An Introduction to Model-based Machine Learning – Homework

Reza Sameni
Department of Biomedical Informatics, Emory University
Email: rsameni@dbmi.emory.edu

Nov. 6, 2024

PLEASE READ THE FOLLOWING GUIDELINES CAREFULLY BEFORE ANSWERING THE QUESTIONS:

- Please select and deliver **only one of the following questions**.
- Please do not combine the question; partial answers from different question sets will not be graded.
- Share the link to a public GitHub repository with your source codes enumerated by the part of the question.
- Write a summary of findings in the `README.md` file of the GitHub repository, including:
  - Your name and contact,
  - The question number you select to answer,
  - Key insights,
  - Comparative model performance,
  - Relevance to model-based machine learning,
  - Suggestions for future modeling improvements.
- Please use Markdown language to write and document your reports. Juputer Notebooks should include figures and images which can be cited in your report. MATLAB figures can be stored in .png or .jpg.
- AI tools, including but not limited to ChatGPT and Copilot, are permitted. However, to ensure integrity, fairness in evaluation and grading, their use must be disclosed. A disclaimer should be included in the `README.md` file as follows:
  - *"Disclaimer: No generative AI (in any form) has been used to complete this homework."*
  - *"Disclaimer: [AI-TOOL-NAME(S)] was/were used to complete HW #[QUESTION-NUMBER].[PART-NUMBER] to [DETAIL-HOW-AI-WAS-USED-&-WHAT-YOU-DID-BEYOND-TO-ANSWER-THE-QUESTION]."*
- If AI is used, include a PDF printout of your prompts and the AI tool responses in your GitHub repository.

Note: Compliance with the above guidelines will be taken into account when grading this homework.

HW 1: SIR and SEIR Model Implementation for Pandemic Spread

Objective: Implement and analyze the *susceptible-infectious-recovered (SIR)* compartmental model to understand the dynamics of infectious disease spread. Next, we expand the model to include an exposed compartment, incorporating the effects of births and deaths. Finally, we evaluate how variations in parameters influence model outcomes and discuss the implications for public health strategies.

A) **Model Implementation:** Write a function in MATLAB or Python (preferably within a Jupyter Notebook) to implement the SIR model using the following system of differential equations:

$$\begin{aligned}\frac{dS}{dt} &= -\beta SI,\\ \frac{dI}{dt} &= \beta SI - \gamma I,\\ \frac{dR}{dt} &= \gamma I,\end{aligned}$$

where:

- S is the number of the *susceptible* (those at risk of infection),
- I is the number of the *infected* population (those currently infectious),
- R is the number of the *recovered* population (those who have recovered and are immune),
- β is the *transmission rate*, and
- γ is the *recovery rate*.

Hint: Use numerical methods (e.g., Runge-Kutta, ODE solvers or discretizing the differential equations) to iteratively solve these equations.

B) **SIR Model Simulation:** Simulate the SIR model over a period of 150 days with the following initial conditions and parameters for a total population of $N = 1000$ individuals:

- Initial populations: $S(0) = 999$, $I(0) = 1$, $R(0) = 0$,
- Transmission rate: $\beta = 0.3 \times 10^{-3}$,
- Recovery rate: $\gamma = 0.1$.

Generate a plot showing the dynamics of S, I, and R over time. Label each curve clearly to indicate the compartment it represents.

C) **Analysis and Interpretation:** Analyze the results of your simulation, focusing on the following aspects:

i. *Infection peak*: Identify the point in time where the number of infected individuals $I(t)$ reaches its maximum value. Discuss the factors contributing to this peak.

ii. *Basic reproductive number* R_0: Calculate and interpret the *basic reproductive number* $R_0 = \frac{\beta}{\gamma}$. Explain how R_0 influences the overall dynamics of the pandemic, particularly in terms of infection spread and control.

iii. *Pandemic dynamics*: Describe the general behavior of the S, I, and R populations over time. Discuss how the interactions between these compartments represent the spread and eventual containment of the infection.

