Modeling of Recent Ebola Epidemic

Zhongyue Zhang
November 29, 2014
AMATH 383 Paper No-W
zhangz6@uw.edu

Abstract

In this paper we explored different models for the recent outbreak of Ebola in West Africa. We fit these models to the dataset (HDX), and determined which can make the most accurate prediction. Our experiment suggest that machine learning algorithm can outperform traditional modeling using differential equations in terms of prediction accuracy. Particularly, elastic net has lowest error, which is significantly better than other model. However, machine learning has a drawback being harder to interpret.

1. Introduction

Recent Ebola outbreak in West Africa have reach unparalleled level (CDC). The outbreak started on March 23, 2014, and cases surged since then (WHO). As of October 5, 2014, there are 8033 reported cases of Ebola (HDX). Ebola is a rare and deadly disease caused by Ebola virus, and it is spread through direct contact with body fluids of an infected person who is already showing symptoms (CDC).

Mathematical models of epidemic give us a quantifiable forecast of the future and may help us make informed decisions when dealing with the epidemic. We use one traditional compartmental model (Kermack) and two machine learning model to predict the future development of the current Ebola outbreak. We examine the accuracy of these models, and seek to determine the best.

2. Methods

2.1 Dataset

The data we used are obtained from The Humanitarian Data Excahnge (HDX), which is compiled from multiple sources including World Health Organization (WHO), national health ministries, etc. The data started from March 23, 2014 and the last data point is on November 26, 2014, containing multiple countries. However, we only use three countries: Guinea, Sierra Leone and Liberia, since other country does not have large amount of cases. For our analysis, we assume these three countries has little interaction with each other, and we can ignore them. There are six categories of data, and we will use one of them for our analysis: Confirmed cases, which represents the cumulative confirmed Ebola cases. For the two machine learning model, we scale the input data to have a mean of 0 and a variance of 1. As for total population of the counties, we collected them from The World Bank (The World Bank).

2.2 Model

2.2.1 Compartmental Model

Compartmental model has been used to describe epidemic for a long time (Kermack). We use a simple SEIR four compartment model. Here is the differential equations govern the system:

$$\frac{dS}{dt} = -\frac{\beta IS}{N}$$
$$\frac{dE}{dt} = \frac{\beta IS}{N} + \alpha E$$
$$\frac{dI}{dt} = \alpha E - \gamma I$$
$$\frac{dR}{dt} = \gamma I$$

Parameters explained:

β: $contact\ rate\ (day^{-1})$

α: $incubation\ period\ (day^{-1})$

γ: $time\ to\ death/recover\ (day^{-1})$

Variables explained:

S: Susceptible, population that can be infected.

E: Exposed, people who are exposed to the virus but not yet infectious.

I: Infectious, people who are infectious.

R: Removed, either recovered and are immune to the virus or died.

2.2.2 Multivariable Regression

We experiment with linear regression model with various feature vector and regularization methods.

The following are the models we use and their objective functions:

Ordinary Least Squares	$\min_w \|Xw - y\|_2^2$
Ridge Regression	$\min_w \|Xw - y\|_2^2 + \lambda\|w\|_2^2$
Lasso	$\min_w \frac{1}{2n_{samples}} \|Xw - y\|_2^2 + \lambda\|w\|_1^2$
Elastic Net	$\min_w \frac{1}{2n_{samples}} \|Xw - y\|_2^2 + \lambda\rho\|w\|_1^2 + \frac{\lambda(1-\rho)}{2}\|w\|_2^2$

X represents the feature vector; w is the weight of each feature, λ and ρ are the hyper parameters.

2.2.3 Support Vector Regression

Support vector regression is extended from support vector classification, a widely used machine learning algorithm (scikit-learn) (Cortes). We use four different kernel: linear, polynomial, radial basis function (RBF) and sigmoid couple with several different feature vectors.

2.3 Model Evaluation and Selection

The objective is predicting the cumulative cases given time, and we use single variable (time being the predictor variable/feature vector) linear regression (ordinary least squares) as the baseline performance. The dataset will be split into two part, the first part is for parameters fitting, and the second part will be used for evaluation of different model. We will call the first part training set, and second part test set from now on. The test set will be the last 60 days of the dataset, specifically from September 27, 2014 to November 26, 2014. For the compartmental model, which is represented by a system of differential equations, we fit its parameters using least-squares optimization on different countries separately, as we assume that the interactions of these country are negligible. For the two machine learning model: multivariable regression and support vector regression, we use cross validation for hyper parameters tuning. Particularly, we train the support vector model on different country separately as well as jointly. When we train the models with different countries jointly, the country attribute of a data point will turn into a three dimension binary vector (one-hot encoding). For the final evaluation of different models, we use squared error across the data from three different countries from test set.

