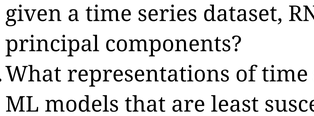
The Dilemma Between Data Transformations and Adversarial Robustness for Time Series Application Systems

Sheila Alemany and Dr. Niki Pissinou
School of Computing and Information Sciences, Florida International University

Background

- Current adversarial machine learning research continuously approaches similar conclusions: **the vulnerability of ML models is highly correlated to how the data is represented** [1, 2].
  - For instance, [3] presents theoretical evidence that datasets with higher intrinsic dimensionality facilitate adversarial attacks.



E.g., Adding color increases images' intrinsic dimensionality. "Harder-to-detect" changes can be made with RGB pixels.

- Adjacently, the presentation of data to a learning model impacts its performance.
  - For example, we have seen this through the use of dimensionality reduction over the years to increase model accuracy.
- Adversarial research has focused primarily on classification problems in computer vision applications.

Problem

ML practitioners in time series fields may be **unknowingly** making more vulnerable models with the use of certain data transformations.

Research Aims

We designed our experimentation to address the following:

1. Could data transformations contribute to any adversary's ability to more easily construct adversarial examples?
2. Is the dimensionality reduction technique, PCA, consistent as a strategy to increase robustness when given a time series dataset, RNN, and varying selected principal components?
3. What representations of time series data contribute to ML models that are least susceptible to adversarial examples?

We explored the effect of 3 classes of linear transformations:

- Dimensionality reduction (e.g., PCA)
- Feature selection (e.g., low variance, random forest selection)
- Trend extraction (e.g., candlestick charting, EMA)

Evaluation Results

Dataset: MHealth Dataset [4] with body motion and vital signs from 22 sensors
Adversarial Attack: Carlini & Wagner l_∞attack [5]
Neural Network: RNN with LSTM layers and average baseline performance over 90%

Data Transformation	Feature Count	Benign Accuracy	Distance (l_∞)	Δ in Robustness
Baseline	22	97.93%	0.51	-
PCA 50%	11	96.71%	0.40	↑ 24.39%
PCA 81%	18	98.80%	0.76	↓ 43.90%
PCA 27%	6	95.00%	0.34	↓ 60.98%
Random Forest	9	96.11%	0.13	↓ 31.71%
Low Variance	11	91.32%	0.15	↓ 65.85%
Candlesticks	22	92.78%	0.11	↓ 60.98%
EMA	22	96.48%	0.51	↓ 7.32%

The most robust RNN used PCA with only half of the principal components and is **only** a consistent defense against adversarial examples if the number of selected principle components approximates the data's intrinsic dimension.

References

[1] Shafahi, A. et al. 2019. Are adversarial examples inevitable? In International Conference on Learning Representations.

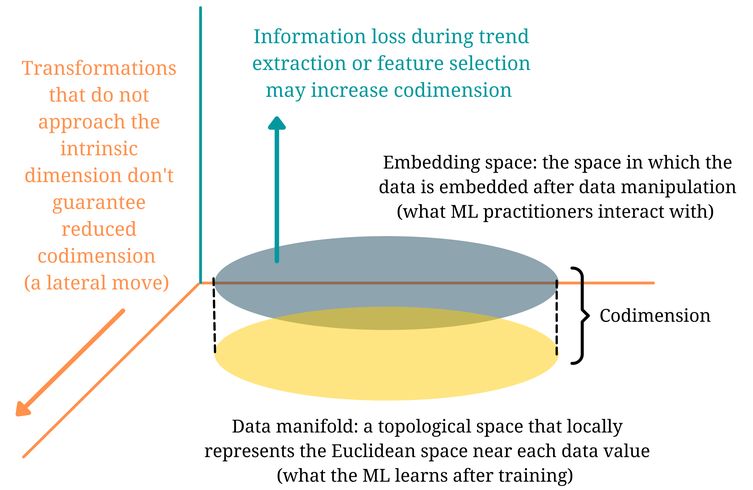
[2] Ilyas, A.; et al. 2019. Adversarial examples are not bugs, they are features. Advances in Neural Information Processing Systems 32.

[3] Amsaleg, L.; et al. 2020. High Intrinsic Dimensionality Facilitates Adversarial Attack: Theoretical Evidence. IEEE Transactions on Information Forensics and Security, 16: 854–865.

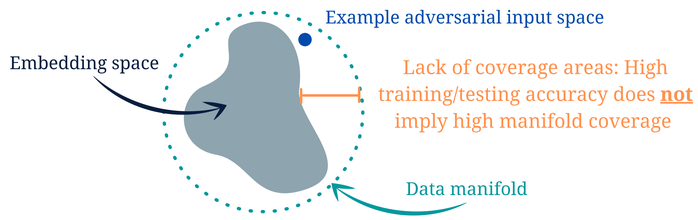
[4] Banos, O.; et al. 2015. mDurance: a novel mobile health system to support trunk endurance assessment. Sensors, 15(6): 13159–13183.

[5] Carlini, N.; and Wagner, D. 2017. Towards evaluating the robustness of neural networks. In 2017 IEEE Symposium on Security and Privacy (SP), 39–57. IEEE.

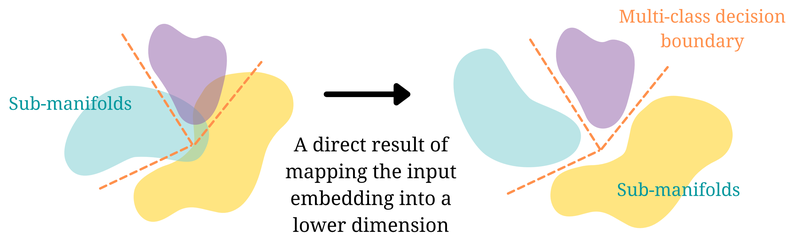
Discussion



- Candlesticks charting contributes to the most vulnerable ML models due to information loss which increases codimension.
- Feature selection techniques may increase codimension and lack manifold coverage resulting in lots of adversarial input spaces.



- PCA creates more well-defined sub-manifolds for each class such that an adversary requires higher perturbation to "trick" an ML model.
  - This is not the case for feature selection and trend extraction since there is no mapping to a lower dimension.



To avoid introducing additional vulnerabilities in ML pipelines, ML practitioners must observe and understand the particular dataset's intrinsic characteristics and ensure any transformation does not stray from the intrinsic dimension.

This work was supported by the Knight Foundation School of Computing and Information Sciences at Florida International University, the National GEM Consortium, and the NSF Graduate Research Fellowship program.