D) **SEIR Model with Births and Deaths:** We now expand the model to include exposed individuals and account for births and deaths.

i. Implement the *susceptible-exposed-infectious-recovered (SEIR)* model with the following differential equations:

$$\frac{dS}{dt} = \mu N - \beta SI - \mu S,$$
$$\frac{dE}{dt} = \beta SI - (\sigma + \mu)E,$$
$$\frac{dI}{dt} = \sigma E - (\gamma + \mu)I,$$
$$\frac{dR}{dt} = \gamma I - \mu R,$$

where E is the exposed population, σ is the rate of becoming infectious, and μ is the birth/death rate.

ii. Simulate for both 365 and 1200 days with $S(0) = 990$, $E(0) = 9$, $I(0) = 1$, $R(0) = 0$, and parameters $\beta = 0.3 \times 10^{-3}$, $\gamma = 0.1$, $\sigma = 0.2$, and $\mu = 0.01$. Plot the compartment populations over time.

iii. Discuss the pattern observed in the number of infections in terms of waves of the pandemic.

iv. Discuss the effect of the exposed compartment and birth/death rates on the pandemic dynamics.

E) **Sensitivity Analysis:** Next, let's assess how variations in parameters impact model outcomes and implications for public health.

i. Conduct a sensitivity analysis on the SEIR model from the previous part by varying β (0.1×10^{-3} to 0.5×10^{-3}) and γ (0.05 to 0.2).

ii. Plot the peak infection and total infections over a year for each β and γ combination.

iii. Discuss the implications for public health interventions, relating β to social distancing and γ to medical treatments.

References and further reading:

1. DOI: 10.1109/JSTSP.2021.3129118
2. DOI: 10.48550/arXiv.2003.11371
3. https://github.com/alphanumericslab/EpidemicModeling

HW 2: Agent-Based Modeling of Pandemic Spread

Objective: Develop an agent-based model to simulate pandemic spread dynamics and explore the impact of intervention strategies, such as social distancing, on infection rates and recovery.

A) **Building the Base Model: Infection Dynamics in a Population**

i. **Define the Environment and Initial Conditions**

- Create a 75×75 voxel grid representing a bounded area where individuals (agents) can move and interact.
- Populate the grid with 100 agents, initialized randomly in the following states:
  - 95 agents as *susceptible* (S) – individuals at risk of infection.
  - 5 agents as *infected* (I) – individuals who can transmit the infection.
  - 0 agents as *recovered* (R) – individuals who have recovered and are immune.

ii. **Define Agent Behaviors**

- *Movement:* Each agent moves to a neighboring cell each time step (up, down, left, right, or stays in place). Movement can be random or follow simple rules, e.g., *random walk (Brownian motion)* or *Levy walk*.
- *Transmission:* If a *susceptible* agent shares a cell with an *infected* agent, there is a probability p that the susceptible agent becomes infected.
- *Recovery:* Infected agents have a probability q of recovering at each time step, after which they transition to the *recovered* state.

iii. **Run the Simulation**

- Simulate the model over 200 time steps, recording the population counts in each compartment (*susceptible, infected, recovered*) at each step.
- Plot the number of agents in each state over time to visualize infection spread, recovery, and eventual immunity.

iv. **Sensitivity Analysis**

- Test different values of p (infection probability) and q (recovery probability). For instance, run simulations with $p = 0.05, 0.1$ and $q = 0.02, 0.05$.
- Analyze how changes in p and q impact infection peaks, time to infection peak, and overall population recovery.
- Rerun the simulation with other random initial conditions. Compare the plots across different runs and plot the average of each population subgroup over time. Do you see some average trend across the random initial conditions?

