Cost Function Analysis under Adversarial Attacks on Logistic Regression

Abstract

This project investigates the impact of adversarial attacks on the cost function and predictions of logistic regression models. We implement the Fast Gradient Sign Method (FGSM) to generate adversarial examples and analyze their influence on model performance. Furthermore, we evaluate mitigation strategies including L2 regularization and adversarial training to improve robustness.

1. Introduction

Machine learning models, including logistic regression, are vulnerable to adversarial examples--inputs subtly perturbed to mislead predictions. This study focuses on analyzing how these perturbations affect the cost function and model accuracy. Additionally, we propose strategies to reduce vulnerability.

2. Methodology

2.1 Dataset

- We used the MNIST dataset, reduced to a binary classification task (digit "0" vs others).
- Data is standardized using `StandardScaler`.

2.2 Baseline Logistic Regression

- A logistic regression model is trained using scikit-learn on clean, standardized data.

2.3 Adversarial Attack Implementation

- FGSM (Fast Gradient Sign Method) is implemented using PyTorch.
- The adversarial input is calculated as: x_adv = x + epsilon sign(_x J(, x, y))
- The model is re-implemented in PyTorch to obtain gradient information.

2.4 Evaluation Metrics

Cost Function Analysis under Adversarial Attacks on Logistic Regression

- Cost Function: Binary Cross Entropy Loss (BCE).
- Prediction Accuracy: Classification accuracy on clean and adversarial data.

3. Results and Visualization

3.1 Cost Function Comparison

- Average BCE loss (Clean): ~0.12
- Average BCE loss (Adversarial): ~0.35

3.2 Prediction Accuracy

- Clean Accuracy: 97%
- Adversarial Accuracy: 65%

3.3 Visualizations

- Bar plots comparing BCE loss (clean vs adversarial).
- Histograms showing cost distributions.
- Scatter plot: clean vs adversarial prediction probabilities.
- PCA-based heatmap of cost surface.

4. Mitigation Strategies

4.1 L2 Regularization

- Training with increased regularization (C=0.1) reduced sensitivity to perturbations.

4.2 Adversarial Training

Cost Function Analysis under Adversarial Attacks on Logistic Regression

- Model retrained with both clean and adversarial data.
- Showed increased robustness: Adversarial accuracy improved to ~78%.

5. Conclusion

Adversarial attacks significantly distort the cost function and reduce model accuracy. Visualization of loss surfaces and prediction shifts reveal how small perturbations can result in high-cost misclassifications. Incorporating regularization and adversarial training provides effective mitigation strategies.

6. Future Work

- Extend to multi-class logistic regression.
- Evaluate stronger attacks (PGD, DeepFool).
- Implement robust optimization frameworks for theoretical guarantees.

References

- Goodfellow et al. "Explaining and Harnessing Adversarial Examples"
- Scikit-learn Documentation
- PyTorch FGSM Tutorial