする。こうした対策の結果、攻撃者は、訓練実行者や判定・予測実行者から、訓練データ、学習モデル、判定・予測エンジン、判定・予測エンジンの入出力を入手することができないとする。ただし、訓練実行者や判定・予測実行者を担う主体が訓練データ提供者あるいはシステム利用者も担う場合には、訓練データ提供者あるいはシステム利用者は、訓練実行者や判定・予測実行者と同様の高度なセキュリティ対策を講じていると想定する。

なお、訓練実行者と判定・予測実行者が講じるセキュリティ対策としては、外部からの不正アクセス等への対策を強化することや、仮にデータが盗取されたとしても、実害が生じないようにしておくことが求められる。例えば、訓練実行者は、訓練データの盗取に備えて、①訓練データを暗号化したまま学習モデルを実行する「準同型暗号等を用いた機械学習」の手法を採用すること

[8] 訓練データ提供者が、意図せず、訓練データとして不適切なデータを選択する場合も考えられる。訓練データの選択・生成は機械学習システムの品質を担保するうえで極めて重要であることから、ここでは、訓練データ提供者が、事前に定められた一定の手順に従って、訓練データの選択・生成を適切に行うものとし、上記のようなケースは検討対象外とする。

（Dowlin *et al.* [2016]、Phong *et al.* [2018]）、②訓練データから個人や組織が識別・特定されないように、訓練データを適切に加工すること（パーソナル・データの保護）などの対策を講じることが考えられる[9]。

各エンティティ間の通信路については、TLS（Transport Layer Security）等の暗号通信プロトコルによって保護され、攻撃者はそのデータを盗取・改変することが困難であると想定する。ただし、攻撃者は、訓練データ提供者やシステム利用者から暗号鍵等を入手することができれば、通信路上のデータを盗聴・改変することができるとする。

（３）攻撃と対応策

本節（2）のとおり、訓練実行者と判定・予測実行者は、サイバー攻撃への対策を十分に講じていると想定し、攻撃者が訓練データ提供者やシステム利用者のデータを悪用するケースについて、想定される攻撃と対応策を整理する（図表 3 を参照）。

イ．攻撃者が訓練データ提供者のデータを悪用する場合

攻撃者は、訓練データを入手し、それを用いて個人・組織に関する秘密の情報を取得する（機密性への攻撃）。また、訓練データを改変する（完全性への攻撃）ことによって、不正な判定・予測エンジンの生成を試行する（Biggio, Nelson, and Laskov [2011, 2012]、Biggio *et al.* [2013]、Barreno *et al.* [2010]、Goodfellow, McDaniel, and Papernot [2018]）。さらに、訓練データを大量に訓練実行者に送信し、訓練実行者の業務を妨害する（可用性への攻撃）。

訓練データの盗取への対策としては、盗取された場合の影響を軽減する観点から、個人・組織の識別・特定につながるデータ等、機密性が求められるデータを訓練データとして利用しない、また、利用する場合には、個人・組織の特定等が困難なようにデータを加工するといった対応が考えられる。訓練データの改変による不正な判定・予測エンジンの生成に対しては、学習モデルに入力する前に、不正な訓練データを検知・排除する、または、それらによる判定・予測エンジンへの影響を軽減する学習モデルを利用することが挙げられる

[9] このほか、訓練実行者や判定・予測実行者の機能を複数のエンティティに分散し、それらの一部がサイバー攻撃を受けたとしても、訓練データや学習モデルが攻撃者の手に渡らないようにする「秘密分散に基づくマルチパーティ計算による秘匿機械学習」の手法も対策の候補となる（Mohassel and Zhang [2017]、Mohassel and Rindal [2018]）。

図表 3. 想定される攻撃・対応策と該当する機械学習システムの構成タイプ

攻撃者が悪用するデータ		攻撃	対応策	構成タイプ		
				1 2 6 7	3 4 10	5 9
訓練データ提供者のデータ	訓練データ	訓練データを盗取。	個人や組織を識別・特定可能な情報等、機密性を有するデータを訓練データに使用しない（必要な加工を実施）。	○	—	○
		不正な判定・予測エンジンを生成。	不正な訓練データを検知・排除。 不正な訓練データによる判定・予測エンジンへの影響を軽減。	○	—	○
		訓練データを大量に送信し、訓練実行者の業務を妨害。	CDN（Contents Delivery Network）のサービス等によって保護。	○	—	○
システム利用者のデータ	判定・予測用データ 判定・予測エンジンの出力	判定・予測エンジンを推定。	推定に必要な情報（判定・予測の確信度等）を入手させないように運用。	○	○	—
		訓練データにかかる情報を推定。	推定に必要な情報（判定・予測の確信度等）を入手させないように運用。 訓練データの推定が困難な学習モデルを採用。	△	○	—
		不正な判定・予測用データによって、誤った判定・予測を誘発。	不正な判定・予測用データを検知・排除。 不正な判定・予測用データによる判定・予測結果への影響を軽減。	○	○	—
		判定・予測用データを大量に送信し、判定・予測実行者の業務を妨害。	CDN のサービス等によって保護。	○	○	—
	還元データ	不正な還元データを介して不正な判定・予測エンジンを生成。	不正な訓練データを検知・排除。 不正な訓練データによる判定・予測エンジンへの影響を軽減。	△	○	—
		還元データを大量に送信し、訓練データ提供者の業務を妨害。	CDN のサービス等によって保護。	△	○	—

備考：1. 「構成タイプ」の欄の「○」は、その欄の構成タイプに左記の攻撃・対応策が該当することを示す。「△」は、訓練データを利用した攻撃が可能であれば、改めて実行する必要がない攻撃であることを示す。
2. 構成タイプ 8、11、12 はいずれの攻撃・対応策も該当しない。

（Carlini and Wagner [2017]）[10]。訓練実行者に対する大量の訓練データの送信に対しては、訓練実行者による CDN（Contents Delivery Network）等のサービスを

[10] 不正な入力データを検知・排除する方法として、ニューラル・ネットワークを利用する手法、主成分分析を利用する手法、入力データの分布差異を利用する手法等が挙げられる。また、判定・予測エンジンへの影響を軽減する方法としては、入力データを正規化する手法がある。

利用した対策や、外部からのデータの受信を制御するゲートウェイ等による対策が考えられる[11]。

ロ．攻撃者がシステム利用者のデータを悪用する場合

攻撃者は、判定・予測用データとそれに対応する判定・予測エンジンの出力を悪用し、以下の攻撃を行うことが想定される。すなわち、①判定・予測エンジンを推定する（機密性への攻撃、Tramèr *et al.* [2016]）、②訓練データにかかる情報を推定する（機密性への攻撃、Shokri *et al.* [2017]、Ateniese *et al.* [2015]、Fredrikson *et al.* [2014]、Fredrikson, Jha, and Ristenpart [2015]）、③誤った判定・予測を誘発する（完全性への攻撃、Szegedy *et al.* [2014]、Nguyen, Yosinski, and Clune [2015]、Sinha, Kar, and Tambe [2016]、Kenway [2018]、Papernot *et al.* [2017b]）、④判定・予測用データを大量に送信し、判定・予測実行者の業務を妨害する（可用性への攻撃）[12]。

また、攻撃者が還元データを悪用するケースも考えられる。不正な還元データを訓練データ提供者に送信し、訓練データのラベル等を変更したうえで再度訓練を実行させ、その不正な訓練データによって（不正な）判定・予測エンジンを生成させることが考えられる（完全性への攻撃）。また、訓練データ提供者に還元データを大量に送信し、訓練データ提供者の業務を妨害する（可用性への攻撃）ことも考えられる。

判定・予測エンジンや訓練データにかかる情報は、判定・予測結果とともに、その確信度に関する情報が利用できる場合により推定されやすくなる（Tramèr *et al.* [2016]、Fredrikson *et al.* [2014]、Fredrikson, Jha, and Ristenpart [2015]）。そのため、システム利用者に判定・予測の確信度を送信しないように運用することが有効な対策手法になると考えられる[13]。これに加えて、訓練データにかかる情報の推定に対しては、PATE（Private Aggregation of Teacher Ensembles）等、推定を困難にする手法の活用も選択肢となる（Abadi *et al.* [2017]、Papernot *et al.* [2017a]、Goodfellow [2018]）[14]。不正な判定・予測用データによる誤った判定・予測の誘

[11] CDN は、インターネット・ユーザーへのコンテンツ配信を、効率的かつ高速に配信するとともに、大量のアクセスを制御する仕組みのこと。

[12] 判定・予測エンジンの推定は、エンジン自体を盗取するものではないが、無権限の第三者に（類似の）エンジンを知られることとなり、エンジンの盗取と同様の結果となることから、ここでは機密性への攻撃と位置付ける。