B) **Extending the Model: Social Distancing and Intervention Strategies**

i. **Introduce Social Distancing Measures:** Modify agent movement behavior to simulate social distancing. For instance:

- When agents have a reduced probability of moving at each step, limiting their interactions. Which parameter of the model should be changed to model this?
- When agents actively move away from cells with infected individuals when possible, simulating avoidance behavior. Which parameter of the model should be changed to model this?
- Run the simulation for 200 steps with social distancing measures in place, recording and plotting the average compartment counts across multiple runs (with random initial states) over time as before.
- Which parameter impacts the maximum number of infected subjects at each time window to prevent the healthcare system overwhelm?

ii. **Analyze and Compare Results:** Compare the results of the simulations with and without social distancing. Focus on metrics such as:

- The peak number of infected individuals.
- The time it takes for the infection to peak and for the population to reach a stable recovered state.
- Overall infection spread and duration.
- Discuss how social distancing impacts infection rates and the timing of infection peaks, and consider its implications for real-world public health interventions.

iii. **Additional Sensitivity Analysis**

- Test varying strengths of social distancing (e.g., different probabilities of movement) to explore how stricter or more relaxed distancing affects infection dynamics.
- Record and discuss changes in infection patterns and recovery rates with each level of intervention.

References and further reading:

1. https://youtu.be/gxAaO2rsdIs
2. DOI: 10.1109/JSTSP.2022.3145622
3. https://youtu.be/pasyQympFGE
4. https://github.com/alphanumericslab/EpidemicModeling

HW 3: Model-based Bias Removal in Machine Learning using Synthetic Blood Pressure Data

Objective: Explore bias from imbalanced datasets and evaluate mitigation methods; generate synthetic data for males and females with provided statistics for systolic blood pressure (SBP) and diastolic blood pressure (DBP) using a bivariate normal model; train a binary classifier to predict sex based on SBP and DBP, varying male/female ratios; evaluate model performance (ROC, F1 score, accuracy) and discuss biases arising from imbalances; reflect on the importance of balanced data, challenges in real datasets, and propose bias mitigation strategies.

A.) **Modeling Blood Pressure as a Function of Age:** Figures 3 and 4 in the [preprint] illustrate age-related trends in SBP and DBP, which we aim to capture with an age-dependent model. Accordingly, SBP typically increases with age, potentially leveling off in older ages; and DBP peaks around middle age and then stabilizes or declines.

i. **Propose Mathematical Models for SBP and DBP:** Develop models to describe these trends, choosing between two suggested model forms, or propose your own:

- *Polynomial Regression Model:*

$$\mathrm{SBP}(a) = c_1 a^2 + c_2 a + c_3$$
$$\mathrm{DBP}(a) = d_1 a^2 + d_2 a + d_3$$

  where a is age (in years) and the parameters c_1, c_2, c_3 (for SBP) and d_1, d_2, d_3 (for DBP) will capture curvature over time.
- *Sigmoidal and Gaussian Models:* a sigmoid or Gaussian curve for the BP (SBP or DBP) trends:

$$\mathrm{BP}(a) = \frac{S_{\max}}{1 + e^{-k(a-a_0)}}$$

$$\mathrm{BP}(a) = D_{\max} \cdot e^{-\frac{(a-a_{\mathrm{peak}})^2}{2\sigma^2}}$$

ii. **Model Fitting:** Use the numeric tables for the mean SBB/DBP across age in the [preprint] to fit the models above to the data points:

- Implement the models in Python or MATLAB.
- For the polynomial model, fit coefficients c_1, c_2, c_3 for SBP and d_1, d_2, d_3 for DBP.
- For the sigmoidal-Gaussian model, fit $S_{\max}, k, a_0$ (SBP) and $D_{\max}, a_{\mathrm{peak}}, \sigma$ (DBP).

iii. **Quantitative Evaluation of Model Fit:** Compute the **Mean Squared Error (MSE)** for each model, and calculate R-squared (R^2) values to assess model fit. Plot model curves for SBP and DBP against the plots in the preprint.

iv. **Interpret Model Parameters**

- For the polynomial model, interpret c_1 and d_1 (curvature terms). What is the physical dimension of each parameter?
- For the sigmoidal-Gaussian model, interpret key parameters:
  - $S_{\max}$: Maximum SBP with age.
  - a_0: Age at which SBP reaches half-maximum.
  - $D_{\max}$: Maximum DBP around middle age.
  - a_{peak}: Age of peak DBP.
  - σ: Spread of the Gaussian curve for DBP.

v. **Discussion and Analysis:** Answer the following

i. Which model captures age trends in SBP and DBP better?
ii. How do model parameters reflect physiological blood pressure changes with age?
iii. Discuss limitations in capturing demographic nuances.

