Adversarial Machine Learning: A review of the "Adversarial Robustness Toolbox (ART)"

Habtamu Desalegn Woldeyohannes

877159

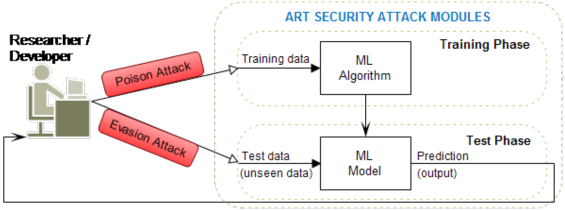
Ca' Foscari University of Venice
Department of Environmental Sciences, Informatics and Statistics

Master's Thesis
10th May 2021

Supervisor: Cloudio Lucchesse

Adversarial Robustness Toolbox (ART) V1.5.1 [1]

- Open source ML Security Library developed by IBM. (current version 1.6.1)
- Written in Python, ML Framework/Library agnostic toolbox: TF, Pytorch, MXNet, Scikit-learn, LightGBM, ...
- Consist of SOTA Adversarial Attacks, Defences and Model Robustness Metrics Algorithms. And supports all data types: Tabular, Images, Video, Audio, ...
- ART Adversarial Attacks
    - Security Attacks on ML (Poisoning Attacks and Evasion Attacks)
    - Privacy Attacks on ML (Inference Attacks and Extraction Attacks))



[1] M.-I. Nicolae, M. Sinn, M. N. Tran, B. Buesser, A. Rawat, M. Wistuba, V. Zantedeschi, N. Baracaldo, B. Chen, H. Ludwig, I. Molloy, and B. Edwards. Adversarial robustness toolbox v1.2.0. CoRR, 2018

ML Models: Attack unaware models

Scikit-learn models: SVM, DT & RF LightGBM model: GBDT

Adult Census Income dataset:
Experimental result on 9k original examples. (best results in boldface)

Baseline models	Precision	Recall	F_1 score	MCC
SVM	0.635	0.325	0.430	0.346
DT	0.609	**0.760**	0.676	0.565
RF	0.689	0.714	0.701	0.605
GBDT	**0.719**	0.707	**0.713**	**0.624**

Experimental result on 1500 original examples.

Baseline models	Precision	Recall	F_1 score	MCC
SVM	0.628	0.350	0.449	0.363
DT	0.592	**0.781**	0.673	0.568
RF	0.667	0.723	0.694	0.599
GBDT	**0.694**	0.714	**0.704**	**0.615**

MNIST handwritten digit database
Experimental result on 14k MNIST test data.

Baseline models	Accuracy
SVM	0.9957
DT	0.9749
RF	0.9922
GBDT	**0.9959**

Experimental result on MNIST 100 original examples.

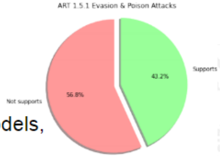
Baseline models	Accuracy
SVM	0.994
DT	0.972
RF	0.986
GBDT	**0.998**

Sklearn.tree. - **DecisionTreeClassifier** - DecisionTreeRegressor - ExtraTreeClassifier	Sklearn.ensemble • AdaBoostClassifier • AdaBoostClassifier • ExtraTreesClassifier • GradientBoostingClassifier • **RandomForestClassifier**	sklearn.linear_model.LogisticRegression sklearn.naive_bayes.GaussianNB **sklearn.svm.SVC,** **sklearn.svm.LinearSVC** **lightgbm.Booster** ~~lightgbm.sklearn~~

ART Adversarial Attack Algorithms

ART 1.5.1: Statistics and Issues

of Issues: 21 Ref: Table 5.1



- ART 1.5.1 library has a total of 37 security attacks on ML Models, including 32 evasion attacks and 5 data poisoning attacks.

