Bilevel Optimization for Defense against Model Extraction
（二段階最適化によるモデル抽出攻撃に対する防御）

数理情報学専攻 48196231　森 雄人
指導教員　武田 朗子 教授

1　概要

モデル抽出攻撃と呼ばれる, 機械学習モデルを入出力から再学習する攻撃に対する防御が機械学習を用いるサービスの発展に伴い喫緊の課題となっている. 本研究では攻撃者と防御者の入力データの分布の違いを踏まえ防御問題を二段階最適化問題として定式化し, 防御が可能である十分条件について簡潔に理論的な考察を与える. さらに具体的に, 攻撃者と防御者がカーネル回帰に基づくモデルを用いる場合, 本防御問題が一つの二次制約付き非凸二次最適化問題に帰着されることを示し, その大域的最適解を多項式時間で求めるアルゴリズムを示す. また, 攻撃者が確率的勾配降下法に基づいて攻撃をする場合においても防御問題を定式化し, 勾配法に基づく防御を行うアルゴリズムを提案する. 二つの設定それぞれについて実データに基づいた数値実験を行い, ともに攻撃者の入力データが防御者の入力データから離れている場合に有効な防御が可能であることを示す. また, 攻撃者の入力データに関する汎化性能を数値実験によって確認する. この汎化性能により, 提案手法は一度防御モデルを構成すれば攻撃者のデータに応じて新たに防御をする必要がないため, 入力一つごとに防御問題を解く既存の防御と比較して実際のサービスの要請により沿うモデルの構築へとつながる.

2　二段階最適化によるモデル抽出攻撃に対する防御 (BODAME)

本研究ではモデル抽出攻撃に対する防御問題を次のような二段階最適化問題として定式化する.

$$\max_{\theta\in\Theta} \quad \mathop{\mathbb{E}}_{X\sim Q}[l^{(\mathrm{o})}(f(X), h_{\tilde{w}(\theta)}(X))], \tag{1}$$

$$\text{s.t.} \quad \tilde{w}(\theta) = \mathop{\arg\min}_{w\in W} \mathop{\mathbb{E}}_{X\sim P}[l^{(\mathrm{a})}(g_\theta(X), h_w(X))], \tag{2}$$

$$\mathop{\mathbb{E}}_{X\sim Q}[l^{(\mathrm{c})}(f(X), g_\theta(X))] \le \varepsilon. \tag{3}$$

ここで, f は防御者が持つ真のモデル, g_θ は防御者が作成する代理モデル, h_w は攻撃者がモデル抽出攻撃によって作成するモデルである. Q は防御者が持つ真の入力データ分布であり, P は攻撃者が持つ入力データ分布である. (2) のように攻撃者は自身の入力データ分布 P 上で, 代理モデル g_θ を教師データとして損失関数 $l^{(\mathrm{a})}$ を最小化することでモデル抽出攻撃を行う. 防御者は真のモデル f と代理モデル g_θ が防御者の分布 Q 上では損失 $l^{(\mathrm{c})}$ の意味で ε ほどしか離れないという条件 (3) の下で, 防御者は攻撃者によって抽出されたモデル $h_{\tilde{w}(\theta)}$ が防御者の分布 Q 上, 損失 $l^{(\mathrm{o})}$ の意味で真のモデルから離れるような代理モデルを作成することを試みる. この最適化問題を「二段階最適化によるモデル抽出攻撃に対する防御」 (Bilevel Optimization for Defense Against Model Extraction, **BODAME**) と呼ぶ. 本研究では, この最適化問題の極端な条件における理論解析を通して攻撃者の分布と防御者の分布の台が異なるという条件が防御のために重要であることを示す.

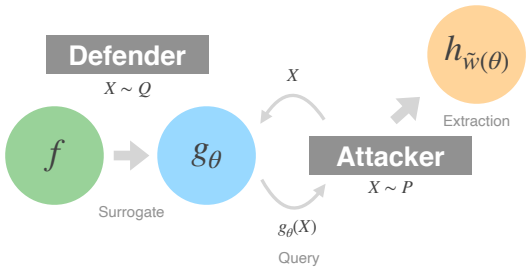

図 1. 本研究における攻撃者と防御者のモデルの概観. 防御者は真の分布 Q と真のモデル f を持っており, 代理モデル g_θ を作成することで直接真のモデル f を公開することを避ける. 攻撃者は P に従ってデータを入力することでモデル抽出攻撃を行い, $h_{\tilde{w}(\theta)}$ を作成する.

3　BODAME–KRR/KR

攻撃者がカーネルリッジ回帰 (Kernel Ridge Regression, KRR) モデルを, 防御者がカーネル回帰 (Kernel Regression, KR) モデルを用いるとき, P や Q からサンプリングした標本と各モデルから導かれる $A, a, \gamma_a, B, b, \gamma_b$ という値を用いて, BODAME は次のような一つの二次制約付き非凸二次最適化問題に帰着される.

$$\max_{\theta\in\mathbb{R}^M} \quad \theta^\top A\theta - 2a^\top\theta + \gamma_a, \tag{4}$$
$$\text{s.t.} \quad \theta^\top B\theta - 2b^\top\theta + \gamma_b \le \varepsilon. \tag{5}$$

本研究では目的関数が凹関数であることに留意し, 一部の手順を簡略化した一般化固有値問題の解法に基づくアルゴリズム [1] を用いることで (4), (5) の大域的最適解が得られることを示す.