[13] 例えば、確信度の値をそのまま提供するのではなく、四捨五入して値を丸めたりすることが考えられる。

[14] PATE は、訓練データを複数の集合に分割したうえで、各集合を訓練データとする判定・予測エンジンを複数生成し、それらのエンジンを集約して最終的な判定・予測エンジンとする手法である。このほか、訓練データが変化したとしても、判定・予測エンジンの出力（の差分）から訓練データの差分にかかる情報の推定を統計的に困難とする手法「差分プ

発への対策としては、判定・予測エンジンに入力する前に、不正な判定・予測用データを検知・排除する、または、それらによる判定・予測エンジンの出力への影響を軽減する学習モデルを利用することが挙げられる。例えば、ある判定・予測エンジンを生成した後、その入出力を再現する判定・予測エンジンを別途生成して最終的なエンジンとする防御的蒸留（defensive distillation、Papernot *et al.* [2016b]）や、誤った判定・予測を引き起こす判定・予測用データを準備して適切なラベルを付け、それらを学習モデルに適用して判定・予測エンジンを生成する敵対的学習（adversarial training）の採用が挙げられる（Szegedy *et al.* [2014]）[15]。

不正な還元データによる不正な判定・予測エンジンの生成に対しては、学習モデルに入力する前にそれを検知・排除する、または、不正な訓練データによる判定・予測エンジンへの影響を軽減する学習モデルを利用することが考えられる。判定・予測用データや還元データをそれぞれ判定・予測実行者や訓練データ提供者に対して大量に送信する攻撃への対策としては、各エンティティがCDN 等のサービスを利用する、あるいは、外部からのデータを受信するゲートウェイで制御するなどが挙げられる。

（４）各構成タイプにおける攻撃と対応策

本節（3）の攻撃と対応策に基づき、機械学習システムの各構成タイプにおいて、どの攻撃を考慮して対応策を講じる必要があるかを示す（図表 3 を参照）。

本節（2）のとおり、訓練実行者や判定・予測実行者を担う主体が訓練データ提供者あるいはシステム利用者も担う場合、それらのエンティティも高度なセキュリティ対策を講じていると想定する。このため、訓練実行者あるいは判定・予測実行者を担う主体が訓練データ提供者およびシステム利用者を担う場合、すなわち、構成タイプ 8、11、12 では、攻撃者はいずれのエンティティのデータも悪用できない。したがって、本節（3）に掲げた攻撃と対応策について考慮する必要はない。

他方、訓練実行者あるいは判定・予測実行者を担う主体が訓練データ提供者を担う場合、すなわち、構成タイプ 3、4、10 では、攻撃者は訓練データ提供者

ライバシー（differential privacy）」も対応策の 1 つと考えることができる（Abadi *et al.* [2016]）。

[15] 蒸留は、生成した判定・予測エンジンの計算量を削減する場合に用いられる圧縮手法の１つ。もとの判定・予測エンジンの入出力を訓練データとして用いて、ニューラル・ネットワークのレイヤー数を少なくしたより軽量な学習モデルで訓練を行うことで、十分な精度を維持しつつ、計算量を軽減した判定・予測エンジンを生成する。蒸留により生成した判定・予測エンジンは、判定・予測エンジンへの入力が多少変化しても出力結果が大きく変化しづらい性質（ロバスト性）が高まることが知られている。防御的蒸留では、蒸留の際に、軽量な学習モデルでなく、もとのモデルを使用する（レイヤー数を削減しない）ことで、ロバスト性をさらに向上させるという手法である。

のデータを悪用できないため、攻撃者がシステム利用者のデータを悪用する場合の攻撃と対応策のみを考慮すればよい。また、訓練実行者あるいは判定・予測実行者を担う主体がシステム利用者を担う場合、すなわち、構成タイプ 5、9 では、攻撃者はシステム利用者のデータを悪用できないため、攻撃者が訓練データ提供者のデータを悪用する場合の攻撃と対応策のみを考慮すればよい。

これに対し、訓練データ提供者とシステム利用者を担う主体が、訓練実行者および判定・予測実行者を担う主体と異なる場合、すなわち、構成タイプ 1、2、6、7 では、攻撃者が訓練データ提供者とシステム利用者の両方のデータを悪用する場合の攻撃と対応策を考慮することが求められる。なお、攻撃者が訓練データを悪用することができれば、システム利用者から得られる情報を用いた攻撃（訓練データにかかる情報の推定、不正な還元データによる不正な判定・予測エンジンの生成）を改めて実行する必要がないほか、還元データを大量に訓練データ提供者に送信する攻撃も不要になると考えられる。

４．機械学習システムを活用するうえでの留意点と課題

本節では、3 節の整理をベースに、機械学習システムに対する攻撃への対策を検討する際に留意すべき事項を考察する。そのうえで、金融分野において機械学習システムを活用することが想定されるケースにおいて、本稿での検討内容がいかに適用されうるかを示す。

（１）実際に想定すべき攻撃の検討

3 節（4）で示したように、機械学習システムのセキュリティ対策を検討する場合には、そのシステムがどの構成タイプに相当するかを明確にしたうえで、前掲の図表 3 を参照しつつ、その構成タイプに想定される攻撃と対応策に焦点を当てる必要がある。図表 3 で示している攻撃は、①訓練データ自体あるいはそれにかかる情報の推定、②判定・予測エンジンにかかる推定、③不正な判定・予測エンジンの生成、④判定・予測エンジンの精度の低下、⑤各エンティティの業務の妨害に集約することができる。これらのうち⑤については、外部のネットワークと接続している情報システムにおいて、通常検討対象となっているものであり、対応策（CDN のサービス等の利用）もよく知られている。したがって、機械学習システムにおいて主に課題となるのは、上記①～④の攻撃のうち、実際に想定すべき攻撃はどれか、そして、これらの攻撃への対応策を実際にどのように行うかである。

まず、想定すべき攻撃を考えるうえでポイントとなるのは、攻撃が成功した場合に、実際にどのような影響や経済的損失が生じうるかである。想定される影響や経済的損失が許容できる場合、特段の対策は不要となる。一方、許容で

きない場合には、影響や経済的損失を許容できるレベルに軽減するための対応策を検討・実施することが求められる。

訓練データ自体あるいはそれにかかる情報の推定については、訓練データが機密性を有するか否かの観点から、流出に伴うリスクを見積もることが求められる。例えば、訓練データとして、金融市場の市況データや金融・経済の統計データ、その他の公表データを利用する場合、訓練データ自体に機密性は認められず、それらが攻撃者の手に渡ったとしても一般的には影響は小さいと考えられる[16]。したがって、こうした攻撃を想定した対応策を特段講じる必要はないと判断するケースもありうる。これに対して、訓練データとして個人の資産や金融取引のデータを利用する場合には、それらがパーソナル・データに該当し、攻撃が成功した場合の影響が大きい可能性があることから、訓練データの保護等、何らかの対応策を講じることが求められる。

判定・予測エンジンにかかる推定については、判定・予測エンジンが外部に流出した場合の経済的損失の多寡が問題となる。例えば、金融市場の動向を分析・予測するためのツールとして機械学習システムが使用されている場合、その判定・予測エンジンは金融機関にとって重要な資産（営業秘密）の 1 つと位置付けられることから、攻撃者による推定を防止するための対応策を検討する必要があるといえる。

また、判定・予測エンジンが攻撃者の手に渡ると、判定・予測エンジン自体の機密性はそれほど高くないとしても、それを手掛りに、訓練データ自体あるいはそれにかかる情報が推定される可能性がある。そのため、判定・予測エンジンにかかる推定への対策方針を検討するうえで、それらの情報が漏洩した場合の影響の有無を検討することも重要となる。例えば、コールセンターにおける「お客様からの問合せへの回答」を機械学習システムによって実施する場合において、質問と回答の関係が特段機密性を有しないケースでは、その機械学習システムの判定・予測エンジンにかかる情報が攻撃者の手に渡ったとしても影響は小さいと考えられる。こうしたケースにおいては、特段の対応策を検討する必要はないと判断する場合もありうる。

不正な判定・予測エンジンの生成については、判定・予測エンジンが攻撃者によって不正なものに改変され、判定・予測結果が不適切であった場合、どのような影響や経済的損失が生じうるかが問題となる。例えば、上記のコールセンターにおける機械学習システムの場合、不適切な判定・予測が生じたとしても、人間による対応に随時切り替えるなどの対応も可能であり、コールセンター