3. Experiments

3.1 Compartmental Model

Country	Parameters	Mean Squared Error
Sierra Leone	516.5097, 0.5540, 499.3509	1.0181e+05
Liberia	10679, 0.0000, 266	2.6692e+06
Guinea	653.5931, 0.0403, 499.8891	6.3659e+03
Average		925791.9667
Weighted Average		744202.0907

We build the model using Matlab and optimized using least square fitting, with a constraint setting all parameters to be greater than or equal to 0. Our model has a particular large MSE on Liberia's data, which may be due to the noisy data. *(Visualization see Fig. 1-3)*

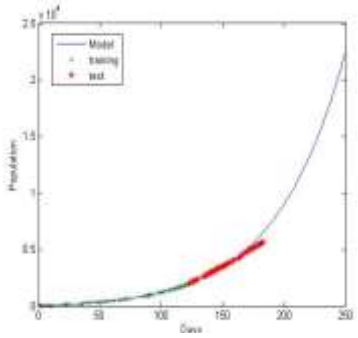


Figure 1 Sierra Leone compartmental model

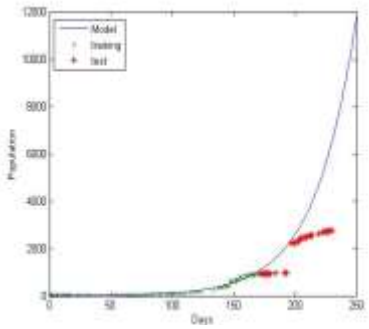


Figure 2 Liberia compartmental model

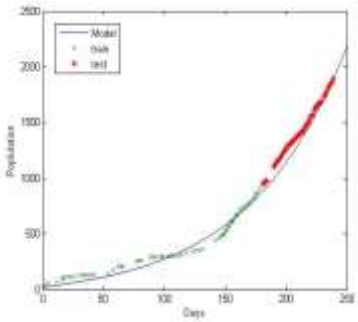


Figure 3 Guinea compartmental model

3.2 Multivariable Regression

The experiment is done using scikit-learn. This type of model essentially fit a linear model to the data using different kind of regularization method (penalize large weight).

3.2.1 Ordinary Least Square

Ordinary least square is significantly worse than our compartmental model.

Country	Mean Squared Error
Sierra Leone	3.434977e+06
Liberia	1.719252e+06
Guinea	3.64039e+05
Average	1839422.666667
Weighted Average	1818966.007143

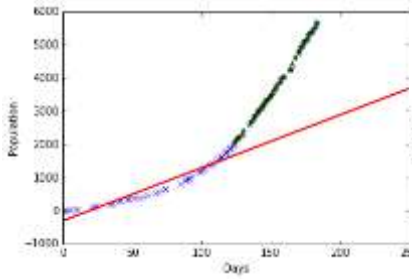


Figure 4 Sierra Leone Ordinary Least Square

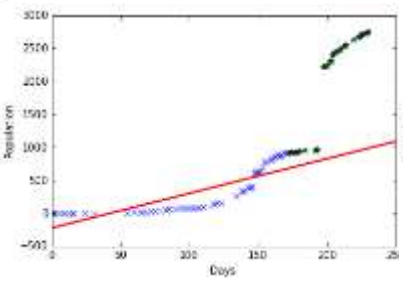


Figure 5 Liberia Ordinary Least Square

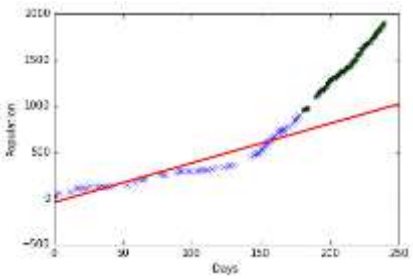


Figure 6 Guinea Ordinary Least Square

3.2.2 Ridge, Lasso, Elastic Net

We use cross validation to pick the best feature vector and parameters. The best feature vector

Country	Ridge MSE	Lasso MSE	Elastic Net MSE
Sierra Leone	3.4987e+04	9.9276e+04	3.8493e+04
Liberia	1.3508e+05	1.4321e+05	2.0044e04
Guinea	6.9809e+05	5.7278e+03	2.6425e+03
Average	289385.6667	82737.9333	13713.1667
Weighted Average	312472	75472.4529	14748.9321

we found is [day, day^2, day^3, logistic (day)], with day^3 and logistic (day) being weighted the most by our algorithm. Logistic (day) is understandable, since the growth of a population is usually

