GVDRA Report and Recommendation

Gun Violence Data Reconstruction Algorithm Analysis

Aydin O'Leary and Jack Greenberg

March 2020

Summary

This report serves as an official guide to the *Gun Violence Data Reconstruction Algorithm*, herein referred to as **GDVRA**. The report was written in response to concerns by **LinAlgCo** about the context and ethics of this algorithm in the real world. Our analysis answers the question: "What are the ethical implications around the use of multivariate linear regression to profile suspects of gun violence incidents?" Our recommendation, in brief, is that, given the error rate of this algorithm and the ethical implications of using such an algorithm, the algorithm is not appropriate in predictions of future events and therefore should not be used in such cases.

Introduction

Background

GDVRA is an algorithm developed by police forces and detectives to aid in the classification, identification, and profiling of suspects. The algorithm is purported to be able to identify the number of perpetrators, as well as their genders and ages. The integration of algorithms into police forces have, however, been problematic in the past, especially when it comes to justice and sentencing. In some cities, like Chicago, police are using algorithms to assign a risk/threat level to arrested people[1]. This algorithm is black-boxed and unaudited, which begs the question, "How can we be sure this algorithm is non-discriminatory?" Indeed, investigations into the Chicago Police Department have shown distinctive patterns of racial profiling. Their algorithm may have some associative blame.

Algorithm Overview

We now explain the mechanisms of the algorithm. The training dataset was scraped from the **Gun Violence Archive** and uploaded to Github[4]. This data was cleaned up and organized to fit the algorithm better using the *one hot* method, which works by encoding a dataset into a set of binary fields[2]. For example, the raw dataset include information about the state in which the incident took place, so one-hot encodes that as a value like `is_California` or `is_Texas` so that data can be encoded numerically and can be used in the algorithm.

After the data is cleaned, the algorithm uses the Moore-Penrose pseudoinverse to calculate the matrix of regression coefficients. Those coefficients are then used to calculate the 7 outcomes of a set of 73 predictor (independent) variables.

Ethics

A major ethical question to consider is the difference between "accuracy" of algorithm and "accuracy" in real life. Since a linear regression analysis doesn't know what happened, any result that's perfectly accurate by the definition of the algorithm may not be accurate by real-life standards.

The broad question we must ask ourselves with an algorithm like this is, "what are the ethical implications of an algorithm that uses historic data to fill in missing data about gun violence incidents?" An algorithm like this in the hands of police could easily lead to increased police presence in areas of high crime, and, if combined with facial recognition algorithms like EIGENFACES, could lead to widespread racial profiling, police mistrust, and the perpetuation of violent crime in low-income areas. We will tackle the sub-question: "what errors arise when applying methods of multivariate linear regression to data about gun violence?"

Methods

The algorithm begins by cleaning the data. A one-hot encoding method is used to transform month, day-of-week, and US state into a logical data value (0 or 1) so that the algorithm can process it. One-hot is used instead of numerical values for month and day of week because linear combinations of weekdays or months wouldn't make very much sense.

The raw data samples include the following fields:

```
incident_id ,date ,state ,city_or_county ,address ,n_killed ,
    n_injured ,incident_url ,source_url ,
    incident_url_fields_missing ,congressional_district ,
    gun_stolen ,gun_type ,incident_characteristics ,latitude ,
    location_description ,longitude ,n_guns_involved ,notes ,
    participant_age ,participant_age_group ,
    participant_gender ,participant_name ,
    participant_relationship ,participant_status ,
```

```
participant_type, sources, state_house_district,
state_senate_district
```

An example entry would be:

```
461105,2013-01-01,Pennsylvania,Mckeesport,1506 Versailles
  Avenue and Coursin Street,0,4,http://www.
  gunviolencearchive.org/incident/461105,http://www.post
  -gazette.com/local/south/2013/01/17/Man-arrested-in-
  New-Year-s-Eve-shooting-in-McKeesport/stories
  /201301170275,False,14,,,"Shot - Wounded/Injured||Mass
  Shooting (4+ victims injured or killed excluding the
  subject/suspect/perpetrator, one location)||Possession
  (gun(s) found during commission of other crimes)||
  Possession of gun by felon or prohibited person
  ",40.3467,,-79.8559,,Julian Sims under investigation:
  Four Shot and Injured,0::20,0::Adult 18+||1::Adult
  18+||2::Adult 18+||3::Adult 18+||4::Adult 18+,0::Male
  ||1::Male||3::Male||4::Female,0::Julian Sims,,0::
  Arrested||1::Injured||2::Injured||3::Injured||4::
  Injured,0::Victim||1::Victim||2::Victim||3::Victim
  ||4::Subject-Suspect,http://pittsburgh.cbslocal.com
  /2013/01/01/4-people-shot-in-mckeesport/||http://www.
  wtae.com/news/local/allegheny/U-S-Marshals-task-force-
  arrests-New-Year-s-party-shooting-suspect/17977588||
  http://www.post-gazette.com/local/south/2013/01/17/Man
  -arrested-in-New-Year-s-Eve-shooting-in-McKeesport/
  stories/201301170275,,
```

We clean this data up into 73 fields.

```
0,4,,5,0,0,3,1,1,4,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,

  0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,
  0,0,0,0,1,0,0,0,0
```

The majority of the data is zeros because of the ones-hot encoding. As we will see in the **Detailed Findings** section, this will become an issue due to the number of added fields.