[16] ただし、公表データであっても、どのようなデータを訓練データとして利用しているかという情報が企業秘密となっている場合が想定される。このような場合には、企業秘密に相当する情報を特定し、それを保護する方策を検討することが求められる。

業務や顧客のサービスに大きな影響を与えることはないケースがありうる。このようなケースであれば、特段の対応は不要と判断することができる。一方、金融機関が投資判断に用いる金融市場の先行き予測に機械学習システムを使用する場合には、誤った先行き予測によって資産運用に問題が発生するなどの可能性があることから、資産運用の規模やリスクに応じて対応策の必要性を検討することが求められる。

判定・予測エンジンの精度の低下についても、不正な判定・予測エンジンの生成と同様に、誤った判定・予測結果による影響や経済的損失に応じて検討の要否を判断することが必要である。

（２）金融分野における応用事例と対応策

金融分野では、預金為替業務、融資業務、投資運用業務、保険業務をはじめとする、さまざまな領域で機械学習システムの活用の検討が進んでいる。主な事例を目的別に分類すると、①事務の効率化、②サービス品質の向上、③判断・予測の支援、④リスク低減の 4 つに整理できる。本節では、目的ごとに、金融機関で用いられる機械学習システムの構成例を取り上げ、図表 2 に示したエンティティの構成タイプ別に、想定される攻撃および対応策を概観する。

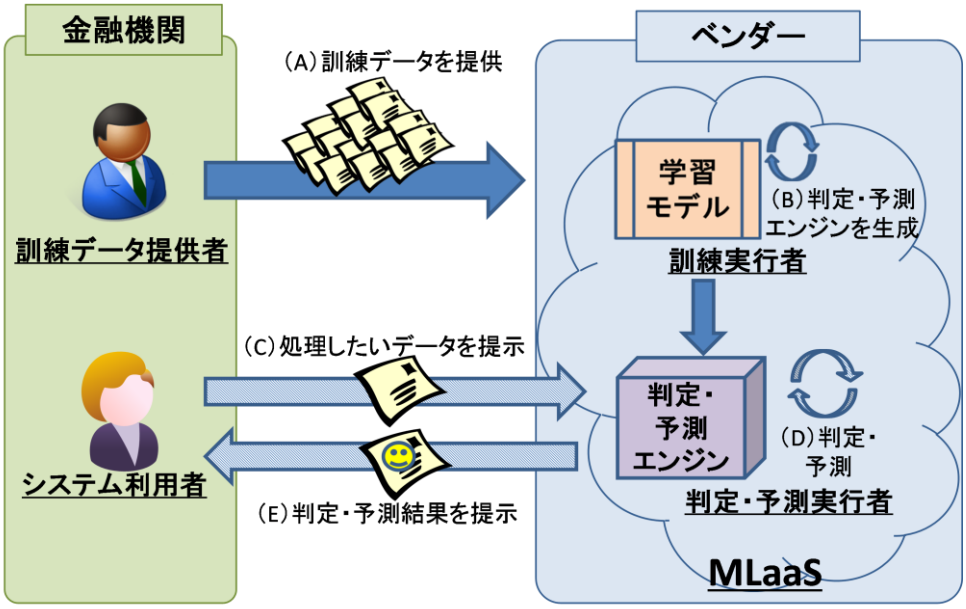
イ．事務の効率化を目的とした機械学習システムの活用

金融機関には、高い専門性を要する業務が多く存在し、過去事例との平仄が重視される業務も多い。例えば、生命保険会社における保険金の支払業務や、銀行における融資関連の契約書作成業務、銀行における振込みや口座振替等で用いられる OCR のデータ処理といった業務が該当する。こうした業務においては、専門知識やノウハウの継承に相応のコストがかかるほか、従来のシステムで代替処理することは困難であった。

一方、機械学習システムは、過去の事例から類似性や規則性を見つけ出すことを得意とするため、こうした事務の一部を代替することが期待される。事務の効率化を目的とした機械学習システムを MLaaS によって構築する場合、このシステムにおける処理の流れは以下のとおりである（図表 4 を参照）。

（A） 金融機関（訓練データ提供者）は、過去に蓄積した事務データ（訓練データ）をベンダー（訓練実行者）に提供する。
（B） ベンダーは、上記（A）で受け取った訓練データを用いてクラウド上の MLaaS で判定・予測エンジンを生成する。
（C） 金融機関（システム利用者）は、新たに処理したいデータをベンダー（判定・予測実行者）に提示する。

図表 4. 事務の効率化を目的とした機械学習システムの構成（概念図）



(D)　ベンダーは、上記（C）で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。

(E)　ベンダーは、判定・予測結果を金融機関に提示する。

このシステムの構成では、訓練データ提供者とシステム利用者は金融機関、訓練実行者と判定・予測実行者はベンダーが担うため、構成タイプ 7 に該当する。

攻撃者は、訓練データ提供者またはシステム利用者のデータを利用することが想定される。まず想定されるのは、訓練データの盗取である。高い専門性や過去事例に関する知識を要する事務では、訓練データとして、過去の顧客データ等の機密性の高いデータが用いられる場合がある。金融機関は、従前より、こうしたデータの管理を厳格に実施してきているが、訓練データとして用いる場合にも、同様の対応が求められる。また、そうした訓練データについて、個人や組織の特定につながる情報を削除するなどの加工を施して、情報漏洩のリスクを低減させることも有用である[17]。

判定・予測エンジンに対する攻撃としては、エンジンの推定、不正なエンジンの生成、エンジンの精度の低下が考えられる。事務の効率化を企図したシステムにおいて、エンジンが推定されたとしても、特段、大きな影響は生じない場合が多いと考えられる。一方、不正なエンジンの生成や精度の低下は、事務の品質の低下につながりうる。不正な訓練データや判定・予測用データを検知・

[17] 訓練データの加工は、判定・予測エンジンの性能に影響を与えることになるため、業務で求められる性能も考慮しつつ加工の方法を検討することになる。

排除するといった技術的な対応に加えて、判定・予測結果の妥当性を人間が確認する運用を検討することが考えられる。

ロ．サービス品質の向上を目的とした機械学習システムの活用

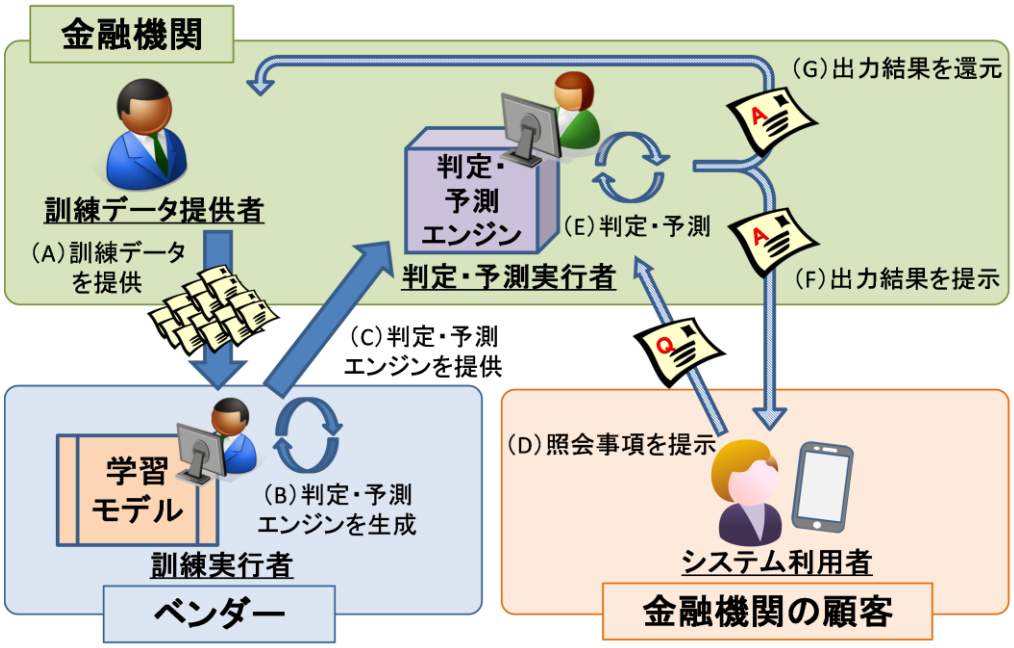
近年、顧客とのコミュニケーションにおけるサービス品質向上の手段として、多くの金融機関でチャットボットを導入する動きがみられる[18]。チャットボットは、コールセンターや SNS、スマートフォン・アプリ、ウェブ上において顧客との対話機能を担い、照会への自動応答や、顧客の状況に合わせた商品提案等を行う。チャットボットを利用すると、均質かつ付加価値のある回答を提示することが可能になると期待されている。