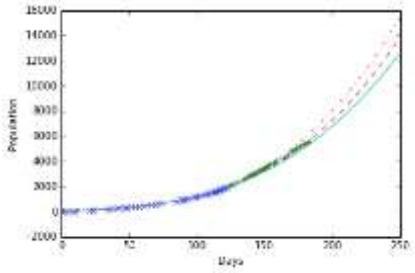


Figure 7 Sierra Leone

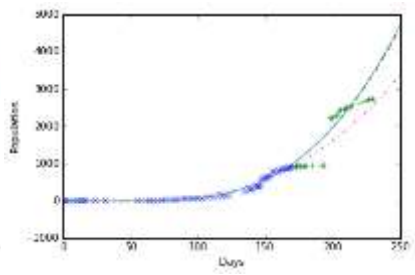


Figure 8 Liberia

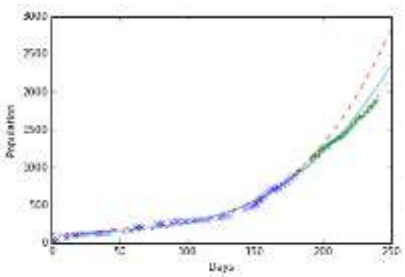


Figure 9 Guinea

associated with logistic function, but heavy weight on day^3 is not so clear. In the visualization, solid line represents ridge regression, dash represents lasso, and dash dot represents elastic

net. All three linear model performs better than compartmental model in terms of MSE. The best model, elastic net is able to achieve 1/67 of MSE of the compartmental model.

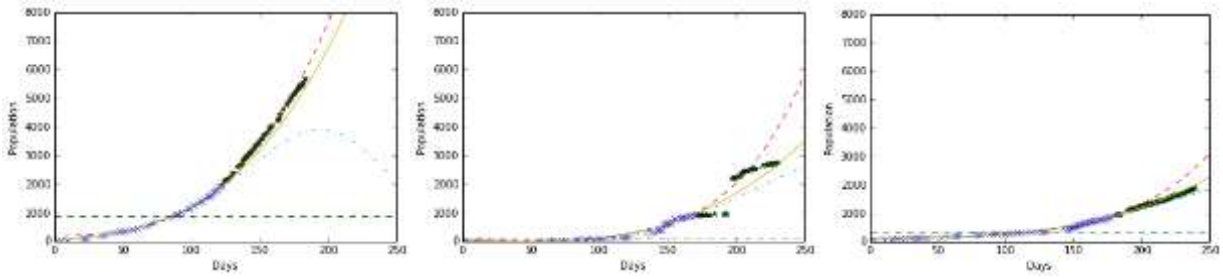
3.3 Support Vector Regression (SVR)

Similar to multivariable regression, support vector regression is also tested using scikit-learn, and we use cross validation to tune the hyper parameters corresponds to each kernel. Previously, we train two our models (compartment and linear) separately on each country. This time we also train our model jointly. The intuition is that the nonlinearity of SVR may be able to extract share information from different countries.

3.3.1 Separate Training

Sigmoid kernel are worse than our baseline, ordinary least square, so we will not discuss here. Linear, polynomial and RBF kernel perform better than compartmental model, however, neither are better than elastic net. Linear kernel has the lowest error, but it is still about 6 times the mean

Country	Linear MSE	Poly MSE	RBF MSE	Sigmoid MSE
Sierra Leone	1.1556e+05	1.3864e+04	9.9430e+05	9.7112e+06
Liberia	1.6707e+05	2.6706e+05	3.4227e+05	3.8929e+06
Guinea	2.8978e+03	1.1243e+05	1.9807+e04	1.2620e+06
Average	95175.9333	131118	668285	4955366.6667
Weighted Average	86522.6671	118094.3571	458486.3857	4974880.7143



squared error of elastic net.

