HIV Infection Percentage Curve Prediction

Leyan Zhu, Hang Zhang, Haonan Song

Github link: https://github.com/abc286d/ECE208-Final-Project

Abstract

HIV, known as human immunodeficiency virus, can cause terrible disease whose final stage is AIDS. It most often spreads through unprotected sex. Therefore, it is meaningful to predict the spread speed of HIV infection within a group of people who may have intimate relationships. In this paper, we introduce a way to predict the spread speed of HIV infection. Also, we find a way to predict the real infection speed through the data of diagnosis data. Which might be meaningful to help control the spread speed of HIV.[1]

Introduction

AIDS, which is also known as acquired immunodeficiency syndrome, is a terrible disease that weakens a person's immune system by destroying important white cells that fight disease and infection.

This terrible disease is caused by HIV (human immunodeficiency virus) and is the final stage of this virus infection. [2] There are basically 3 ways to spread HIV - unprotected sex with a person who has HIV, contact with the blood of a person who has HIV and Women can give it to their babies during pregnancy or childbirth. Most often is that it can spread through unprotected sex.[3]

Therefore, it is meaningful to study the HIV infection curve in a group of people who may have intimate relationship. It can not only help us predict the rate of HIV spreads, but also provide us a very important information about how to prevent them.

In our project, we would predict the percentage of HIV infected people in a group of people given the data before a fixed time point. And for more than one groups, we would like to find the order of their infection percentage increase rates, which indicates whether these group spread fast or not. With all the information, we can decide what actions to take to help these people and how to prevent the infection from spreading.

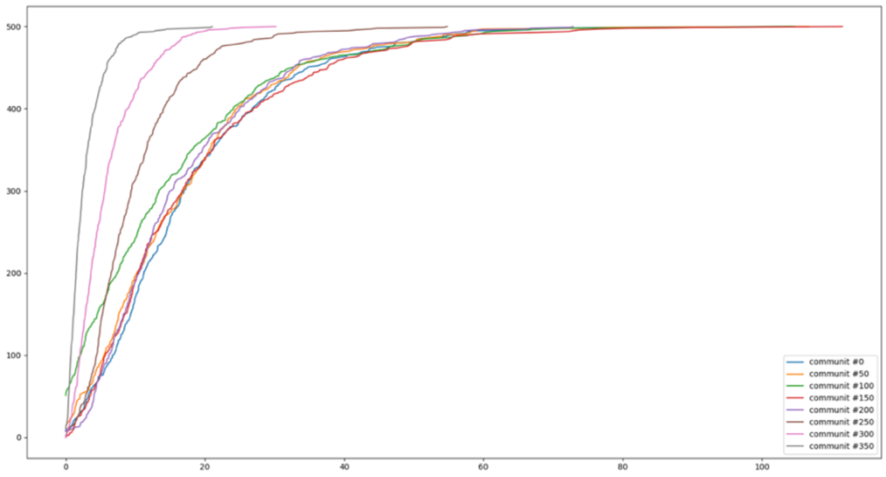
Also, we know that people getting a definite diagnosis are less than the number of people truly having HIV. In our second part. We would also like to use the definite diagnosis percentage curve to predict the real percentage in these groups and also, check these groups' increase rate order at any specific time.

Generating dataset

The dataset we used is generating dataset from FAVITES simulation. It's hard to collect the real data since it may be related to the privacy of some patients. The dataset contains two part, the first part is the transmission network consisting of 400 communities each with 500 individuals. The transmission procedure will proceed until all of the individuals are infected. The expected diagnosis time is 4 years. Communities are Barabasi-Albert graphs and, the expected degree of the contact network is 8. The second part is diagnosis transmission network, which contain diagnosis time and infection time for all individuals of the contact network. The dataset is generated with the help of Sina Malekian.

Model and Error

Model



Picture 1

We can see from picture 1, it's the infection curve for 8 communities we choose. We need to find a way to predict the plot after a specific time. Also, we need a way to calculate the order error. We came up with two ways to predict the result.

First is to predict the plot with polynomial fit. We can see it as a piecewise function, where the first part is quadratic function and the second part is a horizontal function.

$$y = at^2 + bt + c, t < t_0 \tag{1}$$

$$y = d, t \geq t_0 \tag{2}$$

we know all the points of (t, y) before a time t_0. Say that we have n points here, we need to fit the curve so that it has least square error.

Least square error is defined as:

$$e = \frac{1}{n}\sum_{i=1}^{n}(y - y_0)^2 \tag{3}$$

However, with simple test, we find this method does not work very well. Thus, we come up with the second solution.