Evasion Attacks

Estimator (Model) Issues
- AutoAttack
- Auto Projected Gradient Descent Attack
- Threshold Attack
- High Confidence Low Uncertainty Attack
- Pixel Attack
- Spatial Transformation Attack
- Robust DPatch Attack
- ShapeShifter Attack
- Adversarial Patch Attack 'DPatch'
- Frame Saliency Attack
- Adversarial Patch Attack
- Feature Adversaries Attack
- Brendel & Bethge Adversarial Attack

Evasion Attacks

Object has no attribute issues
- Wasserstein Attack
- Simple Black-box Adversarial (SimBA)

Unrecognized input dimension issues
- Square Attack
- Shadow Attack

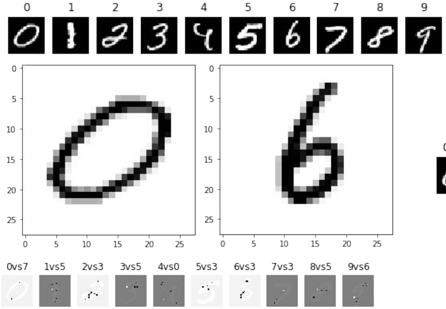
Poison Attacks
- Adversarial Embedding Attack
- Clean-Label Backdoor Attack
- Backdoor Attacks
- Feature Collision Poisoning Attack

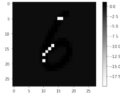
(1) Decision tree-based attack

ART Decision tree Attack (2016)

ART Decision tree Attack '16

- **Traversing the learned tree structure**



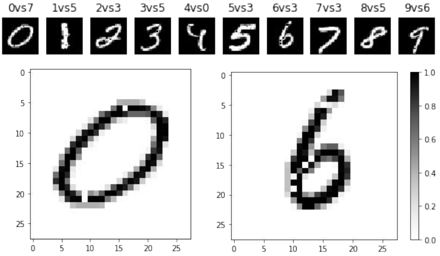


Offset=20

Table 4.10: Experimental results using ART DecisionTree attack against decision trees on census.

Test data	Precision	Recall	F1 score	MCC	Fooling rate(%)
Original	**0.592**	**0.781**	**0.673**	**0.568**	-
Adversarial	0.082	0.214	0.119	-0.478	**92.38**

Untargeted Attack **Offset=0.001**



Evasion Attacks

(2) Gradient-based Attacks

White-Box Attacks

Algorithms		Objec.	Support	Distance Metrics
Fast Gradient Sign Method **(FGSM)**	'14	**T, U**	**SVM**	FGM(L1),FGM(L2),FGSM(L∞)
Basic Iterative Method **(BIM)**	'16	**T, U**	**SVM**	BIM(L∞)
Projected Gradient Descent **(PGD)**	'17	**T, U**	**SVM**	PGD(L1),PGD(L2),PGD(L∞)
Carlini & Wagner Attack **(C&W)**	'16	**T, U**	**SVM**	C&W(L2), C&W(L∞)
Elastic-Net Attack **(EAD)**	'17	**T, U**	**SVM**	EAD(L2), EAD(EN)
Jacobian Saliency Map Attack **(JSMA)**	'16	**U**	**SVM**	JSMA(L0)
NewtonFool	'17	**U**	**SVM**	NewtonFool(L2)
Virtual Adversarial Method **(VAT)**	'15	**U**	**SVM**	VAT(L2)
DeepFool	'15	**U**	**SVM**	DeepFool(L2)
Universal Perturbations **(UP/TUP)**	'16/'19	**U,T**	**SVM**	

Maximum Loss attacks

Regularized attacks

Minimum distance attacks

(2) Gradient-based Attacks

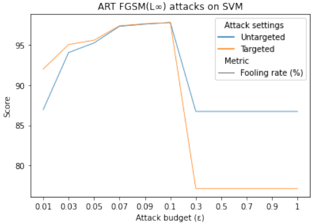
ART FGSM, BIM and PGD Attacks

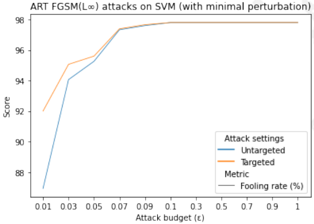
Experimental results on 1500 census adversarial examples in targeted setting.