チャットボットに求められる基本的な機能は、予め想定される内容の質問に対し、自動的に回答を提示する質疑応答機能である。近年では、こうした機能を実現する汎用的な学習モデルが各ベンダーから提供されている。チャットボットを用いた機械学習システムの処理の流れを整理すると、以下のとおりとなる（図表 5 を参照）。

（A） 金融機関（訓練データ提供者）は、過去に蓄積した照会ノウハウや、顧客の特性にあわせた商品情報に関するデータ（訓練データ）をベンダー（訓練実行者）に提供する。
（B） ベンダーは、訓練データを、チャットボット用の学習モデルに適用し、判定・予測エンジンを生成する。
（C） ベンダーは、上記（B）で生成した判定・予測エンジンを金融機関（判定・予測実行者）に提供する。
（D） 金融機関の顧客（システム利用者）は、スマートフォン・アプリや SNS を用いて照会事項を金融機関に提示する。
（E） 金融機関は、上記（D）の照会事項を判定・予測エンジンに適用し、回答内容を出力する。
（F） 金融機関は、上記（E）の出力結果を顧客に提示する。
（G） 金融機関は、必要に応じて、出力結果を還元データとして活用する。

[18] 近年では、顧客の口座残高に関する照会対応のほか、支出状況等の情報を分析して、消費動向に関するアドバイスを行い顧客の経済活動を総合的に支援したり、不正利用や二重払いの可能性を警告したりするスマートフォン・アプリ・サービスが提供されている。例えば、バンク・オブ・アメリカの「Erica」やキャピタル・ワンの「Eno」等がある（Bank of America [2018]、Capital One [2018]）。また、SNS 上におけるチャットボットとのやり取りを通じて、顧客に適した商品の提案や保険料金の見積りを提示するサービスも知られている（ライフネット生命［2018］）。

図表 5. チャットボットを用いた機械学習システムの構成（概念図）



このシステムの構成では、訓練データ提供者、判定・予測実行者は金融機関、訓練実行者はベンダー、システム利用者は顧客が担うため、構成タイプ 4 に該当する。

攻撃者は、システム利用者のデータを利用することが想定される。チャットボットが担う機能が一般的な照会事項への回答や商品の説明である場合には、そのデータの機密性は相対的に低いことが想定される。そのため、訓練データにかかる情報や判定・予測エンジンの推定が成功したとしても、顧客のパーソナル・データが漏洩したり、金融機関の収益に悪影響を及ぼしたりするような、致命的な脅威にはなりにくいと考えられる。もっとも、不正な判定・予測用データによって誤った判定・予測を誘発したり、不正な還元データを介して不正な判定・予測エンジンを生成したりすることによって、顧客の照会に対して不適切な回答が繰り返し生じた場合には、金融機関への信頼低下を招く可能性がある。このため、チャットボットを導入する金融機関は、還元データを用いてチャットボットの回答内容を確認することが望ましい。

ハ. 判断・予測の支援を目的とした機械学習システムの活用

融資審査における顧客の信用度評価システムは、判断・予測の支援を目的とした機械学習システムの代表的な事例の 1 つである[19]。信用度評価システムは、

[19] 個人ローンにかかる信用度評価をウェブ上で簡単に行うことができるサービスが知られている。顧客の年齢や年収、勤務先といった従来の審査項目のほか、その顧客の性格や趣

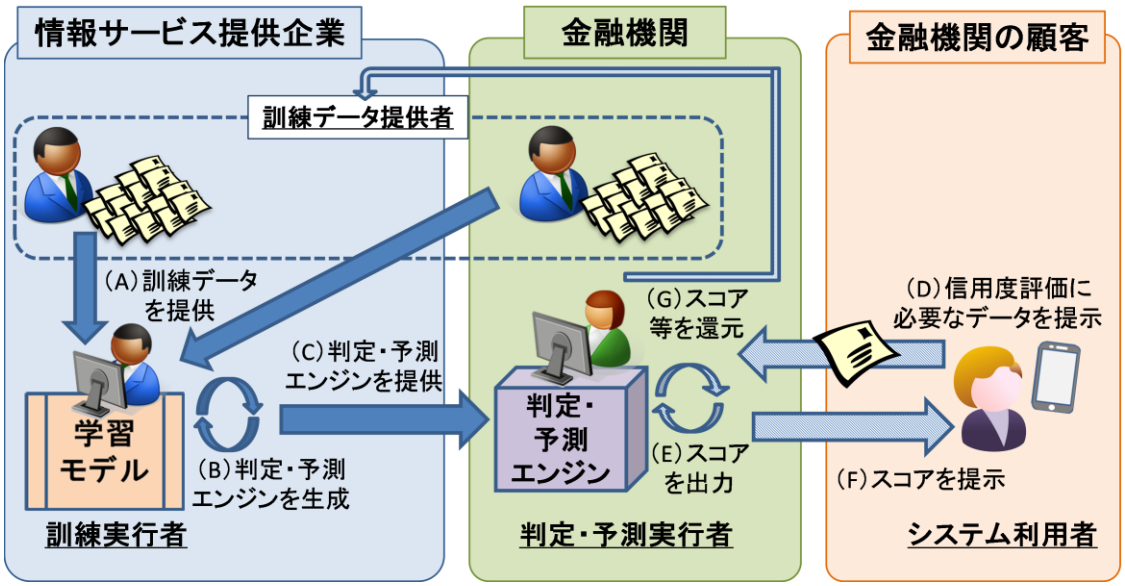
個人ローンを利用する顧客向けのシステムと、融資を行う金融機関向けのシステムに大別される。

（イ）個人ローンの顧客向けの信用度評価システム

金融機関が信用度評価システムを構築するためには、金融機関が有する情報のほか、顧客に関するさまざまなデータ（ビッグデータ）を有する企業から、情報の提供を受ける必要がある。例えば、ビッグデータを有し、機械学習システムを構築するノウハウを有する企業（情報サービス提供企業）と連携するケースが考えられる。そうした機械学習システムの処理の流れは、以下のとおりとなる（図表 6 を参照）。

（A） 金融機関（訓練データ提供者）は、顧客に関するデータ（訓練データ）を情報サービス提供企業（訓練実行者）に提供する。
（B） 情報サービス提供企業は、上記（A）で収集した訓練データに加えて、自社が有するデータを用いて、判定・予測エンジンを生成する。
（C） 情報サービス提供企業は、判定・予測エンジンを金融機関（判定・予測実行者）に提供する。
（D） 金融機関の顧客（システム利用者）は、スマートフォン・アプリや SNS を用いて、信用度評価に必要となるデータを金融機関に提示する。

図表 6. 顧客向けの信用度評価システムの構成（概念図）



味、ライフスタイル、ネットショッピングの実績といった多種多様な情報を基に、機械学習システムを用いて、顧客の信用度を数値化して出力する。金融機関は、この出力結果を融資審査における判定支援ツールとして活用する場合がある。

（E） 金融機関は、上記（D）で受け取ったデータを判定・予測エンジンに適用し、信用度評価の結果のスコアを出力する。
（F） 金融機関は、そのスコアを顧客に提示する。
（G） 金融機関および情報サービス提供企業は、必要に応じて、スコア等を還元データとして活用する。

このシステムの構成では、訓練データ提供者は金融機関と情報サービス提供企業、訓練実行者は情報サービス提供企業、判定・予測実行者は金融機関、システム利用者は金融機関の顧客が担う。金融機関からみた場合、訓練データ提供者と訓練実行者は、いずれも情報サービス提供企業が担っているため、同一のエンティティが担っているものとして整理すると、構成タイプ3に該当する[20]。

攻撃者は、システム利用者のデータを悪用することが想定される。具体的には、判定・予測用データの入出力を取得して、訓練データや判定・予測エンジンを推定する可能性がある。また、不正な判定・予測用データを入力して、誤ったスコアを誘発することも考えられる。

信用度評価システムでは、システム利用者として不特定多数の個人を想定していることから、認証等により攻撃者によるなりすましを防ぐことは困難である。そのため、訓練データを推定する攻撃は起こりうることを前提として、そうした攻撃が起きた場合にも、個人にかかる機密性の高い情報や個人の特定につながる情報が漏洩しないようにすることが求められる。例えば、年齢や収入等のパーソナル・データを訓練データに用いる場合には、そのままの数値ではなく、幅を持ったカテゴリー（年齢は10歳ごと、年収は100万円ごと等）に分類したうえで使用することなどが考えられる。