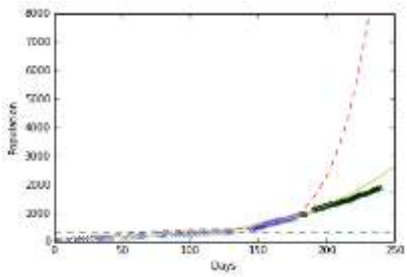


Figure 10 Sierra Leone SVR

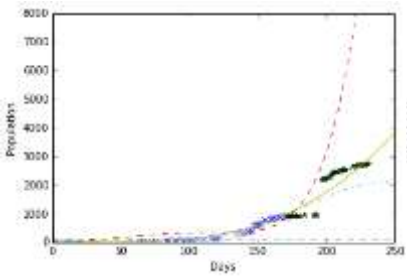


Figure 11 Liberia SVR

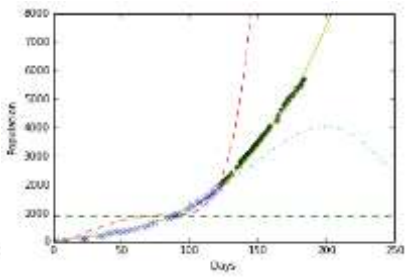
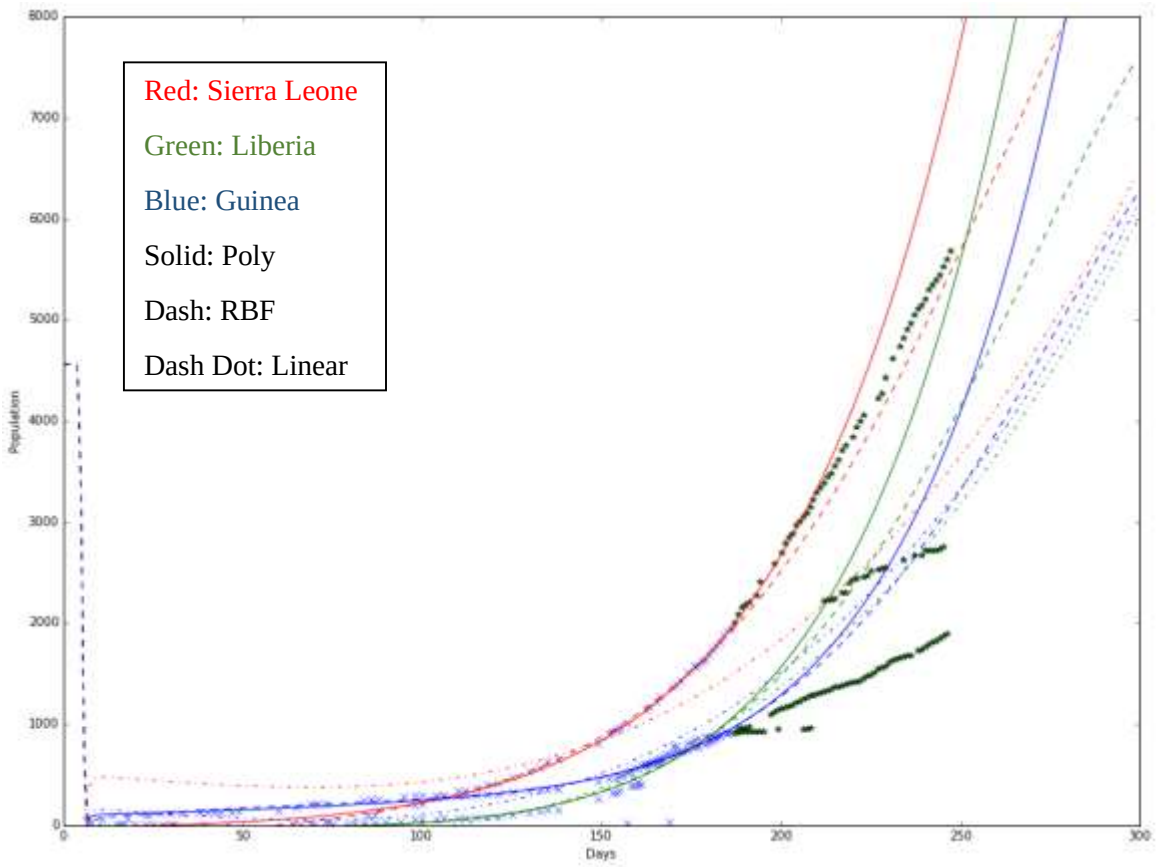


Figure 12 Guinea SVR

3.3.2 Joint Training

We trained linear, polynomial and RBF SVR, and found that RBF perform better. On the other hand, linear and polynomial SVR has larger error. For RBF SVR, the accuracy boost is not very significant, and thus the hypothesis we mentioned aboved, RBF SVR may be able to take advantage of share information among countries, is rejected.