Table 4.11: Parameters for the gradient-based attack algorithms

Attack Algorithm	Parameters
FGM(ℓ_1), FGM(ℓ_2), FGSM(ℓ_∞)	ε=0.1, 0.3, 0.5, 0.7, 0.9, 1.0; ε_{step}=0.1; minimal perturbation=True
BIM(ℓ_∞)	ε=0.1, 0.3, 0.5, 0.7, 0.9, 1.0; ε_{step}=0.1; maximum iteration=2
PGD(ℓ_1), PGD(ℓ_2), PGD(ℓ_∞)	ε=0.1, 0.3, 0.5, 0.7, 0.9, 1.0; ε_{step}=0.1; maximum iteration=2

		FGSM					BIM					PGD				
	ε	Prec.	Rec.	F1	MCC	Fooling rate %	Prec.	Rec.	F1	MCC	Fooling rate %	Prec.	Rec.	F1	MCC	Fooling rate %
L1	0.1	0.06	0.216	0.094	-0.859	95.06	-	-	-	-	-	0.054	0.192	0.084	-0.874	**95.6**
	0.3	0.054	0.192	0.084	-0.874	95.6	-	-	-	-	-	0.054	0.192	0.084	-0.874	**95.6**
	0.5	0.028	0.096	0.043	0.022	97.8	-	-	-	-	-	0.054	0.192	0.084	-0.874	**95.6**
	0.7	0.009	0.032	0.014	-0.979	**99.26**	-	-	-	-	-	0.054	0.192	0.084	-0.874	**95.6**
	0.9	0.009	0.032	0.014	-0.979	**99.26**	-	-	-	-	-	0.054	0.192	0.084	-0.874	**95.6**
	1	0.009	0.032	0.014	-0.979	**99.26**	-	-	-	-	-	0.054	0.192	0.084	-0.874	**95.6**
L2	0.1	0.057	0.204	0.089	-0.866	95.33	-	-	-	-	-	0.033	0.114	0.051	-0.926	**97.39**
	0.3	0.033	0.114	0.051	-0.926	97.39	-	-	-	-	-	0.033	0.114	0.051	-0.926	**97.39**
	0.5	0.009	0.032	0.014	-0.979	99.26	-	-	-	-	-	0.033	0.114	0.051	-0.926	**97.39**
	0.7	0.009	0.032	0.014	-0.979	99.26	-	-	-	-	-	0.033	0.114	0.051	-0.926	**97.39**
	0.9	0.009	0.032	0.014	-0.979	99.26	-	-	-	-	-	0.033	0.114	0.051	-0.926	**97.39**
	1	0.009	0.029	0.014	-0.981	**99.33**	-	-	-	-	-	0.033	0.114	0.051	-0.926	**97.39**
L∞	0.1	0.028	0.096	0.043	-0.937	**97.8**	0.028	0.096	0.043	-0.937	**97.8**	0.028	0.096	0.043	-0.937	**97.8**
	0.3	0.028	0.096	0.043	-0.937	**97.8**	0.028	0.096	0.043	-0.937	**97.8**	0.028	0.096	0.043	-0.937	**97.8**
	0.5	0.028	0.096	0.043	-0.937	**97.8**	0.028	0.096	0.043	-0.937	**97.8**	0.028	0.096	0.043	-0.937	**97.8**
	0.7	0.028	0.096	0.043	-0.937	**97.8**	0.028	0.096	0.043	-0.937	**97.8**	0.028	0.096	0.043	-0.937	**97.8**
	0.9	0.028	0.096	0.043	-0.937	**97.8**	0.028	0.096	0.043	-0.937	**97.8**	0.028	0.096	0.043	-0.937	**97.8**
	1	0.028	0.096	0.043	-0.937	**97.8**	0.028	0.096	0.043	-0.937	**97.8**	0.028	0.096	0.043	-0.937	**97.8**





(2) Gradient-based Attacks

ART FGSM(L∞), BIM(L∞) and PGD(L∞) Attacks

Experimental results on 100 MNIST adversarial examples in targeted and untargeted settings.

Table 4.14: MNIST: Experimental results on 100 adversarial examples FGSM(ℓ_∞), BIM(ℓ_∞), and PGD(ℓ_∞) attacks bounded with different values of ε in the targeted and untargeted settings (best attack success rate in boldface).

