Cloning Machine Intelligence

Mridul Gupta

Indraprastha Institute of Information Technology

mridul15061@iiitd.ac.in

January 13, 2018

Motivation

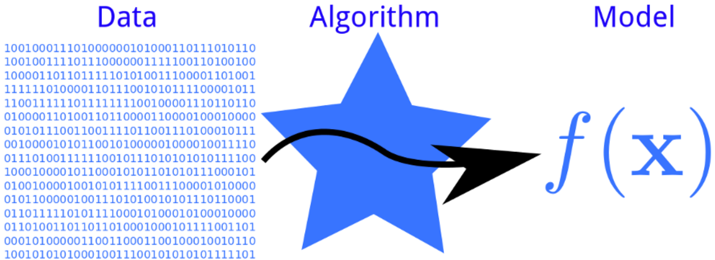
Motivation

1. Exploring Intelligence Hacking

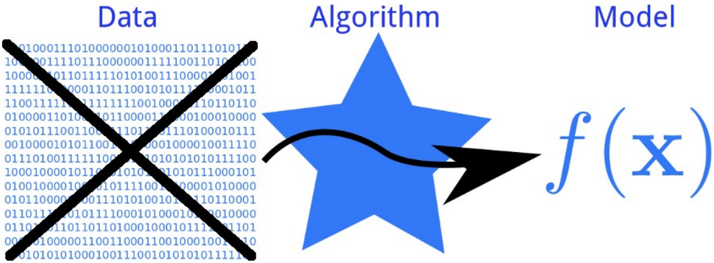
Motivation

1. Exploring Intelligence Hacking
2. A Data-less Approach for Machine Learning Models

General Machine Learning Model



Data-less (Cloned) Machine Learning Model



Introduction

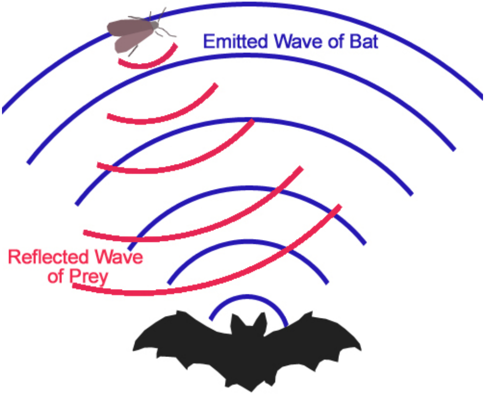
Problem Statement

Given a **Black Box** Machine Learning Model, try an design a **White Box** model that mimics it with a similar level of accuracy.

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

Approach 1



Approach 1

```
Algorithm 1: Algorithm for cloning model using random data
Data: blackbox, whitebox, number of features, number of samples
Result: a cloned model
data = array of zeros of size [number of features, number of samples];
for each feature in the data do
    feature = normal distribution (mean = 0, standard deviation = 1.0);
    label = blackbox.predict(data);
    whitebox.fit(data, label);
    // Checking the performance of the whitebox on original data
    print whitebox.score(original data, original labels);
```

Approach 1

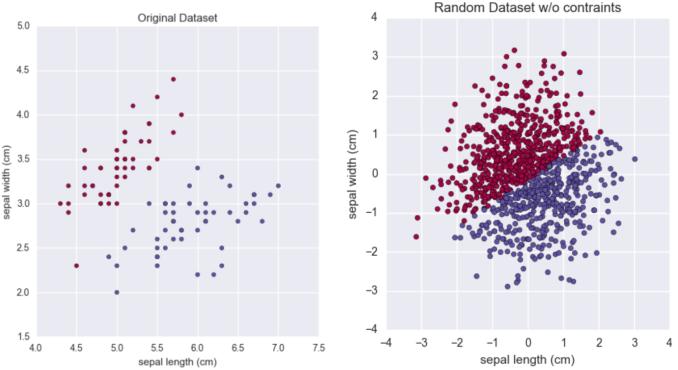
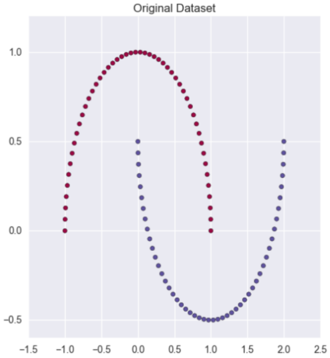


Figure: Original and Randomly Generated Iris Dataset Without Constraints

Approach 1



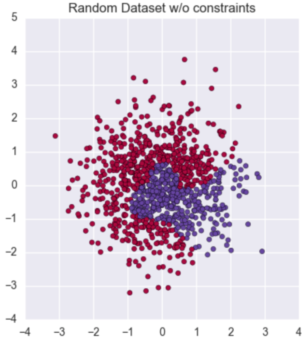


Figure: Original and Randomly Generated Make Moons Dataset Without Constraints

Results of Approach 1

Dataset Description	Number of Features	Black Box	Black Box Accuracy	White Box	**White Box Accuracy**
Iris	2	Logistic Regression	0.99	Logistic Regression	**0.97**
BUPA	7	Logistic Regression	0.59	Logistic Regression	**0.57**
Heart	9	Logistic Regression	0.73	Logistic Regression	**0.54**
Breast Cancer	30	Logistic Regression	0.95	Logistic Regression	**0.53**
Make Moons*	2	Neural Network	0.99	Random Forest	**0.98**

Table: **Cloned model accuracy on original dataset when trained with random dataset generated under constraints**

**The Make Moons Dataset is highly Nonlinear, thus we decided to train it on Neural Networks.*

Approach 2

Constraining Data!

