SCIS 2022 2022 Symposium on
Cryptography and Information Security
Osaka, Japan & Online, Jan. 18 – 21, 2022
The Institute of Electronics,
Information and Communication Engineers

Theoretical security against adversarial examples on Gaussian Processes

前嶋　啓彰 *　　大塚　玲 †
Hiroaki Maeshima　　Akira Otsuka

キーワード ガウス過程, 敵対的サンプル, AI セキュリティ

あらまし

敵対的サンプル (Adversarial Examples) とは，あるオリジナルの画像データに対して，視覚的に区別が難しいがニューラルネットワークに分類させるとオリジナル画像と異なった分類を出力するような画像である [1]．このような画像は，ニューラルネットワークに対する攻撃となりうる．敵対的サンプルを作成する方法や検知する方法については，様々な手法が提案されているが，その安全性を評価する方法として理論基盤が確立している方法がないのが現状である．

本論文においては，敵対的サンプルの安全性評価の理論基盤として，ガウス過程 (Gaussian Processes)[2] を用いる．ガウス過程は，出力が多変量ガウス分布に従う確率過程であり，推論モデルの構築のためにも利用することができる．ガウス過程はカーネル関数を用いて入力を変換することで，より柔軟なモデルとすることが可能である．

ガウス過程は線形回帰モデルを含み，ニューラルネットワークとの対応もあるなど，さまざまな推論モデルとの関連性があることが知られている．特に，ニューラルネットワークとの関係に関しては，中間層のユニット数を無限としたニューラルネットワークがガウス過程と等価となること，ドロップアウトを含めたニューラルネットワークがガウス過程と等価となること，その際に活性化関数の選択がガウス過程のカーネル関数の選択に対応することが知られている．

ガウス過程を用いることで，ある入力点を回帰・分類した際に，その予測結果がガウス分布に従うものとして，平均と分散を与えることができる．このことより，ある入力点が，その点と最も近い点と異なるクラスに入る確率を求めることができる．

本論文においては，ガウス過程を用いて，敵対的サンプルの攻撃の成功確率の上界を求めた．結果として，特定のカーネル関数を用いた場合に，あるデータセットにおける敵対的サンプルの攻撃の成功確率の上界は，敵対的サンプルの作成時のノイズの大きさ，データセット内の最も近い 2 点の類似度およびガウス過程のカーネル関数のみから求められることを示すことができた．

本研究の成果は，ニューラルネットワークに応用することも可能であると考える．本研究の成果を応用することで，特定の活性化関数と特定の構造を持つニューラルネットワークにおいて，敵対的サンプルの攻撃の成功確率の理論的な上限を与えることができる．このことは，ニューラルネットワークにおける敵対的サンプルの安全性研究の理論基盤となると考える．

参考文献

[1] I. J. Goodfellow, J. Shlens & C. Szegedy, “Explaining and Harnessing Adversarial Examples”, arXiv:1412.6572 [cs, stat]. Available at: http://arxiv.org/abs/1412.6572

[2] C. E. Rasmussen and C. K. I. Williams, “Gaussian processes for machine learning.” Cambridge, Mass: MIT Press, 2006.

* 情報セキュリティ大学院大学; NTT テクノクロス株式会社, 横浜, Institute of Information Security; NTT TechnoCross Corporation, Yokohama
† 情報セキュリティ大学院大学, 横浜, Institute of Information Security, Yokohama