Machine Learning and Gradient Descent for Infectious Disease Risk Prediction

Catherine Li
Mentor: Daniel Lazarev

MIT PRIMES Conference

October 14, 2023

Table of Contents

Epidemiology

- Study of incidence, spread, and control of disease
- Source, nature, and risk factors
- Recent emergence of infectious diseases
- Disease Models
  - SIR compartmental model (Susceptible, Infected, Recovered): system of differential equations
  - Maximum Entropy: least-biased probability distribution given constraints

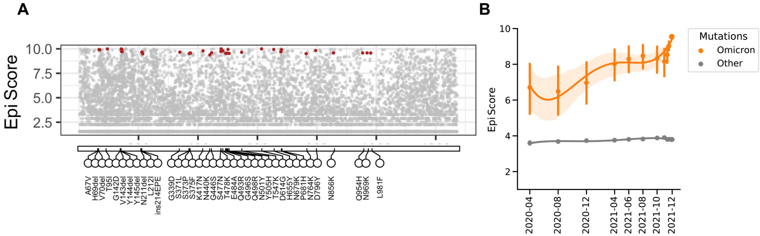
Factors of Transmission

- Temperature
- Humidity
- Vaccination
- Social contact/human mobility patterns
- Host-receptor binding affinity
- Ecological niche of virus
- Viral mutations/escape

Table of Contents

Risk Scores for SARS-CoV-2 Mutations

- Maher et al. combined three epidemiological factors of mutations into Epi Score
  1. Mutation frequency
  2. Fraction of unique haplotypes (group of DNA variations that are inherited together) in which mutation occurs
  3. Number of countries in which mutation occurs
- Forecasts spread of mutations months in advance



Risk Scores for SARS-CoV-2 Mutations, cont.

- For mutation i, let $freq_i, hap_i, count_i$ denote mutation frequency, haplotype occurrence, and country occurrence
- Define f_i, h_i, c_i as percentiles of $freq_i, hap_i, count_i$ (0 to 1)
- Exponential score: Epi $\text{Score}_i = \frac{10^{f_i} + 10^{h_i} + 10^{c_i}}{3}$
  - Exponentials help further differentiate high-risk mutations
- Performed better than any other measure (evolution, immune, etc.)

Table of Contents

Geo Scores

- Risk assignment for geographical regions
  - ZIP Codes in NYC
- Exponential Geo Score calculated from
  1. Vaccination rate
  2. Population density
  3. Socioeconomic status (SES): median annual household income
- 7 scores: all combinations of 1, 2, or 3 variables

Geo Scores, cont.

- Percentiles v_i, d_i, s_i in ZIP Code i

$$\text{Geo Score } 1_i = 10^{v_i},$$
$$\text{Geo Score } 2_i = 10^{d_i},$$
$$\text{Geo Score } 3_i = 10^{s_i},$$
$$\text{Geo Score } 4_i = \frac{10^{v_i} + 10^{d_i}}{2},$$
$$\text{Geo Score } 5_i = \frac{10^{v_i} + 10^{s_i}}{2},$$
$$\text{Geo Score } 6_i = \frac{10^{d_i} + 10^{s_i}}{2},$$
$$\text{Geo Score } 7_i = \frac{10^{v_i} + 10^{d_i} + 10^{s_i}}{3}.$$

Geo Score Performance

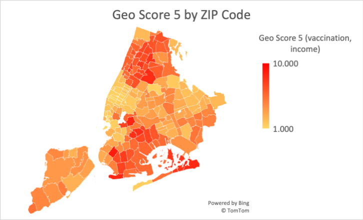
- Compared against 2 ground-truth metrics: test positive rate, death rate
  - Same exponential percentiles method to compare scores with metrics on a 1-10 scale
- Geo Score 5 (vaccination rate and socioeconomic status) performed best in Mean Absolute Error

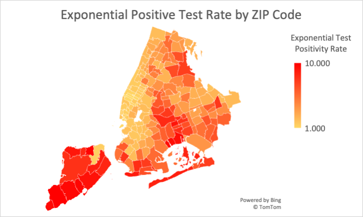
	Test Positive Rate	Death Rate
Geo Score 1	2.001	2.225
Geo Score 2	3.093	2.908
Geo Score 3	2.254	1.969
Geo Score 4	2.261	2.224
Geo Score 5	1.881	1.833
Geo Score 6	2.444	2.187
Geo Score 7	2.102	1.979

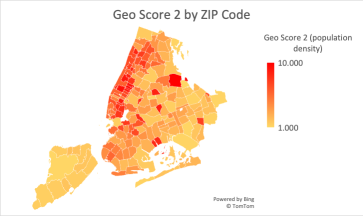
Geo Score Performance, cont.

