2F-05

機械学習システムの構成に基づくセキュリティ分析

井上 紫織† 宇根 正志†

日本銀行金融研究所情報技術研究センター†

1. はじめに

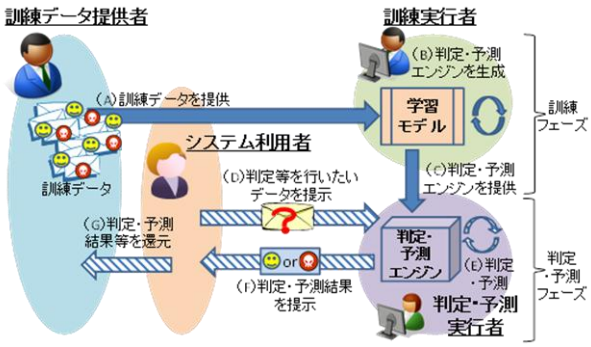
近年，様々な領域で人工知能（artificial Intelligence：AI），とりわけ機械学習システムの活用にかかる検討が進んでいる．AI は，一般に，推論や認識，判断等，人間と同様の知的な処理能力を持つコンピュータ・システムやその技術分野を指し，その機能を実現するツールとして用いられる技術が機械学習である．機械学習を実装したシステム（以下，機械学習システム）では，大量のサービス要求による機能低下等，情報システム一般に存在する脆弱性に加え，特有の脆弱性も存在する[1][2]．こうした脆弱性が悪用されると，機械学習システムにおいて処理されるデータや学習モデル，判定・予測エンジンが，盗取され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，機械学習システムを構成する機能の担い手（構成タイプ）を整理・分類したうえで，各構成タイプにおいて想定される脆弱性と攻撃，対応策を整理する[3]．

本稿で示されている意見は，筆者ら個人に属し，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ら個人に属する．

2. 機械学習システムの構成と分類

2.1 機械学習システムの構成

機械学習システムは，次の 4 つのエンティティにより構成されるものとする．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図表 1 のとおりである．



図表 1　機械学習システムの構成（イメージ）

Classification and Security Analysis of Machine Learning Systems
† Shiori Inoue and Masashi Une, Center for Information Technology Studies (CITECS), Institute for Monetary and Economic Studies (IMES), Bank of Japan

2.2 機械学習システムの構成の分類

機械学習システムでは，2.1 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．ただし，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構成は想定しづらいことから，実際に想定される機械学習システムの構成は図表 2 に示す 12 のタイプ（構成タイプ）になる．

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	■	■	■	■	1
12	■	■	■	■	

備考：同じパターンのセルは同じエンティティが担う．

3. 攻撃と対応策

3.1 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は，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）を確保することである[4]．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．

3.2 セキュリティと攻撃者の能力の前提

攻撃者は，機械学習システムの第三者であり，3.1 に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．訓練実行者，判定・予測実行者および各通信路は，サイバー攻撃への対策を十分に講じており，攻撃者による直接の攻撃は受けず，攻撃者が悪用しうるデータは，訓練データ提供者やシステム利用者のデータであるとする．また，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

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	構成タイプ 3,4,10	構成タイプ 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[5].	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[6].			
		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・予測エンジンへの入出力	判定・予測エンジンを推定[7].	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[8].	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 [10].			
		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 [9].	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[11].			
		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.

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おり，攻撃者によりデータを悪用されないものとする.

3.3 攻撃と対応策

3.2 の前提を踏まえ，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おいて想定される攻撃と対応策，および各攻撃と対応策に該当しうる構成タイプを纏めると，図表 3 のとおりとなる.

4. 考察と今後の課題

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，まず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，想定される攻撃とそれへの対応策を特定する必要がある．そのうえで，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を検討する．それらが許容できる場合には，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．また，攻撃と対応策にかかる技術は日々進展することを踏まえ，講じるべきセキュリティ対策を定期的に見直すことが肝要である.

参考文献

[1] 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. IMES Discussion Paper Series. 2018, no.2018-J-16.

[2] 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, 49〜53 頁

[3] 井上紫織・宇根正志, 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. IMES Discussion Paper Series. 2019, no.2019-J-1.

[4] Papernot, N. et al., Towards the Science of Security and Privacy in Machine Learning. arXiv: 1611.03814v1.Cornell University Library. 2016.

[5] Goodfellow, I. et al., Making Machine Learning Robust against Adversarial Inputs, Communications of the ACM, 61(7), Association for Computing Machinery, 2018, pp. 56-66.

[6] Carlini, N. et al., Adversarial Examples Are Not Easily Detected: Bypassing Ten Detection Methods, Proceedings of ACM Workshop on Artificial Intelligence and Security (AISec), Association for Computing Machinery, 2017, pp. 3-14.

[7] Tramèr, F. et al., Stealing Machine Learning Models via Prediction APIs, Proceedings of USENIX Security Symposium, Advanced Computing Systems Association,2016, pp.601-618.

[8] Shokri, R. et al., Membership Inference Attacks Against Machine Learning Models, Proceedings of IEEE Symposium on Security and Privacy, 2017, pp.3-18.

[9] Szegedy, C. et al., Intriguing Properties of Neural Networks, Proceedings of International Conference on Learning Representations (ICLR), arXiv: 1312.6199v4, Cornell University Library, 2014.

[10] Abadi, M. et al., On the Protection of Private Information in Machine Learning Systems: Two Recent Approaches, Proceedings of IEEE Computer Security Foundations Symposium, 2017, pp.1-6.

[11] Papernot. et al., Distillation as a Defense to Adversarial Perturbations against Deep Neural Networks, Proceedings of IEEE Symposium on Security and Privacy, 2016, pp. 582-597.