判定・予測エンジンの推定や誤ったスコアの誘発は、攻撃者による信用度評価の不正操作を可能ならしめ、その結果、本来よりも緩い条件での不適切な融資が実行されたり貸倒れに至るリスクが高まったりする可能性がある。さらに、誘発されたスコアが訓練データに還元されることによって判定・予測エンジンが不正に改変されると、他の顧客の信用度も正しく判定できなくなる可能性がある。対応策として、不正な判定・予測用データおよび訓練データを検知・排除する機能や、それらがスコアに与える影響を低減する工夫を、判定・予測エンジンに組み込むことが考えられる。

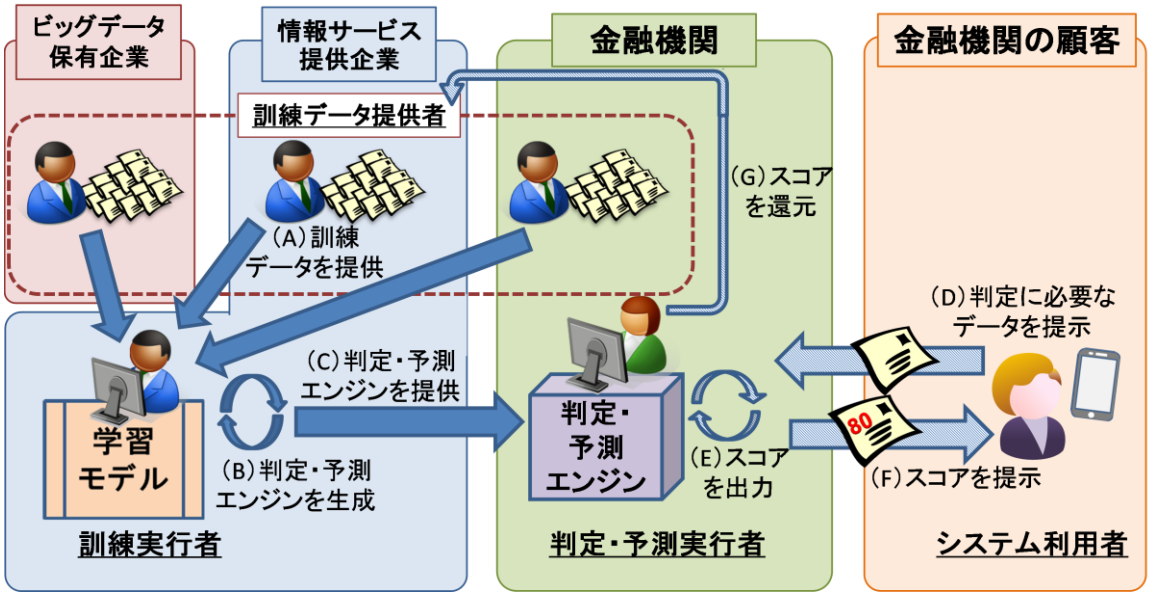
[20] 同様のサービスを、金融機関と情報サービス提供企業が一体となって設立した合弁会社が提供する場合には、金融機関と情報サービス提供企業を、それぞれ合弁会社に置き換えて解釈すればよい。つまり、訓練データ提供者、訓練実行者、判定・予測実行者の 3 つのエンティティを合弁会社が担い、システム利用者は金融機関の顧客が担うため、構成タイプ10に該当する。例えば、みずほ銀行（金融機関）とソフトバンク（情報サービス提供企業）が設立した（株）J.Score（合弁会社）が該当しうる（J.Score [2017a]）。

上記のシステムにおいて、更なる判定・予測精度の向上を企図して、顧客に関する他のデータを活用する場合も考えられる。そうしたビッグデータを保有する企業（ビッグデータ保有企業）は、金融機関、情報サービス提供企業とともに訓練データ提供者としての役割を担うことになる（図表 7 を参照）[21]。

エンティティの構成を整理すると、訓練データ提供者は金融機関、情報サービス提供企業およびビッグデータ保有企業、訓練実行者は情報サービス提供企業、判定・予測実行者は金融機関、システム利用者はその顧客が担う。ここで、金融機関からみた場合に、情報サービス提供企業は訓練データ提供者と訓練実行者の両者を担うものの、訓練データ提供者の役割をビッグデータ保有企業も担っている。そのため、訓練データ提供者と訓練実行者は、それぞれ異なるエンティティが担うものとして整理すると、構成タイプ 1 に該当する。

システム利用者のデータを悪用した攻撃に加え、ビッグデータ保有企業が提供する訓練データも攻撃対象となりうる。訓練データには、利用者の年齢や収入、性格や趣味といった機密性の高い情報が含まれる可能性がある。ビッグデータ保有企業の訓練データが攻撃者に盗取された場合にも、そのデータに対応する個人の特定が困難になるよう、真に必要な情報のみを抽出するなどの加工を予め施す必要がある。

図表 7. 複数のビッグデータを活用する信用度評価システムの構成（概念図）



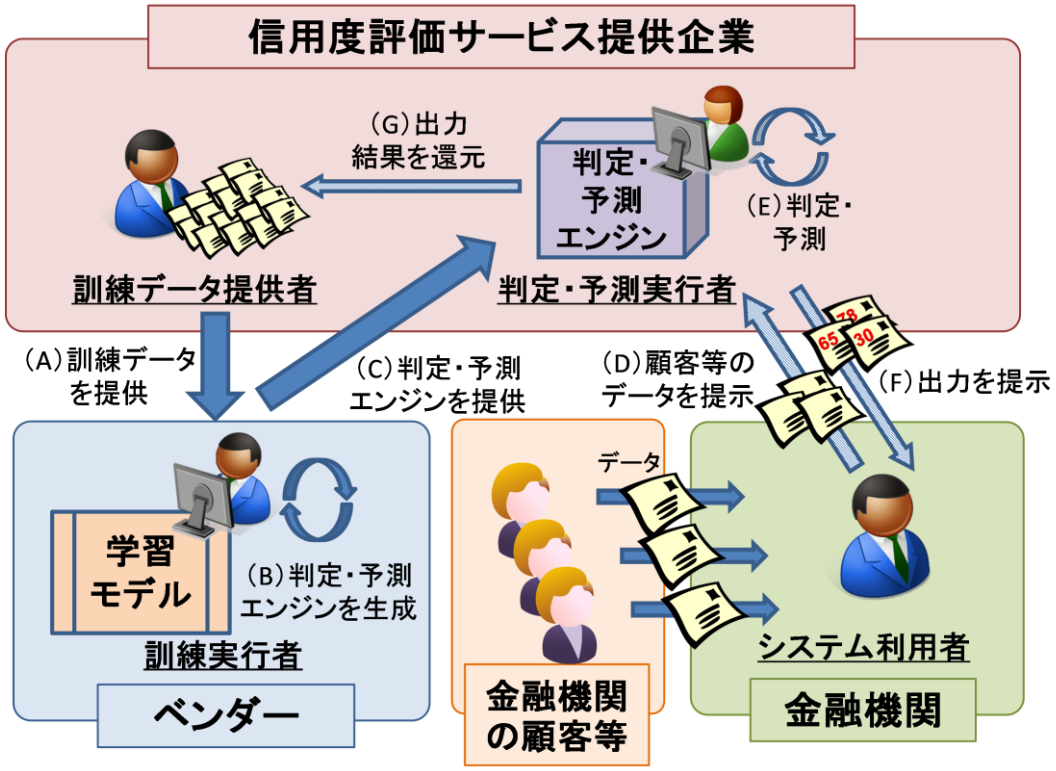
[21] このケースも、金融機関と情報サービス提供企業が合弁会社を設立する場合が想定される。訓練データ提供者は合弁会社とビッグデータ保有企業、訓練実行者および判定・予測実行者は合弁会社、システム利用者は金融機関の顧客が担うため、構成タイプ 6 に該当する。前述の（株）J.Score を例にとると、ビッグデータ保有企業はヤフー（株）となりうる（J.Score [2017b]）。

さらに、攻撃者は、システム利用者のデータのみを悪用できる場合と比較して、判定・予測エンジンを改変する攻撃を行いやすくなる。すなわち、攻撃者がシステム利用者のデータのみを悪用できる場合には、還元データを介してのみ攻撃可能（還元データにより再度訓練が行われない場合には、攻撃が不可能）であるが、訓練データを悪用できる場合には、その訓練データを改変することにより攻撃することも可能である。そのため、他の顧客のスコアが不正に操作されるリスクは、一層高くなる。