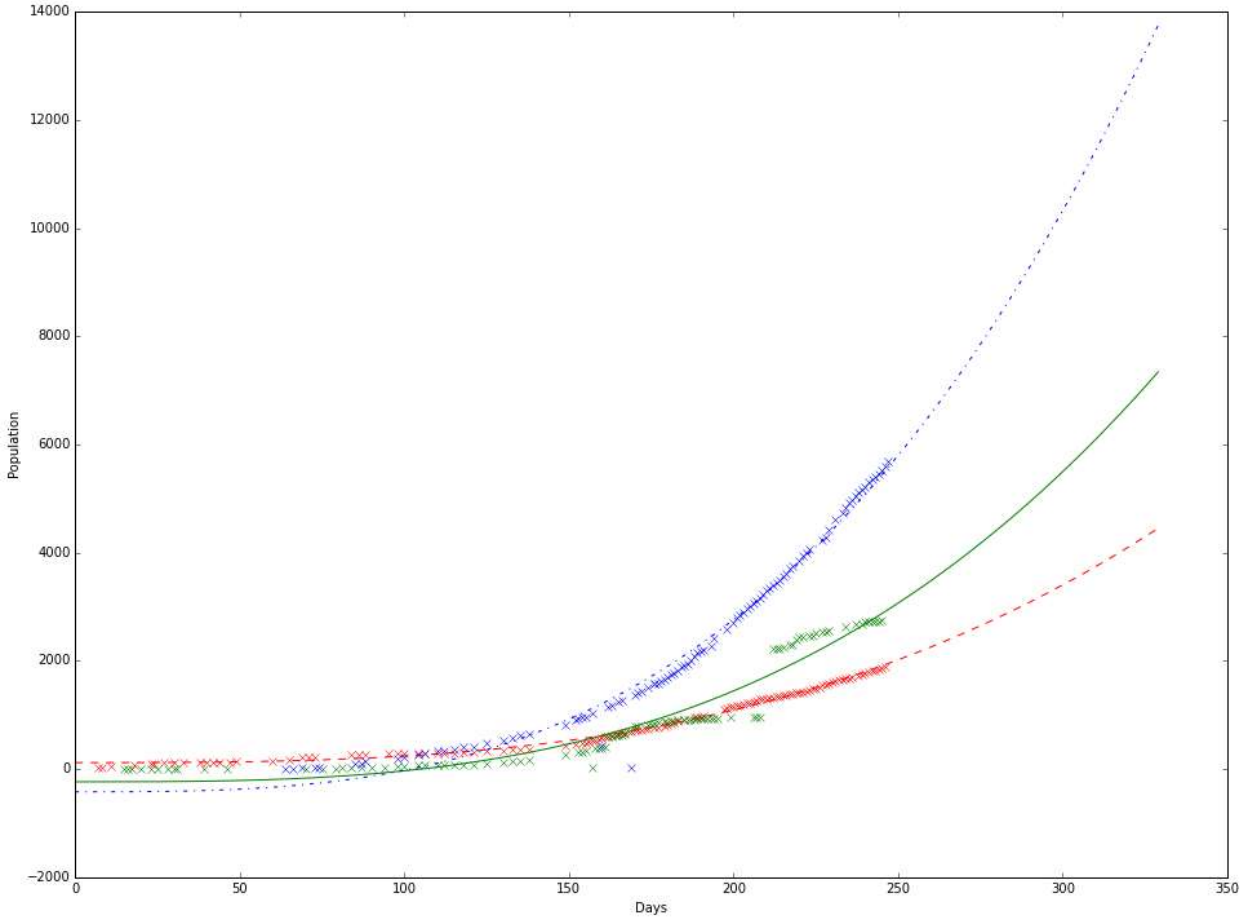
Country	Linear MSE	Poly MSE	RBF MSE
Weighted Average	1.0152e+06	6.9779e+05	2.3682 +05

3.4 Final Prediction

Based on our experiment, we determined that elastic net is the best estimator, and we use it to estimate the next 60 days development of the ongoing outbreak. The result shows that the cases will continue to grow with no sign of slowing.

Predicted Cases at Day 310 (Approx. 1/27/2015)

Sierra Leone	Liberia	Guinea
11436	6096	3743



4. Conclusion

Our experiment result shows elastic net has the best predicting ability among all the models we tested, and by a wide margin, 1/6 of the error of the 2nd best. The traditional differential equations performs better than ordinary least square, but is worse than other machine learning techniques. The benefit of using differential equations is that it is much easier to interpret the result in terms of real world meaning. Machine learning models has the advantages of being more accurate, and it is also much faster when we try to fit the data. We've tried to fit a 6 equations 11 parameters ODE system, but it took too long using Matlab to fit data, and we aborted it switching to a simpler SEIR model. To further improve the accuracy of our model, we may design a model that incorporate historical epidemic data and information of each counties health care system.