Geo Scores

Metrics







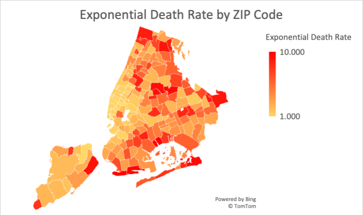


Table of Contents

Tunable Weights

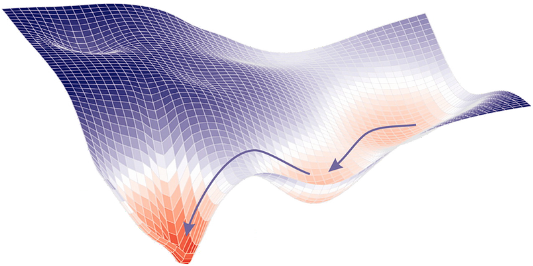
- Let p_1, p_2, p_3 be the distributions of the exponential scores for vaccination rate, population density, and SES across the ZIP codes
- Find parameters $0 \leq \alpha, \beta, \gamma \leq 1$ such that $\alpha + \beta + \gamma = 1$ and

$$p = \alpha p_1 + \beta p_2 + \gamma p_3$$

  best predicts test positive/death rate distributions
- Minimize L_1 (total absolute error) or L_2 distance (squared error)

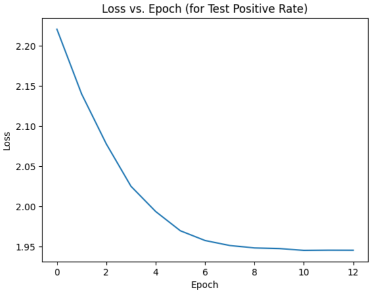
Gradient Descent

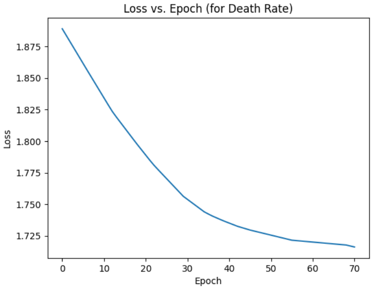
- Optimization algorithm often used to train machine learning models
- Loss function f
- Gradient: $\langle f_x, f_y \rangle$ (direction of steepest ascent)
- Learning rate/step size



Results

- Split dataset in half: training and evaluation
- Compared against linear regression and neural network
- $\beta \approx 0$; $\alpha \approx 0.5$ for test positive, $\alpha \approx 0.7$ for death





Summary

- Geographical risk assignment with exponential scores
- Gradient descent algorithm performs better than linear regression and neural network models
    - Provides interpretable results

Acknowledgements

I would like to thank:

- My mentor, Daniel Lazarev
- Dr. Tanya Khovanova, Prof. Patel Etingof, Dr. Slava Gerovitch, and the MIT PRIMES-USA Program
- My family

References

- M. C. Maher, I. Bartha, S. Weaver, J. Iulio, E. Ferri, L. Soriaga, F. A. Lempp, B. L. Hie, B. Bryson, B. Berger, D. L. Robertson, G. Snell, D. Corti, H. W. Virgin, S. Pond, and A. Telenti. Predicting the mutational drivers of future SARS-CoV-2 variants of concern. *Sci. Transl. Med.*, 14 (633), 2022.
- C. Bishop. *Pattern Recognition and Machine Learning*. Springer Science+Business Media, 2006.
- PRIMO.ai. *Gradient Descent Optimization and Challenges*. 2023.