（ロ）金融機関向け信用度評価サービスへの応用

融資審査を行う際に、外部の信用度評価サービスを利用する金融機関もある。信用度評価サービスを提供する企業（信用度評価サービス提供企業）が、過去に蓄積した信用度評価の事例を訓練データとして機械学習システムを構築し、信用度評価の精度向上を試みるケースを考える。信用度評価サービス提供企業は、機械学習システムの開発ノウハウを有しておらず、判定・予測エンジンの生成をベンダーに委託すると仮定する。また、システム利用者である金融機関は、融資相手となる一般の個人の顧客や取引先企業等に関する、多数のデータを、判定・予測用データとして提示する。この場合、機械学習システムの処理の流れは以下のとおりとなる（図表 8 を参照）。

図表 8. 金融機関向け信用度評価システムの構成（概念図）



（A） 信用度評価サービス提供企業（訓練データ提供者）は、過去に蓄積した信用度評価の事例（訓練データ）をベンダー（訓練実行者）に提供する。
（B） ベンダーは、上記（A）の訓練データを用いて、判定・予測エンジンを生成する。
（C） ベンダーは、判定・予測エンジンを信用度評価サービス提供企業（判定・予測実行者）に提供する。
（D） 金融機関（システム利用者）は、顧客等から収集したデータを信用度評価サービス提供企業に提示する。
（E） 信用度評価サービス提供企業は、上記（D）のデータを判定・予測エンジンに適用し、信用度評価に関する出力を得る。
（F） 信用度評価サービス提供企業は、上記（E）の出力を金融機関に提示する。
（G） 信用度評価サービス提供企業は、必要に応じて、出力を還元データとして活用する。

このシステムの構成では、訓練データ提供者および判定・予測実行者は信用度評価サービス提供企業、訓練実行者はベンダー、システム利用者は金融機関が担うため、構成タイプ 4 に該当する。

攻撃者は、システム利用者である金融機関になりすます、あるいは金融機関の内部の者と結託することによって、システム利用者のデータを悪用することが想定される。なりすましは、金融機関が、信用度評価サービス提供企業との間で、判定・予測の入出力にかかる一連のデータをやり取りする際に、認証等の一般的なセキュリティ対策を講じることによって防ぐことができる。一方、金融機関の内部の者との結託は、金融機関において、判定・予測エンジンの入出力にアクセス可能な職員を限定したり、職員ごとのアクセス履歴を記録したりすることにより防ぐことが求められる。

ニ．リスク低減を目的とした機械学習システムの活用

金融機関の経営リスクを低減する手段として、近年、機械学習システムを用いた金融市場の異常検知やクレジットカード等の不正取引検知が注目されている。金融市場の異常検知では、過去の注文、市場流動性、価格変動といった情報から、金融市場の正常状態を学習することで、異常を検知する。また、クレジットカード等の不正取引検知では、過去の不正取引のデータから、そのパターンや特徴を学習することで、類似の不正取引を検知する[22]。

[22] 金融機関向けにクラウド上で提供される不正取引検知サービス等が知られている。

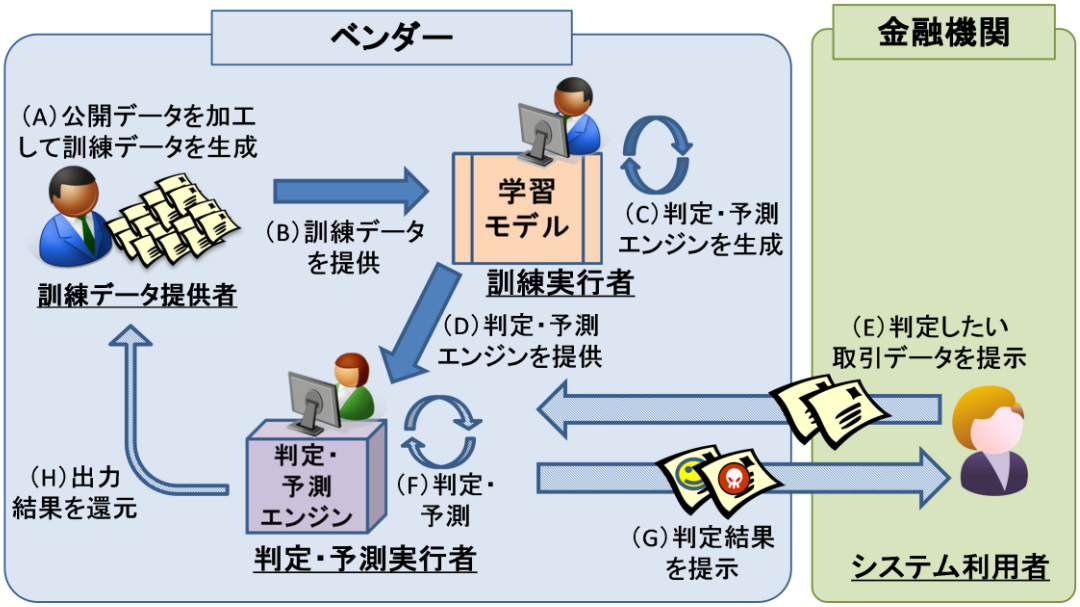
ベンダーが、公開された金融市場のデータや取引データを訓練データとして機械学習システムを構築し、異常や不正取引を検知するサービスを金融機関向けに提供する場合、処理の流れは以下のとおりとなる（図表 9 を参照）[23]。

（A）　ベンダー（訓練データ提供者）は、公開データを訓練データとして活用するために、適宜の形式に加工する。
（B）（C）（D）　ベンダー（訓練実行者）は、上記（A）で加工した訓練データを用いて、判定・予測エンジンを生成する。
（E）　金融機関（システム利用者）は、判定したい取引データをベンダー（判定・予測実行者）に提示する。
（F）（G）　ベンダーは、上記（E）の取引データを判定・予測エンジンに適用して、判定結果を出力し、それを金融機関に提示する。
（H）　ベンダーは、必要に応じて、判定結果を還元データとして活用する。

このシステムの構成では、訓練データ提供者、訓練実行者および判定・予測実行者はベンダー、システム利用者は金融機関が担うため、構成タイプ 10 に該当する。

攻撃者は、システム利用者になりすましたり、システム利用者と結託したり

図表 9. 異常や不正取引を検知するための機械学習システムの構成（概念図）



[23] ここでは、訓練データとして公開データを用いる場合を例に取り上げたが、不正取引検知用の機械学習システムでは、金融機関が所有する過去の取引データを訓練データとして用いる場合もある。この場合、訓練データ提供者は金融機関となり、処理の流れは、事務の効率化を目的とした機械学習システムの構成（図表 4）と同様となる。

することによって、システム利用者が有する情報を用いた攻撃を行う。具体的には、判定・予測用データの入出力に関するデータを取得して、訓練データや判定・予測エンジンを推定する。また、不正な判定・予測用データを入力して、誤った判定結果を誘発したり、不正な判定・予測エンジンを生成したりすることも考えられる。

上記のシステムでは、訓練データとして公開データを用いるため、訓練データを推定されても問題はない。また、判定・予測エンジンについても、ベンダーにとっては重要な資産であるが、金融機関にとっては、エンジンが推定されたとしても特段の影響は生じない。一方、不正な判定・予測用データによる攻撃の影響は、システムの利用目的により異なる。金融市場の異常を検知する場合、攻撃により、異常を検知できない、または、正常時に異常と判断するといった誤判定が誘発される可能性がある。金融機関が、こうしたシステムの判定結果を基に金融取引を行う場合、その取引により経済的損失を被ったり不正な取引を実行したりするなどのリスクが生じる。

また、クレジットカード等の不正取引検知システムの場合には、不正取引にかかるデータが検知されないようにする攻撃が考えられる。こうした不正な出力結果が還元データとして用いられると、判定・予測エンジンが改変され、攻撃者が攻撃に用いたデータ以外のデータについても正しく判定できなくなる可能性がある。対応策として、不正な判定・予測用データおよび還元データを検知・排除したり、それらが出力結果に与える影響を低減したりする機能を、判定・予測エンジンに組み込むことが有用であると考えられる。

５．おわりに

金融分野における機械学習システムの利活用は始まったばかりの段階にあり、深刻なセキュリティ被害の報告はまだ聞かれていない。しかし、機械学習システムには、従来の情報システムが有する脆弱性に加えて、機械学習システムに特有の脆弱性も存在する。機械学習システムを中長期的に、安全かつ安定的に活用していくうえでは、こうしたリスクが顕在化する前に、予め、システムに潜む脆弱性やセキュリティ上のリスクを十分に把握し、対策を講じることが重要である。

