작성자: 윤정현*

작성 날짜: 2024-08-25

GAN 실습 보고서

목 차

* 중앙대학교 산업보안학과

Ⅰ. 서 론

본 보고서는 Membership Inference Attack 실습과 위 3가지 Robustness 향상기법 실습을 통해 결과를 얻고, 수치 분석을 통해서 각 방식의 효과성을 평가하고자 합니다.

Ⅱ. Membership Inference Attack 실습 결과

실습 종류	정확도	정밀도	재현율	F1 점수
전체 Shadow Models	0.1384	0.1022	0.9760	0.1850
최고 Shadow Model	0.1437	0.1026	0.9741	0.1856

Membership Inference 공격의 성능을 높이는 방법을 찾기 위해서 다수의 Shadow Model을 학습 시킨 Attack Model하고, Shadow Model 3개를 학습 시키고 그중에 제일 우수한 성능을 가진 Shadow Model로 Attack Model을 학습시키고 성능을 비교해 보았다. 실험 결과에 따르면, Shadow Model 3개중에 우수한 성능을 가진 Shadow Model로 Attack Model을 학습 시킨게 성능이 더 우수했다.(정확도, 정밀도, F1 점수 기준)

Ⅲ. Robustness 향상기법 실습 결과 및 비교

a. Data Augemented

데이터 증강은 기존 데이터셋을 변형 또는 증강하여 학습 데이터의 양을 늘리는 기술입니다. 이는 모델이 다양한 데이터 패턴을 학습하게 하여, 과적합(overfitting)을 방지하고 일반화(generalization) 성능을 향상시킵니다.

실습 종류	정확도	정밀도	재현율	F1 점수
개선 전	0.4987	0.0991	0.4950	0.1652
개선 후	0.5031	0.0995	0.4920	0.1656

결과에 따르면 데이터 증강을 적용한 이후로 공격 모델의 성능이 재현율을 제외하고 모든 면에 성능이 좋아진 모습을 볼수가있다. 통계적 오차일수도있으나, 데이터 증강이 Robustness 향상하는데 이 실습에서는 오히려 역효과를 일으키는 것을 볼수가 있다.

b. Adverserial Regularization

적대적 정규화는 모델 학습 과정에서 악의적인 입력(Adversarial Examples)에 대해 모델의 강건성을 높이는 정규화 기술입니다. 이는 모델이 의도적으로 변형된 입력에 대해서도 정확한 예측을 하도록 하여 모델의 강건성을 향상시킵니다.

실습 종류	정확도	정밀도	재현율	F1 점수
노이즈 없음	0.1481	0.1017	0.9581	0.1839
노이즈 추가	0.1536	0.1017	0.9511	0.1838

결과에 따르면 적대적 정규화를 적용 시키면 공격 모델의 성능이 떨어지는 것을 보아서, 적대적 정규화가 멤버십 추론 공격을 방어하는데 효율적이라는 것을 볼수가 있다.

c. Differencial Privacy

차분 프라이버시는 모델 학습 과정에 노이즈를 추가하여 모델의 출력이 특정 데이터 포인트에 의존하지 않도록 합니다. 이는 사용된 데이터와 그렇지 않은 데이터의 confidence vector를 모호하게 만들어 멤버십 추론 공격을 어렵게 합니다.

실습 종류	정확도	정밀도	재현율	F1 점수
노이즈 없음	0.5224	0.1037	0.4930	0.1714
노이즈 추가	0.5178	0.1029	0.4940	0.1703

결과에 따르면 차분 프라이버시 적용 시키면 공격 모델의 성능이 떨어지는 것을 보아서, 차분 프라이버시가 멤버십 추론 공격을 방어하는데 F1 점수 기준으로는 데이터 증강, 적대적 정규화 중에 제일 우수한 성과를 낸다는 것을 관찰할 수가 있다.

Ⅳ. 논문초록

【초록】

머신러닝 모델의 멤버십 추론 공격과 방어 기법의 효과 분석

윤정현[1)]

머신러닝 모델의 프라이버시 취약점인 멤버십 추론 공격(Membership Inference Attack, 약자: MIA)의 효과와 이에 대한 다양한 방어 기법들의 효과성을 실험적으로 비교 분석하였습니다. MIA의 성능을 평가하고, 데이터 증강(Data Augmentation), 적대적 정규화(Adversarial Regularization), 차분 프라이버시(Differential Privacy) 세 가지 방법을 적용하여 모델의 강건성 향상 정도를 평가하였습니다.

MIA 실험 결과, 전체 Shadow Models에서 정확도 0.1384, 정밀도 0.1022, 재현율 0.9760, F1 점수 0.1850을 기록했다. 최고 성능의 Shadow Model은 정확도 0.1437, F1 점수 0.1856으로 약간 더 높은 성능을 보였다. 이는 MIA가 높은 재현율을 가지고 있어 훈련 데이터셋의 멤버를 식별하는데 효과적이고, 많은 Shadow Models 학습하는것보다 최고 성능의 Shadow Model 3개 중 제일 우수한 성능을 가진 모델을 선정하고 그 Shadow Model로 Attack Model을 생성하는게 Attack Model의 성능을 높이는데 더 효과적이라는 것을 알수가 있다.

그리고 방어 기법 적용 결과, 데이터 증강 기법은 정확도를 0.4987에서 0.5031로, F1 점수를 0.1652에서 0.1656으로 소폭 개선 효과를 보였다. 적대적 정규화의 경우, 노이즈 추가 후 정확도가 0.1481에서 0.1536으로 증가했으나 F1 점수는 0.1839에서 0.1838로 미세하게 감소했다. 차분 프라이버시 적용 시에는 정확도가 0.5224에서 0.5178로, F1 점수가 0.1714에서 0.1703으로 소폭 하락하였다.

이러한 결과는 MIA가 모델의 프라이버시에 실질적인 위협이 될 수 있음을 보여주며, 동시에 현재의 방어 기법들이 MIA에 대한 강건성 향상에 어느정도 효과를 가짐을 시사한다.

키워드: 멤버십 추론 공격, 데이터 증강, 적대적 정규화, 차분 프라이버시, 머신러닝 보안

---

1) Undergraduate degree course, Chung-Ang University