Objective	Epsilon(ε)	FGSM(ℓ_∞) Acc.	FGSM(ℓ_∞) Fooling rate(%)	BIM(ℓ_∞) Acc.	BIM(ℓ_∞) Fooling rate(%)	PGD(ℓ_∞) Acc.	PGD(ℓ_∞) Fooling rate(%)
Targeted	0.1	0.97	0	0.97	0	0.97	0
	0.2	0.85	5	0.83	**4**	0.83	3
	0.3	0.85	5	0.83	4	0.72	14
	0.4	0.18	**75**	0.83	4	0.61	**21**
Untargeted	0.1	0.89	12	0.88	13	0.88	13
	0.2	0.59	42	0.56	**45**	0.55	46
	0.3	0.59	42	0.56	45	0.38	63
	0.4	0.1	**91**	0.56	45	0.28	**73**

Figure: FGSM (L∞) ε=0.4 with untargeted attack

(2) Gradient-based Attacks

ART FGSM(L∞), BIM(L∞) and PGD(L∞) Attacks (Cont.)

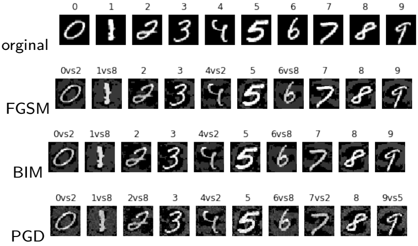


Fig: Image of original and perturbed images generated by ART FGSM, BIM, and PGD attacks under ℓ_∞ norm bounded by $\epsilon = 0.3$ on MNIST and predicted class labels by SVM model.

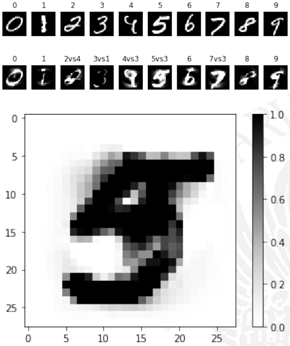


Fig: ART FGSM attacks under ℓ_2 norm bounded by $\varepsilon = 10$ on MNIST.

(2) Gradient-based Attacks
ART FGSM(L∞) (Cont.)

Increasing ε value will increase misclassification, but the adversarial image is very different from the original image.

Figure: $\varepsilon = 1$,ε_step=0.1, 98%

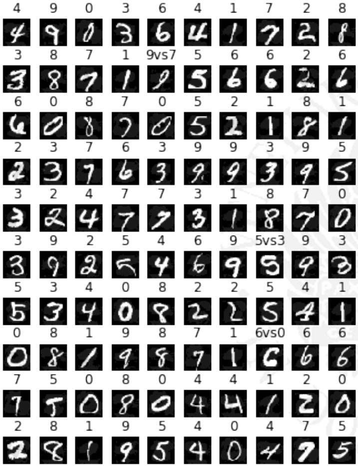


Figure: $\varepsilon = 0.1$,ε_step=0.01, 1%

(2) Gradient-based Attacks

ART FGSM(L∞) with $\varepsilon = 0.3$ (Cont.)