We thought of a better fit curve for this problem:

$$y = 1 - a\, e^{-b\,t} \tag{4}$$

where a, b are the two parameters we need to learn.

This equation is not a simple polynomial function, but we can use some trick to transform it.

With (4), we can get:

$$a\, e^{-b\,t} = 1 - y \qquad (5)$$

take log base e to both side of (5), we can get:

$$\log(a) - b\ t = \log(1 - y) \qquad (6)$$

Let $\log(a) = a', \log(1 - y) = y'$, we can get:

$$a' - b\ t = y' \qquad (7)$$

Where all the points of (t, y') are given.
With this method, we turn an exponent function like (4) into a linear function like (7) and we need to find curve with a' and b that has least square error, which is calculated through (3).

For the second part of this problem, we are asked to use the diagnosis data to predict the real infection data. We can use the same function for infection and diagnosis. Since there are 400 groups, we can find 400 pairs of a and b for infection part, we use a[n], b[n] to denote these parameters.

For diagnosis part, we also have 400 pairs of a' and b', we use a'[n] and b'[n] to denote these parameters, to calculate a, b with a' and b', we need to find a relationship between a and a', b and b'.

We assume that the relationship between a and a', b and b' are both linear,

$$a = \mathrm{m}a' + n \qquad (8)$$

$$b = \mathrm{l}\ b' + s \qquad (9)$$

we fit these with the same way as finding the least square error like (3).

Error

To find a way to calculate the error, we must clarify what is the result.

For different groups, we would like to calculate the increase percentage divided by the original infection percentage which is $\frac{\Delta y}{y}$. This value indicates how likely is one person in this community tending to affect others. And then for these groups, we make an order based on this value, and compare the predict order with the real order to calculate error.

We used two ways to calculate the order: MAE and Kendall Tau distance.
MAE is defined like this:

$$e = \frac{1}{n}\sum_{i=1}^{n} |y - y_0| \qquad (10)$$

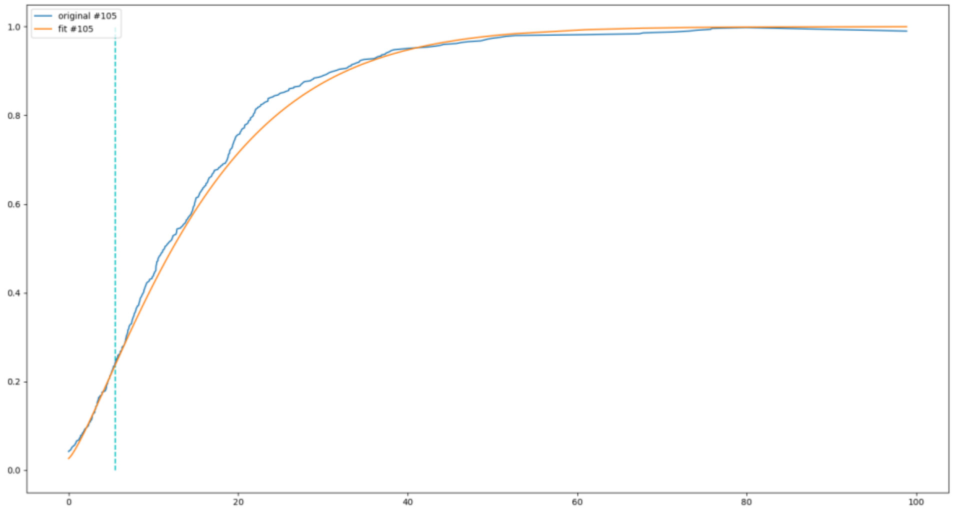
The Kendall Tau distance between two list l_1 and l_2 is:

$$K(l_1, l_2) = |(i,j): i < j, (l_1(i) < l_1(j) \cap l_2(i) > l_1(j)) \cup (l_2(i) > l_1(j) \cap l_2(i) < l_2(j))|$$

Where $l_1(i)$ and $l_2(i)$ are the ranking of element I in l_1, l_2respectively.
Also, we need to normalize it by divide by n(n-1)/2, where n is the number of total numbers of element in list l.