本稿でみたとおり、機械学習システムの構成は 12 の構成タイプに分類されるが、各構成タイプにおいて、どのエンティティを金融機関が担うかによって、さらにバリエーションは多様化する。機械学習システムを導入する際には、そのシステムの構成が、どの構成タイプに該当するかを整理し、考えうる脆弱性を洗い出すことが肝要である。そのうえで、当該システムにおいて各エンティティが取り扱うデータの重要性等を考慮し、脆弱性が顕在化した際に金融機関

に与える影響の多寡を見極めて対策を検討することになる。

情報システムの技術が日々進化しているのと同様に、それらを狙った攻撃手法も日々巧妙化しており、これまで対処可能であった攻撃に対して、さらなる対策が必要となる場合もある。セキュリティ対策を考える際には、最新の攻撃手法とそれへの対策手法について、技術の進展を踏まえつつ検討していくことが求められる。

以　上

【参考文献】

宇根正志、「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」、金融研究所ディスカッション・ペーパーNo. 2018-J-16、日本銀行金融研究所、2018 年

――――・廣川勝久、「モバイル端末による金融サービスの安全性を高めるために : セキュア・エレメント等の活用」、金融研究所ディスカッション・ペーパーNo. 2017-J-15、日本銀行金融研究所、2017 年

吉岡信和、「機械学習システムがセキュリティに出会うとき」、『第 1 回機械学習工学ワークショップ（MLSE2018）論文集』、機械学習工学研究会、2018 年、49～53 頁

ライフネット生命、「ライフネット生命 LINE 公式アカウント」、ライフネット生命、2018 年（https://www.lifenet-seimei.co.jp/sph/line/、2018 年 12 月 6 日）

J.Score、「みずほ銀行とソフトバンクの合弁会社 J.Score が日本初の FinTech サービス『AI スコア・レンディング』を本日より提供開始」、J.Score、2017 年 a（https://www.jscore.co.jp/company/news/2017/0925_01/、2018 年 12 月 6 日）

――――、「J.Score と Yahoo!JAPAN の業務提携契約の締結に関するお知らせ」、J.Score、2017 年 b（https://www.jscore.co.jp/company/news/2017/1222_01/、2018 年 12 月 6 日）

NTT データ、「AI を活用したチャットボットの試行提供を開始～AI 技術『corevo®』を活用し、金融機関向け共同利用型チャットボットを実現～」、NTT データ、2017 年
（http://www.nttdata.com/jp/ja/news/services_info/2017/2017060901.html、2018 年 11 月 16 日）

Abadi, Martin, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang, “Deep Learning with Differential Privacy,” *Proceedings of ACM SIGSAC Conference on Computer and Communications Security (CCS) 2016*, Association for Computing Machinery, 2016, pp. 308-318.

――――, Úlfar Erlingsson, ――――, ――――, ――――, Nicolas Papernot, Kunal Talwar, and Li Zhang, “On the Protection of Private Information in Machine Learning Systems: Two Recent Approaches,” *Proceedings of IEEE Computer Security Foundations Symposium 2017*, IEEE, 2017, pp. 1-6.

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, pp.

137-150.

Bank of America, “Erica Makes Banking Easier than Ever,” Bank of America, 2018 (available at: https://promo.bankofamerica.com/erica/、2018 年 11 月 19 日).

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.

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

Capital One, “Eno, An Intelligent Assistant from Capital One,” Capital One, 2018 (available at: https://www.capitalone.com/applications/eno/、2018 年 12 月 6 日).

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, 2017, pp. 3-14.

Dowlin, Nathan, Ran Gilad-Bachrach, Ran, 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.

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.

Goodfellow, Ian, "Security and Privacy of Machine Learning," presentation at RSA Conference 2018, RSA, 2018 (available at: http://www.iangoodfellow.com/slides/2018-04-rsa.pdf、2018 年 11 月 19 日).

————, Patrick McDaniel, and Nicolas Papernot, "Making Machine Learning Robust against Adversarial Inputs," *Communications of the ACM*, 61(7), Association for Computing Machinery, 2018, pp. 56-66.

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

Mohassel, Payman, and Peter Rindal, "ABY^3: A Mixed Protocol Framework for Machine Learning," *Proceedings of ACM SIGSAC Conference on Computer and Communications Security (CCS) 2018*, Association for Computing Machinery, 2018, pp. 35-52.

————, and Yupeng Zhang, "SecureML: A System for Scalable Privacy-Preserving Machine Learning," *Proceedings of IEEE Symposium on Security and Privacy (SP) 2017*, IEEE, 2017, pp. 19-38.

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, Martín Abadi, Úlfar Erlingsson, Ian Goodfellow, and Kunal Talwar, "Semi-Supervised Knowledge Transfer for Deep Learning from Private Training Data," Talk at International Conference on Learning Representations (ICLR) 2017, OpenReview.net, 2017a (available at: https://openreview.net/pdf?id=HkwoSDPgg、2018 年 12 月 7 日).

————, 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 (ASIACCS) 2017*, Association for Computing Machinery, 2017b, pp. 506-519.

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

————, ————, Xi Wu, Somesh Jha, and Ananthram Swami, "Distillation as a Defense to Adversarial Perturbations against Deep Neural Networks,"

Proceedings of IEEE Symposium on Security and Privacy (SP) 2016, IEEE, 2016b, pp. 582-597.

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 Security 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.

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.

補論．機械学習システムの構成のバリエーション

（１）15 の構成タイプ

機械学習システムの構成のバリエーションは、単一または複数の主体が 4 つのエンティティの役割をどう担うかという観点から分類することができる。各エンティティの役割をそれぞれ異なる主体が担う場合、4 つの主体が機械学習システムに関与する。また、すべてのエンティティの役割を同一の主体が担う場合には、その主体のみが機械学習システムに関与する。1 つのエンティティの役割を 1 つの主体が担うと仮定すれば、機械学習システムに関与する主体の数は、最大で 4、最小で 1 となり、これらの中間として 2 つあるいは 3 つの主体が関与するケースもあることから、機械学習システムの構成として、論理的には 15 のバリエーション（構成タイプ）が想定される（図表 A1 を参照）[24]。

（２）各構成タイプの分析

イ．構成タイプ 1

この構成タイプでは、各エンティティをそれぞれ異なる主体が担う。例えば、複数のマルウェア検知エンジンによるオンラインでの検査サービスが想定される。こうしたエンジンを提供する複数のセキュリティ・ベンダー（訓練実行者）

図表 A1．機械学習システムの構成の 15 のバリエーション

構成タイプ	各エンティティを担う主体				主体数
	訓練データ提供者	システム利用者	訓練実行者	判定・予測実行者	
1	▲	○	■	◇	4
2	▲		■	◇	3
3	▲	○	▲	◇	
4	▲	○	■	▲	
a	▲	○		◇	
5	▲	○	■	○	
6	▲	○	■		
7	▲		□		2
8	▲	○	▲	○	
b	▲	○		▲	
9	▲	○			
10	▲	○	▲		
11	▲		■	▲	
c	▲			◇	
12	▲				1

備考：1. 各エンティティの役割を担う主体を▲、○、■、◇等の記号で表示。
2. 有色（白以外）かつ同色のセルのエンティティは同一の主体が担う。

[24] 例えば、訓練データ提供者の役割を複数の主体が担う場合が考えられる。こうした場合については、それらの複数の主体の集合を 1 つの主体とみなすこととする。また、訓練データ提供者が訓練データを訓練実行者に提供しつつその処理の一部を実行する場合も想定される（Phong [2017]）。この場合については、学習モデルの出力となる判定・予測エンジンを最終的に生成する主体が訓練実行者の役割を担うとみなすこととする。

が、さまざまなマルウェアの情報をインターネット上のハニーポット等の管理者（訓練データ提供者）から入手・解析し、マルウェア検知エンジンを生成する。マルウェアの検査サービスを提供するポータル（判定・予測実行者）は、マルウェア検知エンジンを取得した後、マルウェア検査の依頼者（システム利用者）からインターネット経由で送信された検査対象のデータを検査し、その結果を依頼者に返信する。このような構成でのサービスが実際に提供されているか否かは定かでないものの、類似のサービスが既に提供されており、機械学習システムの文脈でも同様のサービスが提供される可能性はあるとみられる[25]。