Figure: Untargeted setting: Attack success rate of 42%

Figure: Targeted setting: Attack success rate of 5%

(2) Gradient-based Attacks

Summary

ART Attack Algorithms	Targeted				Untargeted			
	SVM	DT	RF	GBDT	SVM	DT	RF	GBDT
TABULAR DATA TYPE								
FGM(ℓ_1)	++	−−	−−	−−	++	−−	−−	−−
FGM(ℓ_2)	++	−−	−−	−−	++	−−	−−	−−
FGSM(ℓ_∞)	++	−−	−−	−−	++	−−	−−	−−
BIM(ℓ_∞)	++	−−	−−	−−	++	−−	−−	−−
PGD(ℓ_1)	++	−−	−−	−−	++	−−	−−	−−
PGD(ℓ_2)	++	−−	−−	−−	++	−−	−−	−−
PGD(ℓ_∞)	++	−−	−−	−−	++	−−	−−	−−
UP(ℓ_∞)	−−	−−	−−	−−	+	−−	−−	−−
UAP(ℓ_∞)	+	−−	−−	−−	−−	−−	−−	−−
JSMA(ℓ_0)	−−	−−	−−	−−	++	−−	−−	−−
C&W(ℓ_2)	++	−−	−−	−−	++	−−	−−	−−
C&W(ℓ_∞)	++	−−	−−	−−	+	−−	−−	−−
EAD(EN)	−−	−−	−−	−−	++	−−	−−	−−
NewtonFool(ℓ_2)	−−	−−	−−	−−	−−	−−	−−	−−
DeepFool(ℓ_2)	−−	−−	−−	−−	−−	−−	−−	−−
VAT(ℓ_2)	−−	−−	−−	−−	−	−−	−−	−−
IMAGE DATA TYPE								
FGM(ℓ_1)	−−	−−	−−	−−	−−	−−	−−	−−
FGM(ℓ_2)	−−	−−	−−	−−	−−	−−	−−	−−
FGSM(ℓ_∞)	++	−−	−−	−−	++	−−	−−	−−
BIM(ℓ_∞)	−	−−	−−	−−	+	−−	−−	−−
PGD(ℓ_1)	−−	−−	−−	−−	−−	−−	−−	−−
PGD(ℓ_2)	−−	−−	−−	−−	−−	−−	−−	−−
PGD(ℓ_∞)	−	−−	−−	−−	+	−−	−−	−−
UP(ℓ_∞)	−−	−−	−−	−−	+	−−	−−	−−
UAP(ℓ_∞)	+	−−	−−	−−	−−	−−	−−	−−
JSMA(ℓ_0)	−−	−−	−−	−−	++	−−	−−	−−
C&W(ℓ_2)	+	−−	−−	−−	++	−−	−−	−−
C&W(ℓ_∞)	++	−−	−−	−−	++	−−	−−	−−
EAD(EN)	−−	−−	−−	−−	**++**	−−	−−	−−
NewtonFool(ℓ_2)	−−	−−	−−	−−	−	−−	−−	−−
DeepFool(ℓ_2)	−−	−−	−−	−−	−	−−	−−	−−
VAT(ℓ_2)	−−	−−	−−	−−	−	−−	−−	−−

Attack Algorithm	Parameters
FGM(ℓ_1), FGM(ℓ_2), FGSM(ℓ_∞)	ε=0.1, 0.3, 0.5, 0.7, 0.9, 1.0; ε_{step}=0.1; minimal perturbation=True
BIM(ℓ_∞)	ε=0.1, 0.3, 0.5, 0.7, 0.9, 1.0; ε_{step}=0.1; maximum iteration=2
PGD(ℓ_1), PGD(ℓ_2), PGD(ℓ_∞)	ε=0.1, 0.3, 0.5, 0.7, 0.9, 1.0; ε_{step}=0.1; maximum iteration=2
UP(ℓ_1), UP(ℓ_2), UP(ℓ_∞)	ε=0.1, 0.3, 0.5, 0.7, 0.9, 1.0; ε_{step}=0.1; maximum iteration=1
UAP(ℓ_1), UAP(ℓ_2), UAP(ℓ_∞)	ε=0.1, 0.3, 0.5, 0.7, 0.9, 1.0; ε_{step}=0.1; maximum iteration=1
C&W(ℓ_2), C&W(ℓ_∞)	ε=0.1, 0.3, 0.5, 0.7, 0.9, 1.0; ε_{step}=0.1; maximum iteration=2
JSMA(ℓ_0)	θ=0.1, 0.3, 0.5, 0.7, 0.9, 1.0; γ=0.1; maximum iteration=2
NewtonFool(ℓ_2)	η=0.1, 0.3, 0.5, 0.7, 0.9, 1.0; maximum iteration=2
DeepFool(ℓ_2)	ε=0.1, 0.3, 0.5, 0.7, 0.9, 1.0; nb_grads=10; maximum iteration=2
EAD(ℓ_1), EAD(ℓ_2), EAD(EN)	ε=0.1, 0.3, 0.5, 0.7, 0.9, 1.0; maximum iteration=2
VAT(ℓ_2)	ε=0.1, 0.3, 0.5, 0.7, 0.9, 1.0; finite_diff=$1e-6$; maximum iteration=2

(3) Score-based Attack
ART Zeroth-order optimization ZOO(L2) (2017)

