Machine Learning and Gradient Descent for Infectious Disease Risk Prediction

Catherine Li
Mentor: Daniel Lazarev

MIT PRIMES Conference

October 14, 2023
Table of Contents

1. Introduction
2. Exponential Risk Scores
3. Geographic Risk Model
4. Tunable Weights and Gradient Descent
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

1. Introduction
2. Exponential Risk Scores
3. Geographic Risk Model
4. Tunable Weights and Gradient Descent
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.
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: $\text{Epi Score}_i = \frac{10^{f_i} + 10^{h_i} + 10^{c_i}}{3}$

- Exponentials help further differentiate high-risk mutations.
- Performed better than any other measure (evolution, immune, etc.)
# Table of Contents

1. Introduction
2. Exponential Risk Scores
3. Geographic Risk Model
4. Tunable Weights and Gradient Descent
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 $i$

Geo Score $1_i = 10^{v_i}$,

Geo Score $2_i = 10^{d_i}$,

Geo Score $3_i = 10^{s_i}$,

Geo Score $4_i = \frac{10^{v_i} + 10^{d_i}}{2}$,

Geo Score $5_i = \frac{10^{v_i} + 10^{s_i}}{2}$,

Geo Score $6_i = \frac{10^{d_i} + 10^{s_i}}{2}$,

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

<table>
<thead>
<tr>
<th></th>
<th>Test Positive Rate</th>
<th>Death Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Geo Score 1</td>
<td>2.001</td>
<td>2.225</td>
</tr>
<tr>
<td>Geo Score 2</td>
<td>3.093</td>
<td>2.908</td>
</tr>
<tr>
<td>Geo Score 3</td>
<td>2.254</td>
<td>1.969</td>
</tr>
<tr>
<td>Geo Score 4</td>
<td>2.261</td>
<td>2.224</td>
</tr>
<tr>
<td>Geo Score 5</td>
<td>1.881</td>
<td>1.833</td>
</tr>
<tr>
<td>Geo Score 6</td>
<td>2.444</td>
<td>2.187</td>
</tr>
<tr>
<td>Geo Score 7</td>
<td>2.102</td>
<td>1.979</td>
</tr>
</tbody>
</table>
Geo Score Performance, cont.

**Geo Scores**

**Geo Score 5 by ZIP Code**

**Geo Score 2 by ZIP Code**

**Metrics**

**Exponential Positive Test Rate by ZIP Code**

**Exponential Death Rate by ZIP Code**

Catherine Li  Mentor: Daniel Lazarev

Machine Learning and Gradient Descent for Infectious Disease Risk Prediction
Table of Contents

1. Introduction
2. Exponential Risk Scores
3. Geographic Risk Model
4. Tunable Weights and Gradient Descent
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
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

