Is Machine Intelligence Hackable?

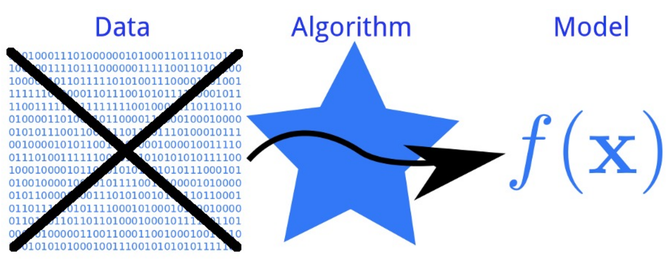
Mridul Gupta

Dept. of Computer Science, Indraprastha Institute of Information Technology, New Delhi
mridul15061@iiitd.ac.in

Introduction

With the scope of Machine Learning increasing, it becomes necessary to check for the security issues that can arise when applying a machine learning model to a given problem.
In this experiment we, for the first time, probe into the security vulnerabilities of machine intelligence. Some simple experiments reveal that we can clone parametric models without necessitating access to the training data. We hope our findings will trigger research towards blocking the security loopholes in learning algorithms.

Motivation



- **Exploring Intelligence Hacking.**
- **A Data-less approach for generating Machine Learning Models.**

Terminology

- **Black Box** - A trained Machine Learning Model provided to us with no access to its data or parameters of the learned function.
- **White Box** - An artificially cloned Model created which mimics the behaviour of the Black Box

Algorithm

Algorithm 1: Algorithm for cloning model using random data under constraints

Data: blackbox , whitebox, number of features, number of samples, range of features
Result: a cloned model

```
data = array of zeros of size [number of features, number of samples];
for each feature in the data do
    feature = uniform distribution under constraints(range of feature);
label = blackbox.predict(data);
whitebox.fit(data, label);
// Checking the performance of the whitebox on original data
print whitebox.score(original data, original labels);
```

Datasets

Source: UCI Machine Learning Repository

- Iris
- Make Moons
- Breast Cancer
- Heart Disease
- Adult Census
- Pittsburgh Bridges
- Abalone
- Car Evaluation

Methodology

Random Probing Under Constraints
It is quite common that we derive the data from natural phenomena and each of the feature value thus recorded lies under a minimum and maximum value. We hypothesize that exploiting this range to generate random data can bolster the performance of our white box even in high dimensions.
Example: Considering the Iris dataset, the sepal width and sepal length may well be under some specific range.

Figures and Tables

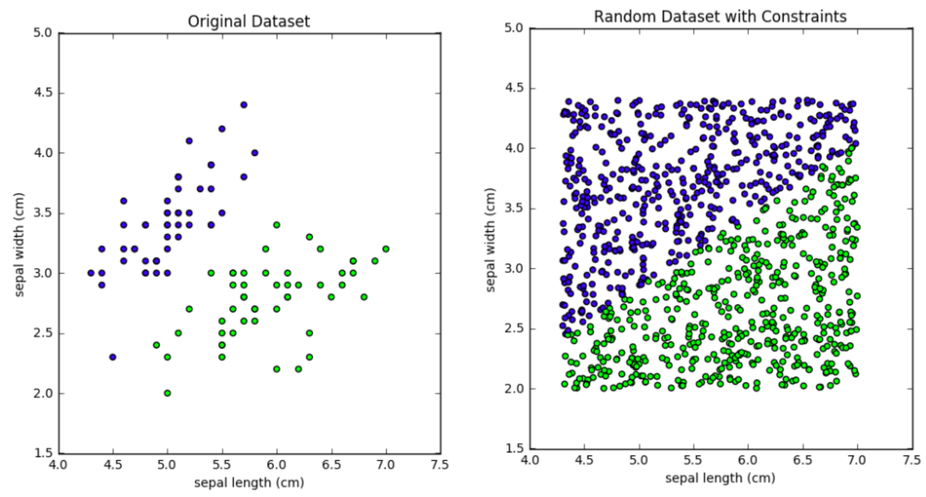


Fig. 1: Iris Dataset in Two Dimensions (a) Original dataset (b) Randomly generated data under constraints

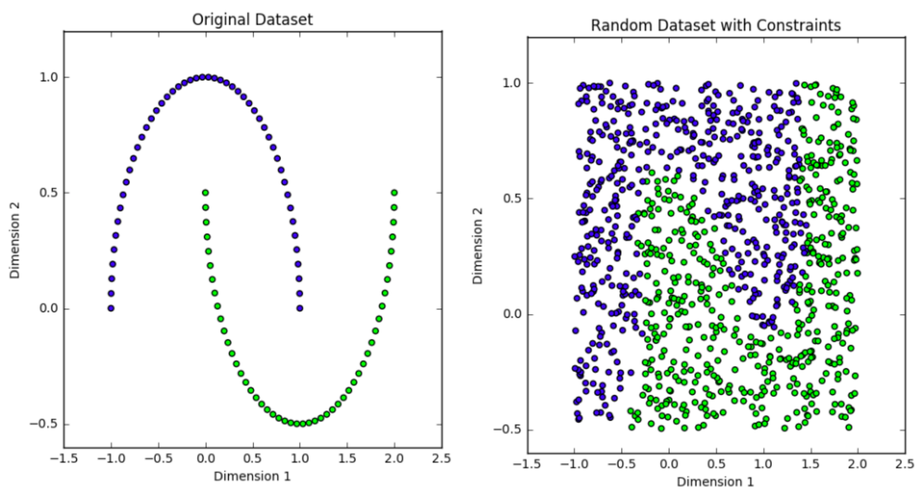


Fig. 2: Make Moons Dataset in Two Dimensions (a) Original dataset (b) Randomly generated data under constraints

Dataset Used	Number of Features	Black Box Accuracy	White Box Accuracy
Make Moons	2	0.95	0.95
Heart	9	0.69	0.66
Bridges	29	0.91	0.90
Breast Cancer	30	0.62	0.61
Census	107	0.76	0.75

Table 1: Performance of Neural Network as Black Box and White Box model on different Datasets

Dataset	Number of Features	Black Box Accuracy	White Box Accuracy
Make Moons	2	0.99	0.99
Heart	9	0.96	0.70
Bridges	29	0.93	0.65
Breast Cancer	30	0.94	0.45
Census	107	0.84	0.74

Table 2: Performance of Random Forests as Black Box and White Box model on different Datasets

Conclusion

- We analyse the cloning procedure and its accuracy using different datasets of varying dimensionality and sizes.
- Ability to learn without data, and without model details.
- Successfully able to mimic parametric models, results still vary for non-parametric models.

Can we clone a Machine Learning Model? ***Yes***
But can we clone *all* Machine Learning Models? ***NO***

- **Attempting to classify Machine Learning Models on whether they can be cloned or not.**