Model	eps. (ε)	Targeted Pre.	Rec.	F1	MCC	Fooling rate(%)	Untargeted Pre.	Rec.	F1	MCC	Fooling rate(%)
SVM	0.1	0.094	0.35	0.148	-0.767	**92**	0.229	1	0.373	0	**86.73**
	0.3	0.094	0.35	0.148	-0.767	92	0.229	1	0.373	0	86.73
	0.5	0.094	0.35	0.148	-0.767	92	0.229	1	0.373	0	86.73
	0.7	0.094	0.35	0.148	-0.767	92	0.229	1	0.373	0	86.73
	0.9	0.094	0.35	0.148	-0.767	92	0.229	1	0.373	0	86.73
DT	0.1	0.089	0.21	0.125	-0.361	**67.26**	0.194	0.397	0.261	-0.078	**65.66**
	0.3	0.089	0.21	0.125	-0.358	67	0.195	0.397	0.262	-0.075	65.4
	0.5	0.09	0.21	0.126	-0.353	66.6	0.195	0.394	0.261	-0.073	64.93
	0.7	0.091	0.21	0.127	-0.343	65.73	0.198	0.391	0.263	-0.066	64
	0.9	0.107	0.21	0.142	-0.259	57.93	0.234	0.379	0.289	0.009	55.93
RF	0.1	0.024	0.044	0.227	-0.405	**61.8**	0.181	0.306	0.227	-0.09	**61.46**
	0.3	0.028	0.044	0.241	-0.355	56.46	0.204	0.294	0.241	-0.042	55.86
	0.5	0.028	0.044	0.239	-0.355	56.46	0.202	0.292	0.239	-0.045	55.8
	0.7	0.028	0.044	0.24	-0.353	56.2	0.204	0.292	0.24	-0.042	55.53
	0.9	0.028	0.044	0.24	-0.351	56.06	0.204	0.292	0.24	-0.04	55.4
GBDT	0.1	0.162	0.114	0.134	-0.07	33.73	0.212	0.099	0.135	-0.013	**30.33**
	0.3	0.162	0.114	0.134	-0.07	33.73	0.316	0.146	0.200	0.072	27.93
	0.5	0.154	0.111	0.129	-0.078	34.2	0.254	0.105	0.149	0.019	28.59
	0.7	0.178	0.134	0.153	-0.055	33.93	0.287	0.125	0.174	0.046	28.46
	0.9	0.152	0.111	0.128	-0.082	**34.46**	0.31	0.143	0.196	0.067	28.06

Census: Experimental results on 1500 adversarial examples generated by ART ZOO(L2) attacks with different values of ε in the targeted and untargeted settings (best attack success rate in boldface).

Attack Algorithm	Parameters
ZOO(ℓ_2)	Step size ($\varepsilon = [0.1, 0.3, 0.5, 0.7, 0.9]$) Maximum number of iterations (max_iter=2) confidence=0 learning_rate=0.01 binary_search_steps=2 initial_const=0.001 abort_early=True batch_size=1

(3) Score-based Attack

ART Zeroth-order optimization ZOO(L2) (cont...)

Model	eps. (ε)	Targeted Avg. Time (per attack)	Targeted Acc.	Targeted Fooling rate(%)	Untargeted Avg. Time (per attack)	Untargeted Acc.	Untargeted Fooling rate(%)
SVM	0.1	**91.29 sec**	0.98	0	**91.41 sec**	0.97	3
	0.3	**72.95 sec**	0.98	0	**72.98 sec**	0.97	3
	0.5	**74.07 sec**	0.98	0	**74.06 sec**	0.97	3
	0.7	**76.59 sec**	0.98	0	**76.59 sec**	0.97	3
	0.9	**79.72 sec**	0.98	0	**79.72 sec**	0.97	3
DT	0.1	2.69 sec	0.65	26	3.83 sec	0.23	**83**
	0.3	3.69 sec	0.66	25	3.76 sec	0.27	76
	0.5	3.72 sec	0.68	23	3.75 sec	0.27	74
	0.7	3.61 sec	0.68	23	3.87 sec	0.27	74
	0.9	3.56 sec	0.75	16	3.42 sec	0.3	71
RF	0.1	2.87 sec	0.67	28.9	2.86 sec	0.26	**76**
	0.3	2.88 sec	0.73	23	3 sec	0.3	72
	0.5	2.85 sec	0.75	21	3 sec	0.31	71
	0.7	2.85 sec	0.74	21	2.88 sec	0.3	72
	0.9	2.88 sec	0.72	23	2.87 sec	0.34	68
GBDT	0.1	7.46 sec	0.95	3	7.19 sec	0.82	**19**
	0.3	7.37 sec	0.95	3	6.96 sec	0.86	15
	0.5	7.35 sec	0.94	4	6.48 sec	0.9	11
	0.7	7.42 sec	0.95	3	5 sec	0.84	16
	0.9	7.29 sec	0.95	3	4.95 sec	0.8	21