Result

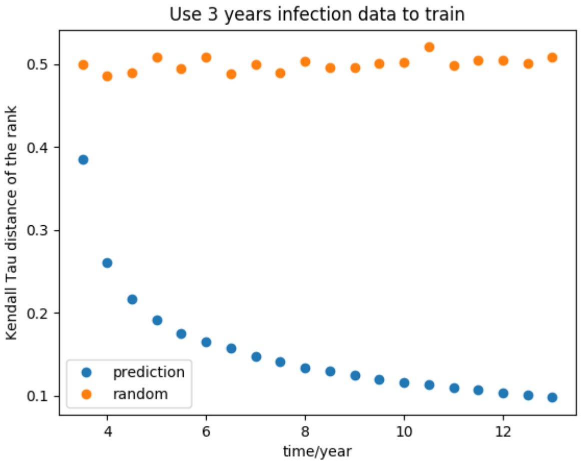
Curve fitting



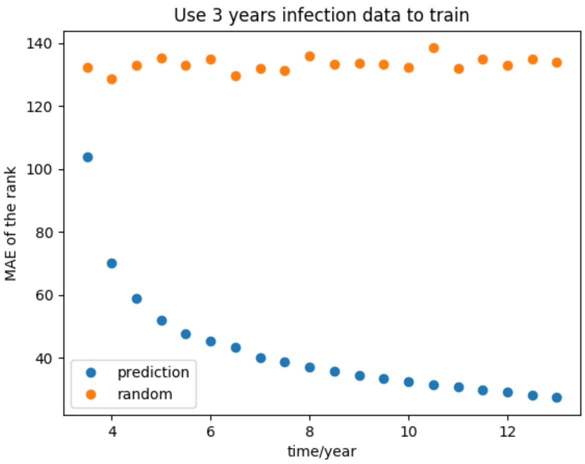
The above figure is an example of our fitter curve. We use the left part of the vertical line to train (the first 5 years' data). The fitting curve is very close to the original curve, which is satisfying.

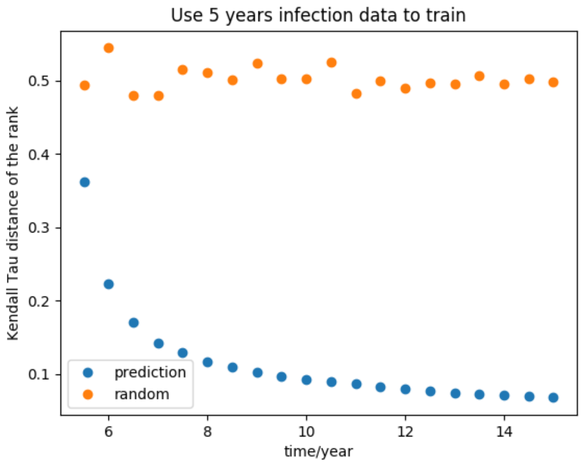
Problem1:

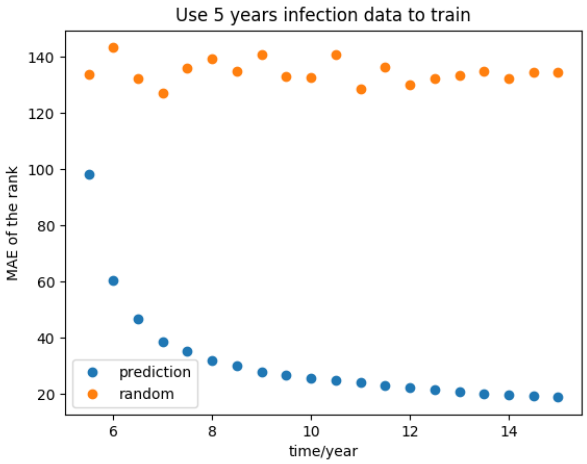
Kendall Tau Distance

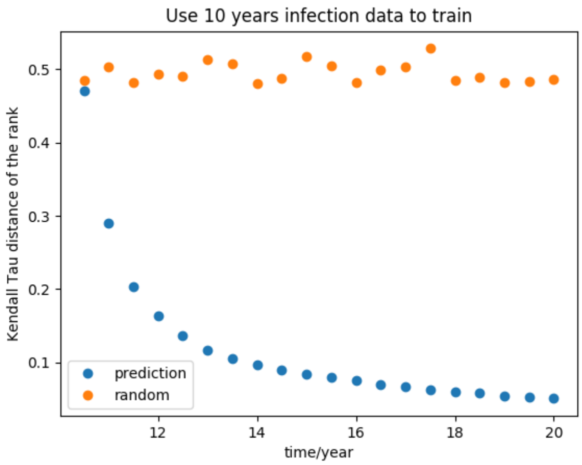


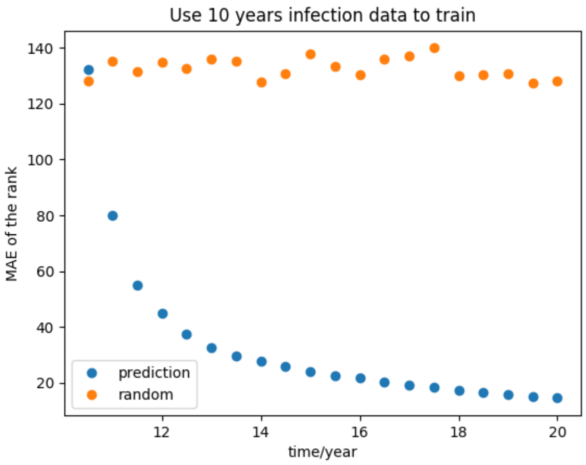
MAE





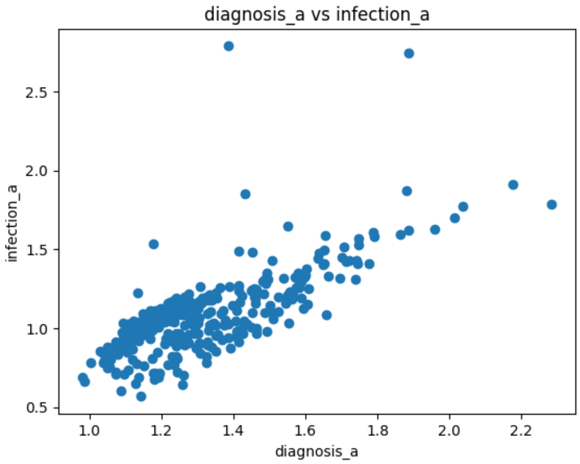


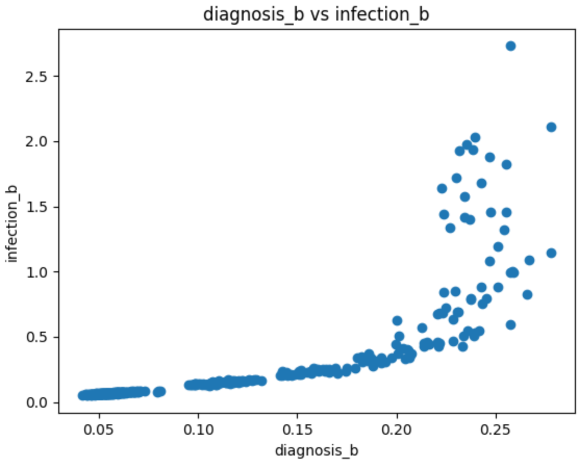




As the figures shown above, we use two ways—MAE and Kendall Tau distance to evaluate the error and we compare our results with the random guess. We use 3, 5, 10 years' infection data to train separately, and predict for the next 10 years. The more training data are used, the more accurate is the result. Also, the prediction is more accurate when predicting many years later. This is because the increase is large when many years later, thus the error proportion becomes small.

Relationship between a, a’ and b, b’



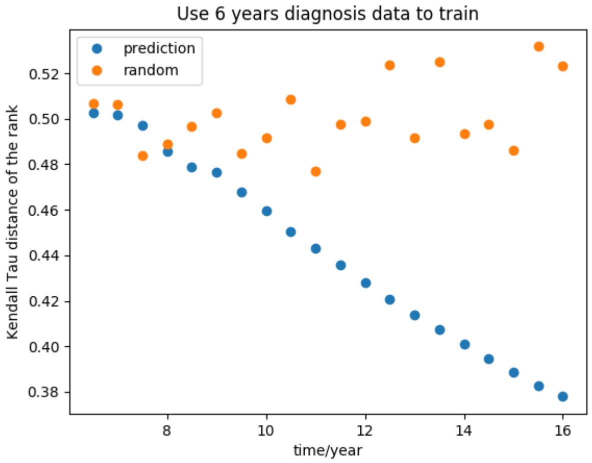


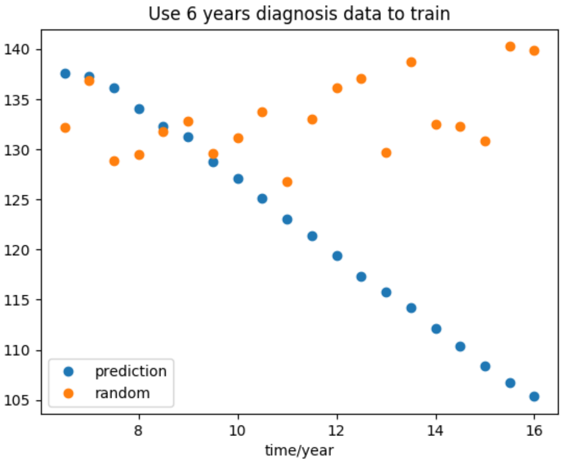
As we mentioned in the model part, for the infection curve we have the a, b parameters. For the diagnosis curve, we have the a’, b’ parameters. We need to find the relationship between these parameters. We hope it will be linear and verify on that. From the graph above we can see that, the b and b’ tends to have a good linear property when b is not so large. Thus, the method we just said is feasible.