ロ．構成タイプ 2

この構成タイプでは、訓練データ提供者とシステム利用者の役割を同一の主体が担い、その主体とは別の複数の（異なる）主体が、訓練実行者と判定・予測実行者の役割をそれぞれ担う。例えば、金融機関（訓練実行者）が、自社のモバイル金融取引のサービスを利用する顧客（訓練データ提供者、システム利用者）から、顧客のプロファイルや金融資産等にかかるデータを訓練データとして利用する場合が想定される。金融機関は、数多くの顧客のデータを用いて判定・予測エンジンを生成し、それを各顧客のモバイル端末内部の安全な領域（例えば、セキュア・エレメント＜Secure Element＞やトラステッド・エグゼキューション・エンバイロメント＜Trusted Execution Environment＞）に格納する[26]。ここで、その領域が、セキュア・エレメント等の管理者やモバイル通信キャリアであるとすれば、これらの主体が判定・予測実行者といえる。顧客は、自分のプロファイルに合致した（モバイル端末内部の）判定・予測エンジンを用いて、判定・予測結果を得る。

ハ．構成タイプ 3

この構成タイプでは、訓練データ提供者と訓練実行者の役割を同一の主体が担い、その主体とは別の複数の（異なる）主体が、システム利用者と判定・予測実行者の役割をそれぞれ担う。例えば、インターネット上でのオンライン商取引の事業者（訓練データ提供者、訓練実行者）が、自社の商取引にかかるデー

[25] こうした構成による類似のサービスの例として「virustotal」が挙げられる。

[26] セキュア・エレメントは、暗号処理等のセキュリティ機能を有するとともに、外部からの物理的な攻撃に対しても高い安全性を有するモジュールの総称であり、ハードウェアとソフトウェアを組み合わせて実現される。トラステッド・エグゼキューション・エンバイロメントは、セキュア・エレメントに関連する実行環境であり、主にソフトウェアによって通常の実行環境から分離された安全な実行環境を実現する。セキュア・エレメントやトラステッド・エグゼキューション・エンバイロメントについては、宇根・廣川［2017］を参照されたい。

タ（例えば、購買履歴）を訓練データとして判定・予測エンジン（例えば、推奨する商品を検索・抽出するエンジン）を生成する場合が考えられる。上記の事業者は、生成した判定・予測エンジンを電子商取引のウェブサイトの管理者（判定・予測実行者）に提供し、そのウェブサイトにアクセスしたユーザー（システム利用者）に対して、推奨する商品を提示する。

ニ．構成タイプ4

この構成タイプでは、訓練データ提供者と判定・予測実行者の役割を同一の主体が担い、その主体とは別の（複数の）異なる主体が、システム利用者と訓練実行者の役割をそれぞれ担う。例えば、金融機関（訓練データ提供者、判定・予測実行者）が、自社のデータを訓練データとして機械学習のベンダー（訓練実行者）に提供する場合が想定される。上記のベンダーは、判定・予測エンジンを生成して金融機関に送り、金融機関はそれを用いた判定・予測のサービスを顧客（システム利用者）に提供する。

ホ．構成タイプa

この構成タイプでは、システム利用者と訓練実行者を同一の主体が担い、訓練データ提供者と判定・予測実行者を、上記の主体とは別の主体がそれぞれ担う。この構成タイプでは、システム利用者を担う主体は、訓練実行者として判定・予測エンジンを生成することができるため、他のエンティティと通信することなく、その判定・予測エンジンを使用することができる。したがって、判定・予測エンジンのみを別の主体（判定・予測実行者）に渡すことは考えにくい。これは、判定・予測エンジンのみを別の主体に渡すこととすれば、判定・予測時にその主体と通信するなどの追加的な処理を実施することが必要となるためである。このように考えると、この構成タイプに相当する機械学習システムを想定しづらいといえるため、検討対象から除外する。

ヘ．構成タイプ5

この構成タイプでは、訓練データ提供者と訓練実行者の役割をそれぞれ異なる主体が担い、これらの主体とは異なる単一の主体が、システム利用者と判定・予測実行者の役割を担う。例えば、セキュリティ・ベンダー（訓練実行者）がハニーポット等の運営者（訓練データ提供者）からマルウェアについてのデータを訓練データとして入手し、判定・予測エンジンを生成する場合が想定される。セキュリティ・ベンダーは、その判定・予測エンジンを金融機関（システム利用者、判定・予測実行者）に提供する。金融機関はそれを用いて自社内部でマルウェア対策を実施する。

ト．構成タイプ 6

この構成タイプでは、訓練データ提供者とシステム利用者の役割をそれぞれ異なる主体が担うほか、これらの主体とは別の単一の主体が、訓練実行者と判定・予測実行者の役割を担う。例えば、金融機関（訓練データ提供者）が、自社のデータを訓練データとして外部事業者（訓練実行者と判定・予測実行者）に提供し、外部事業者がその訓練データを用いて判定・予測エンジンを生成する場合が想定される。金融機関の顧客（システム利用者）は、外部事業者にアクセスして判定・予測エンジンの使用を依頼し、判定・予測結果を得る。

チ．構成タイプ 7

この構成タイプでは、訓練データ提供者とシステム利用者を同一の主体が担うとともに、訓練実行者と判定・予測実行者を、上記の主体とは別の単一の主体が担う。例えば、クラウドによって提供される機械学習のサービスが想定される。機械学習向けのクラウドを運営する外部事業者（訓練実行者と判定・予測実行者）は、そのサービスのユーザー（訓練データ提供者とシステム提供者）から訓練データを入手して学習モデルを実行し、判定・予測エンジンを生成する。ユーザーは判定・予測したいデータを外部事業者に送信し、外部事業者は判定・予測結果をユーザーに送信する。

リ．構成タイプ 8

この構成タイプでは、訓練データ提供者と訓練実行者を同一の主体が担うとともに、システム利用者と判定・予測実行者を、上記の主体とは別の単一の主体が担う。例えば、システム・ベンダー（訓練データ提供者と訓練実行者）が、ユーザー（システム利用者と判定・予測実行者）の依頼に応じて、自社のデータを訓練データとして用いて判定・予測エンジンを生成する場合が考えられる。システム・ベンダーは、その判定・予測エンジンをユーザーに提供する。

ヌ．構成タイプ b

この構成タイプでは、システム利用者と訓練実行者を同一の主体が担うほか、訓練データ提供者と判定・予測実行者を、上記の主体とは別の単一の主体が担う。この構成タイプでは、上記の構成タイプ a と同様に、システム利用者を担う主体は、訓練実行者として判定・予測エンジンを生成することができるため、他のエンティティと通信することなく、その判定・予測エンジンを使用することができる。したがって、判定・予測エンジンのみを別の主体（判定・予測実行者）に渡すことは考えにくく、この構成タイプに相当する機械学習システム

を想定しづらいといえることから、検討対象から除外する。

ル．構成タイプ 9

この構成タイプでは、システム利用者と訓練実行者と判定・予測実行者の役割を同一の主体が担い、その主体とは別の主体が訓練データ提供者の役割を担う。例えば、金融機関（システム利用者と訓練実行者と判定・予測実行者）が、外部の事業者からデータを入手し（例えば、市況データ）、それを訓練データとして自ら判定・予測エンジンを生成する場合が想定される。金融機関は、その判定・予測エンジンを用いて市況予測等を行う。

ヲ．構成タイプ 10

この構成タイプでは、訓練データ提供者と訓練実行者と判定・予測実行者の役割を同一の主体が担い、その主体とは別の主体がシステム利用者の役割を担う。例えば、金融機関（訓練データ提供者と訓練実行者と判定・予測実行者）が、自社のデータを訓練データとして自ら判定・予測エンジン（例えば、信用度や融資判定のエンジン）を生成し、金融機関の顧客（判定・予測実行者）に提供する場合が想定される。その顧客は、判定・予測エンジンを自身のスマートフォンにアプリとして格納・使用する。

ワ．構成タイプ 11

この構成タイプでは、訓練データ提供者とシステム利用者と判定・予測実行者を同一の主体が担い、訓練実行者を、上記の主体とは別の主体が担う。例えば、金融機関（訓練データ提供者とシステム利用者と判定・予測実行者）が、自社のデータを訓練データとして機械学習のベンダー（訓練実行者）に提供し、ベンダーが判定・予測エンジンを生成して金融機関に提供する場合が想定される。

カ．構成タイプ c

この構成タイプでは、訓練データ提供者とシステム利用者と訓練実行者を同一の主体が担い、判定・予測実行者を上記の主体とは別の主体が担う。上記の構成タイプ a、b と同様に、システム利用者を担う主体は、訓練実行者として判定・予測エンジンを生成するため、他のエンティティと通信することなく判定・予測エンジンを使用することができる。したがって、判定・予測エンジンのみを別の主体（判定・予測実行者）に渡すことは考えにくいことから、この構成タイプに相当する機械学習システムを想定しづらく、検討対象から除外する。

ヨ. 構成タイプ 12

例えば、1 つの金融機関が、自社内のデータを訓練データとし、学習モデルを用いて判定・予測エンジンを生成・使用する場合、このケースに相当する。