Adversarial Robustness and Anomaly Detection in Financial ML Systems

Introduction

In financial machine learning systems, robustness against adversarial inputs and the ability to detect anomalies are crucial for safeguarding assets and maintaining market integrity. Adversarial robustness focuses on defending models from maliciously crafted inputs intended to cause mispredictions, while anomaly (out-of-distribution, OOD) detection identifies inputs that differ significantly from the model's training data distribution.

Examples include:

- Fraudulent transactions designed to bypass detection.
- Abnormal trading signals indicating manipulation.
- Sudden market regime shifts.

OOD Detection Algorithm

A practical approach for detecting OOD anomalies involves using the Mahalanobis distance or autoencoder-based reconstruction error.

Mahalanobis Distance

Mahalanobis distance is defined as:

$$D_M(x) = \sqrt{(x-\mu)^\top \Sigma^{-1} (x-\mu)},$$

where μ and Σ are the mean vector and covariance matrix estimated from in-distribution data.

A high Mahalanobis distance indicates that a point is far from the learned distribution, suggesting it might be anomalous.

Autoencoder Reconstruction Error

An autoencoder is a neural network trained to compress and then reconstruct input data. The reconstruction error:

$$E(x) = \|x - \hat{x}\|^2$$

is small for in-distribution data. Large errors suggest OOD inputs.

Advantages

- Mahalanobis distance provides an analytical score without training a separate neural model.
- Autoencoders can capture complex, non-linear structures in financial data.

Conclusion

Combining adversarial robustness and OOD detection strengthens financial ML models against attacks and unexpected data shifts, improving trust and regulatory compliance.