MNIST: Experimental results on 100 adversarial examples generated by ART ZOO(L2) attacks with different values of ε in the targeted and untargeted settings (best attack success rate in boldface).

Fig: Perturbed images generated by ART ZOO(L2) attack against DT on MNIST with 10 iterations.

Fig: Perturbed images generated by ART ZOO(L2) attack against RF on MNIST with 10 iterations.

(4) Decision-based Attacks

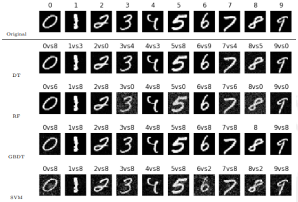
ART Boundary Attack (2018) — BA(ℓ_2) with $\varepsilon = 0.01$ *and* $\delta = 0.01$

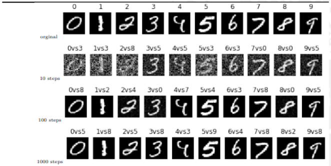
Census: max_iter=2

Model	Objective	Avg. time (per attack)	Prec.	Rec.	F1	MCC	Fooling rate
SVM	Targeted	0.26 sec	0	0	0	-0.128	**28.13**
	Untargeted	0.33 sec	0	0	0	0	13.26
DT	Targeted	0.01 sec	0.123	0.472	0.195	-0.68	**89.2**
	Untargeted	0.01 sec	0.18	0.697	0.286	-0.326	81.13
RF	Targeted	1.29 sec	0.1	0.376	0.158	-0.749	**91.4**
	Untargeted	1.6 sec	0.177	0.688	0.282	-0.345	86.4
GBDT	Targeted	0.01 sec	0.108	0.408	0.171	-0.727	**90.66**
	Untargeted	0.01 sec	0.184	0.729	0.294	-0.328	85.93

MNIST: max_iter=100

Model	Objective	Avg. Time (per attack)	Accuracy	Fooling rate (%)
SVM	Targeted	16.73 sec	0.66	33
	Untargeted	34.83 sec	0.02	**98**
DT	Targeted	0.49 sec	0.39	57.99
	Untargeted	0.83 sec	0.01	**100**
RF	Targeted	0.27 sec	0.79	20
	Untargeted	1.24 sec	0.1	**89**
GBDT	Targeted	3.97 sec	0.66	31
	Untargeted	11.46 sec	0.02	**99**





(4) Decision-based Attacks

ART HopSkipJump Attack (2019) — HJSA(ℓ_∞), HJSA(ℓ_2)

max_iter=2

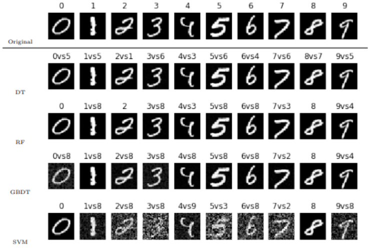


Image of original and perturbed images generated by HSJA(ℓ_∞) attack against DT, RF, GBDT, and SVM models on MNIST with 10 iterations and $\epsilon = 0.01$. As a result, the predicted class labels for adversarial images by DT, RF, GBDT, and SVM models.