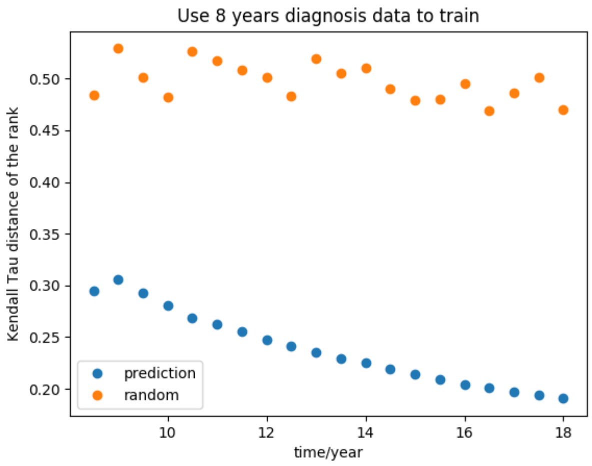
Problem2

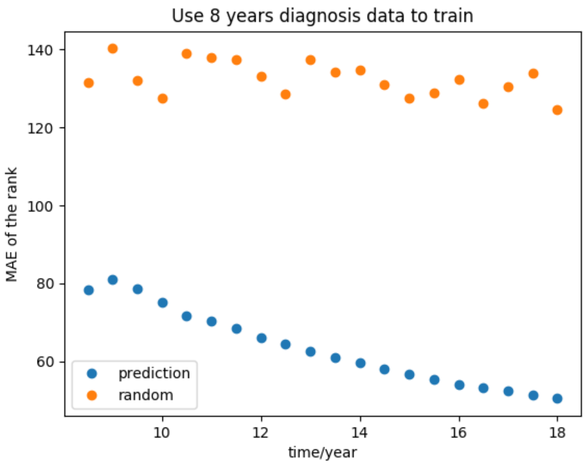
Kendall Tau Distance

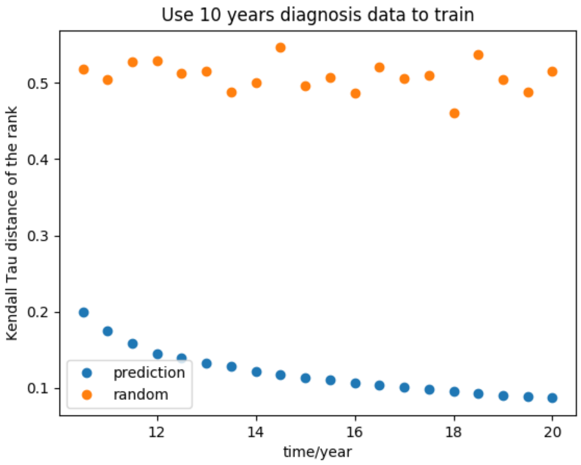
MAE

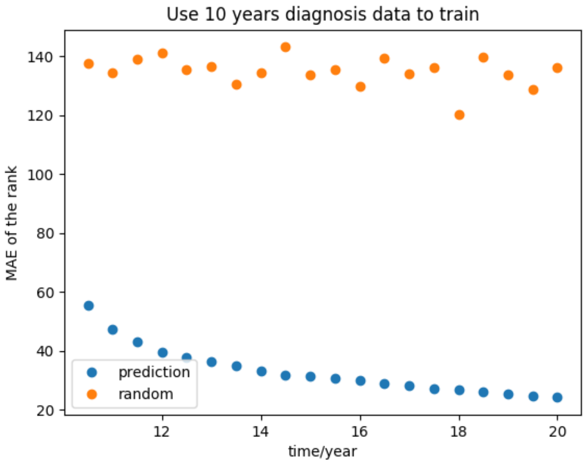












These are the results of using diagnosis data to predict real infection data. The prediction result is worse than Problem 2, especially when training set is small. For example, when we use first 6 years diagnosis data to train, some of the prediction is even worse than random guess. However, when the training set gets larger, the result soon become much better.

References

[1] Niema Moshiri, Manon Ragonnet-Cronin, Joel O Wertheim, Siavash Mirarab, FAVITES: simultaneous simulation of transmission networks, phylogenetic trees and sequences, Bioinformatics, Volume 35, Issue 11, 1 June 2019, Pages 1852–1861, https://doi.org/10.1093/bioinformatics/bty921
[2] https://medlineplus.gov/hivaids.html
[3] https://www.cdc.gov/hiv/basics/