4 BODAME–SGD/SGA

実際の機械学習モデルの学習では, 攻撃者は (2) において大域的最適解を得ることは少ない. 例えば, ニューラルネットワークなどのカーネル法に比べて複雑なモデルを用いる場合, 攻撃者は有限ステップの確率的勾配降下法 (Stochastic Gradient Descent, SGD) に基づく最適化を行う. そこで, (2) の条件を陽に勾配降下法などの微分可能な写像 Φ_t を用いて各ステップ t ごとにパラメータ更新するという条件に置き換えた次の最適化問題を考える.

$$\max_{\theta\in\Theta} \quad \mathop{\mathbb{E}}_{X\sim Q}[l^{(\mathrm{o})}(f(X), h_{w^{(T)}(\theta)}(X))], \tag{6}$$
$$\text{s.t.} \quad w^{(0)} = w_0, \tag{7}$$
$$w^{(t+1)} = \Phi_t(w^{(t)}, \theta) \quad (t = 0\ldots, T-1), \tag{8}$$
$$\mathop{\mathbb{E}}_{X\sim Q}[l^{(\mathrm{c})}(f(X), g_\theta(X))] \le \varepsilon. \tag{9}$$

本研究では代理モデル g_θ が θ について微分可能であることを仮定し, 確率的勾配上昇法 (Stochastic Gradient Descent, SGA) に基づくアルゴリズムを提案する. このとき, 目的関数 (6) に現れる $h_{w^{(T)}(\theta)}$ の θ についての勾配は Hypergradient [2, 3] というハイパーパラメータ最適化を対象とする手法を用いることで計算を可能とし, 対数障壁を用いた内点法を用いることで制約 (9) 内で解を探索する.

5 数値実験

数値実験を通して, BODAME–KRR/KR, BODAME–SGD/SGA それぞれの状況設定において, 攻撃者と防御者の分布が異なる場合に有効な防御が可能であることを示す. また, 防御者の代理モデルを構成したのちに固定し, 代理モデルを構成する際に用いなかった攻撃者の入力データを用いても代理モデルの構成に用いた入力データと同様の防御が可能であることを示し, 提案する手法が攻撃者の入力データに関する汎化性能を持つことを数値的に確認する.

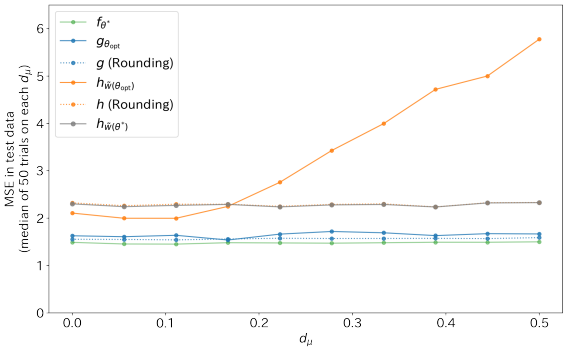

図 2. BODAME–KRR/KR に対する数値実験. 横軸 d_μ は攻撃者の分布と防御者の分布の平均の差を示し, 縦軸は各モデルのテストデータにおける平均二乗誤差 (50 試行の中央値) を示す. 橙の実線で示される提案手法による防御の上で抽出された攻撃者のモデルは分布が離れるにつれて平均二乗誤差が悪化する.

6 関連研究

モデル抽出攻撃に対する防御として防御者の出力の桁数を丸める Rounding [4] や, 勾配に基づく攻撃を防ぐ Maximum Angular Deviation [5] が考案されているが, これらは防御者の真の分布 Q 上での損失関数の最大化を直接的に定式化していない. また, 攻撃者の入力一つ一つごとに防御を行う必要があり, この点は本研究が提案する手法と異なる点である.

参考文献

[1] Satoru Adachi, Satoru Iwata, Yuji Nakatsukasa, and Akiko Takeda. Solving the trust-region subproblem by a generalized eigenvalue problem. *SIAM Journal on Optimization*, Vol. 27, No. 1, pp. 269–291, 2017.

[2] Dougal Maclaurin, David Duvenaud, and Ryan Adams. Gradient-based hyperparameter optimization through reversible learning. In *32nd International Conference on Machine Learning*, pp. 2113–2122, 2015.

[3] Luca Franceschi, Michele Donini, Paolo Frasconi, and Massimiliano Pontil. Forward and reverse gradient-based hyperparameter optimization. In *34th International Conference on Machine Learning*, pp. 1165–1173, 2017.

[4] FlorianTramèr, Fan Zhang, Ari Juels, Michael K Reiter, Thomas Ristenpart. Stealing machine learning models via prediction APIs. In *25th USENIX Security Symposium*, pp. 601–618, 2016.

[5] Tribhuvanesh Orekondy, Bernt Schiele, and Mario Fritz. Prediction poisoning: Towards defenses against DNN model stealing attacks. In *8th International Conference on Learning Representations*, 2020.