Census:

Model	Objective	Dist.	Avg. Time (per attack)	Prec.	Rec.	F1	MCC	Fooling rate
SVM	Targeted	ℓ_∞	0.4 sec	0.094	0.35	0.148	-0.767	**92**
		ℓ_2	0.33 sec	0.094	0.35	0.148	-0.767	**92**
	Untargeted	ℓ_∞	0.49 sec	0.229	1	0	0	86.73
		ℓ_2	0.4 sec	0.229	1	0	0	86.73
DT	Targeted	ℓ_∞	0.03 sec	0.184	0.738	0.295	-0.348	**80.66**
		ℓ_2	0.02 sec	0.186	0.741	0.297	-0.318	**79.86**
	Untargeted	ℓ_∞	0.03 sec	0.23	0.956	0.371	0.012	67.2
		ℓ_2	0.02 sec	0.228	0.95	0.368	-0.004	68.73
RF	Targeted	ℓ_∞	**7.82 sec**	0.176	0.682	0.280	-0.349	**80.26**
		ℓ_2	**5.98 sec**	0.175	0.679	0.278	-0.362	**80.66**
	Untargeted	ℓ_∞	**8.44 sec**	0.229	0.939	0.368	0.003	71.86
		ℓ_2	**6.73 sec**	0.229	0.927	0.367	0.006	71.86
GBDT	Targeted	ℓ_∞	0.11 sec	0.176	0.685	0.28	-0.361	**80.66**
		ℓ_2	0.09 sec	0.167	0.644	0.276	-0.405	**81.73**
	Untargeted	ℓ_∞	0.06 sec	0.233	0.959	0.375	0.037	72.66
		ℓ_2	0.05 sec	0.231	0.959	0.372	0.02	73.46

MNIST:

Model	Objective	Distance	Avg. Time (per attack)	Accuracy	Fooling rate (%)
SVM	Targeted	ℓ_∞	0.17 sec	0.79	19
		ℓ_2	0.14 sec	0.82	17
	Untargeted	ℓ_∞	0.71 sec	0.1	**91**
		ℓ_2	0.59 sec	0.09	**92**
DT	Targeted	ℓ_∞	0.01 sec	0.8	11
		ℓ_2	0.01 sec	0.8	10
	Untargeted	ℓ_∞	0.04 sec	0.12	**87**
		ℓ_2	0.03 sec	0.11	**92**
RF	Targeted	ℓ_∞	0.08 sec	0.86	11
		ℓ_2	0.08 sec	0.81	15
	Untargeted	ℓ_∞	0.56 sec	0.12	**87**
		ℓ_2	0.42 sec	0.09	**92**
GBDT	Targeted	ℓ_∞	0.02 sec	0.79	20
		ℓ_2	0.01 sec	0.79	20
	Untargeted	ℓ_∞	0.06 sec	0.11	**89**
		ℓ_2	0.04 sec	0.11	**89**

Poisoning Attack

ART Poisoning Attack on SVM

Attack Algorithms	Parameters
Poisoning Attacks on SVM	ε= 0.3 or ε= 1, ε_{step}=0.1, maximum iteration=10 15 examples (Attack data points on census data) 315 training sets÷180 test sets (For census data) 10 examples (Attack data points on MNIST data) 1169 training sets÷565 test sets (For MNIST data)

Experimental results on clean and poison SVM model on 180 **census** original examples.

Trained SVM model	Precision	Recall	F1 score	MCC
Clean	**0.667**	**0.4**	**0.5**	**0.404**
Poison($\varepsilon = 0.3$)	0.6	0.2	0.3	0.244
Poison($\varepsilon = 1$)	0.667	0.178	0.281	0.257

Experimental results on clean and poison SVM model on 565 **MNIST** original examples.

Trained SVM model	Accuracy
Clean	**0.9947**
Poison($\varepsilon = 0.3$)	0.9733
Poison($\varepsilon = 1$)	0.9760

Conclusions and Future work

- We are identifying Adversarial Attacks from ART that supports our chosen Machine Learning Models.
- We shows the strength and weakness of these attack algorithms on chosen ML models for the untargeted and targeted cases in tabular data (Adults Census Income dataset) and Images (MNIST).

Future work:

- We extend this work to incorporate Adversarial training of GBDT [2] And ART defense mechanisms such as adversarial training methods and evaluate adversarial examples generated by ART evasion and poisoning attacks on the resilient ML models.

[2] Stefano Calzavara, Claudio Lucchese, and Gabriele Tolomei. Adversarial training of gradient-boosted decision trees. In CIKM '19: Proceedings of the The 28th ACM International Conference on Information and Knowledge Management, 2019.