The data set is split into predictor and response data sets (independent and dependent variables, respectively). The following equation used for the regression model:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \epsilon$$

Where $\mathbf{Y}$ is the matrix of dependent variables of the training data, $\mathbf{X}$ is the matrix of independent variables, $\boldsymbol{\beta}$ is the matrix of regression coefficients, and ϵ is the associated error of the equation.

The algorithm ignores the error generated. If we take $\mathbf{K}$ as 73 independent (known) variables of the training data and $\mathbf{U}$ as the 7 dependent (unknown) variables of the data matrix with 100,000 entries we have:

$$\mathbf{K}_{100000,73}\boldsymbol{\beta}_{73,7} = \mathbf{U}_{100000,7}$$

We can see from this that $\boldsymbol{\beta}$ is the matrix that correlates known variables to unknown variables. To solve, the Moore-Penrose inverse is then calculated as

$$\boldsymbol{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

$\boldsymbol{\beta}$ is then used with a different matrix of known variables $\mathbf{A}$ to get its associated matrix of unknown variables $\mathbf{B}$:

$$\mathbf{B} = \mathbf{A}\boldsymbol{\beta}$$

In this way, we can approximate a set of dependant variables after training our algorithm on a full set of data. We accomplished this transformation using a linear regression function from Python's `scikit-learn` library.

Detailed Findings

Below is a table of RMSE values for the comparison of the reconstructed test data to the original test data.

Variable	**Error (RMSE)**
Number of Guns Involved	.523
Number of Adults 18+	1.395
Number of Teens 12-17	.628
Number of Children 0-11	.291
Number of Males	1.305
Number of Females	.741
Number of Suspects	.988

It's obvious that the error is far too large to actually be usable in a real setting: even though the error looks small, the numbers used in calculation are small enough that the error is too large to be useful. For example, the error could mean that our algorithm says there are two suspects when in fact there are anywhere between one and three suspects inclusive, which is far too much error for something this sensitive.

We believe the error is due in large part to the one-hot encoding: many variables are encoded as Boolean values, so things like state, month, and weekday are given enormous weight in the algorithm because there are 50 Boolean variables (one for each state).

If we take an example, we can see that if we put in the following data, our results have questionable accuracy:

```
IN:
    STATE:                      Florida
    MONTH:                      January
    WEEKDAY:                    Thursday
    NUMBER OF INJURED:          3
    NUMBER OF KILLED:           0

ALGORITHM OUT:
    NUMBER OF GUNS INVOLVED:    0.752
    NUMBER OF ADULTS:           2.602
    NUMBER OF TEENS:            0.287
    NUMBER OF CHILDREN:         0.007
    NUMBER OF MALES:            2.384
    NUMBER OF FEMALES:          0.512
    NUMBER OF SUSPECTS:         0.572

ACTUAL OUT:
    NUMBER OF GUNS INVOLVED:    1
    NUMBER OF ADULTS:           3
    NUMBER OF TEENS:            0
    NUMBER OF CHILDREN:         0
    NUMBER OF MALES:            4
    NUMBER OF FEMALES:          1
    NUMBER OF SUSPECTS:         3
```

In this example, the algorithm gets the number of males, females, and suspects wrong, with error ranging from moderately close to drastically wrong. Given what we know about the nature of violent crimes and incidents involving guns, algorithms with error of this scale are unfit to be used in the field without massive changes to internal workings and data processing.

Recommendations

Many of the criticisms leveled against criminal profiling of serial killers can also be applied to this algorithm. For example, criminal profiling is often accused of correlating variables that don't affect each other [3]. That criticism applies to the GDVRA algorithm as well: it correlates variables like `is_Wyoming` and `number_of_suspects`, which almost certainly have no relation to each other. For this reason, the ethical issues raised in the introduction, and the massive error, the GDVRA algorithm should not be used for any reason.

References

[1] Andrew Guthrie Ferguson. The police are using computer algorithms to tell if you're a threat.

[2] Wikimedia Foundation. One hot.

[3] Maurice Godwin. Reliability, validity, and utility of criminal profiling typologies. *Journal of Police and Criminal Psychology*, 17(1), 2002.

[4] James Ko. Gun violence data.