Approach 2

Algorithm 2: 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);
```

Approach 2

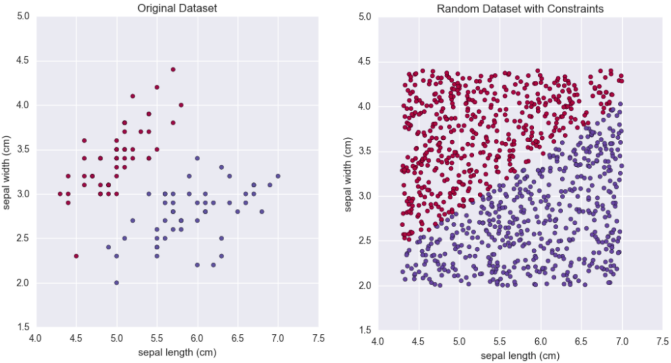


Figure: Original and Randomly Generated Iris Dataset With Constraints

Approach 2

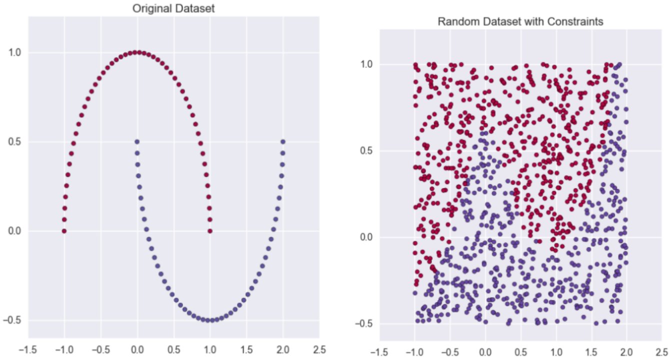


Figure: Original and Randomly Generated Make Moons Dataset With Constraints

Results of Approach 2

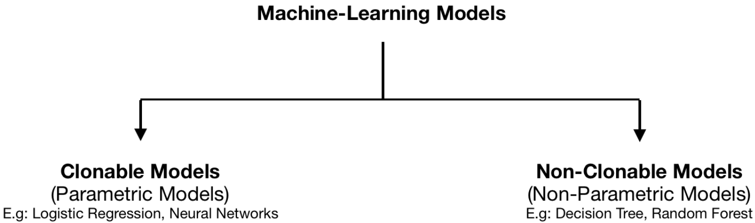
We observe an improvement in the White Box Accuracy.

Dataset Description	Number of Features	Black Box	Black Box Accuracy	White Box	**White Box Accuracy**
Iris	2	Logistic Regression	0.99	Logistic Regression	**0.97**
BUPA	7	Logistic Regression	0.59	Logistic Regression	**0.57**
Heart	9	Logistic Regression	0.73	Logistic Regression	**0.72**
Breast Cancer	30	Logistic Regression	0.95	Logistic Regression	**0.94**
Make Moons*	2	Neural Network	0.99	Random Forest	**0.98**

Table: **Cloned model accuracy on original dataset when trained with random dataset generated under constraints**

**The Make Moons Dataset is highly Nonlinear, thus we decided to train it on Neural Networks.*

Observation



Machine-Learning Models

Clonable Models
(Parametric Models)
E.g: Logistic Regression, Neural Networks

Non-Clonable Models
(Non-Parametric Models)
E.g: Decision Tree, Random Forest

Drawbacks of This Approach

Although there has been an improvement in the results, we have made some big assumptions!

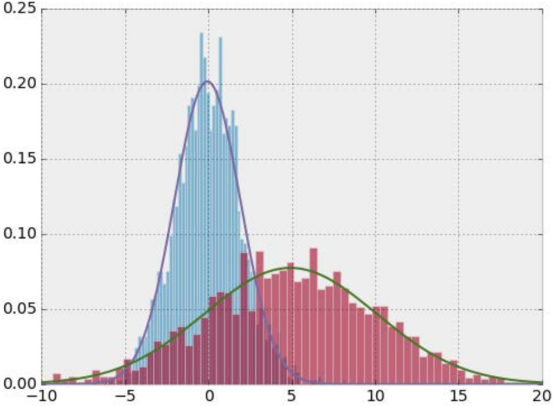
- Kills the whole idea of a “black box”!
- Assumption of Feature Independence.
- Very time consuming to generate constraints for large number of features

E.g. *Gene expression data can have* $\geq$*30,000 features!*

Current Approach

Current Approach

Using Multivariate Gaussian Mixture Models



Conclusion

- **Ability to learn without data, and without model details**
- Important Observation in terms of security.
- We attempt a new basis to classify models, based on whether they can be cloned or not.

Future Work

- Work on the current approach using Multivariate GMM
- Compare results for a lot more data and models.

Questions?

Acknowledgements

This project couldn't be done without the immense help of

- **Najeeb Khan**, B.Tech Student at Jamia Milia Islamia.
- **Dr. Debarka Sengupta**, Assistant Professor at IIIT Delhi.

References

Here are links to the datasets used:

- **Iris**, `archive.ics.uci.edu/ml/datasets/iris`
- **Make Moons**, `scikit-learn.org/stable/modules/generated/sklearn.datasets.make_moons.html`
- **BUPA**, `archive.ics.uci.edu/ml/datasets/Liver+Disorders`
- **Heart**, `archive.ics.uci.edu/ml/datasets/heart+Disease`
- **Breast Cancer**, `archive.ics.uci.edu/ml/datasets/breast+cancer`
- **Adult Census**, `archive.ics.uci.edu/ml/datasets/breast+cancer`
- **Abalone**, `archive.ics.uci.edu/ml/datasets/breast+cancer`
- **Mushrooms**,