B.) **Model-based Bias Removal in Machine Learning using Synthetic Blood Pressure Data:** In the second part, we seek to understand and demonstrate the impact of dataset imbalance on machine learning model performance for BP classification between males and females, and explore strategies to mitigate these biases using synthetic data.

i.) **Synthetic Blood Pressure Data Generation**

- Using a bivariate normal model, generate synthetic SBP and DBP data for a large population of patients. Assume the following statistics for SBP and DBP:
  - **Male:** SBP mean $\mu_{sbp}^{(M)} = 133.0$ mmHg, SD $\sigma_{sbp}^{(M)} = 18.6$ mmHg; DBP mean $\mu_{dbp}^{(M)} = 78.8$ mmHg, SD $\sigma_{dbp}^{(M)} = 12.6$ mmHg; correlation $\rho^{(M)} = 0.45$.
  - **Female:** SBP mean $\mu_{sbp}^{(F)} = 125.8$ mmHg, SD $\sigma_{sbp}^{(F)} = 19.0$ mmHg; DBP mean $\mu_{dbp}^{(F)} = 74.8$ mmHg, SD $\sigma_{dbp}^{(F)} = 12.4$ mmHg; correlation $\rho^{(F)} = 0.5$.
- Generate a dataset of 100,000 samples with M male and F female entries ($M + F = 100,000$) to define the prevalence of each sex.
- Assign a binary label indicating male (1) or female (0).

ii.) **Binary Classification**

- **Data Splitting:** Divide the dataset into training (80%) and testing (20%) sets.
- **Model Training:** Build a binary classifier (e.g., logistic regression, SVM, decision tree) to estimate sex based on SBP and DBP values, only.
- **Evaluation:**
  i. Train the classifier on datasets with varying male-to-female ratios (M and F).
  ii. For each model, plot the ROC curve and compute the F1 score and accuracy.
- **Analysis:**
  i. Discuss classifier performance changes with different male/female data ratios.
  ii. Identify and examine potential biases introduced by varying prevalences.

iii.) **Discussion**

i. Reflect on the importance of balanced datasets in healthcare.
ii. Discuss implications and challenges with real-world unbalanced datasets (by changing M and F).
iii. Suggest and explore strategies to address these challenges in practice, based on the example provided in the class.

iv.) **Bias Mitigation in Training:** Modify your training strategy (e.g., loss function) to systematically reduce sex bias. Explain the reasoning behind the chosen method.

References and further reading:

1. https://arxiv.org/pdf/2306.08451.pdf
2. https://arxiv.org/pdf/2402.01598
3. https://doi.org/10.1016/j.jelectrocard.2022.07.007

HW 4: Biopotential Modeling and Synthetic ECG Generation

Objective: To understand and simulate biopotential signals, focusing on action potential and ECG generation models. We explore dynamic models for action potentials, apply McSharry-Clifford's synthetic ECG model, and evaluate multichannel signal generation, gaining insights into biosignal modeling and physiological implications for machine learning applications.

A.) **Action Potential Dynamics and Oscillations:**

i. **Exploration of Differential Equation for Action Potentials:** The second-order differential equation

$$\ddot{x}(t) + 2\alpha\dot{x}(t) + \omega_0^2 x(t) = 0$$

models oscillations in action potentials. Analyze how varying the parameter α affects stability and oscillation type (damped, unstable, and stable oscillations). Implement this equation in Python/MATLAB for different values of α and ω_0:

- $\alpha > 0$: damped oscillations.
- $\alpha < 0$: unstable oscillations.
- $\alpha = 0$: pure oscillatory behavior.

Plot the phase-plane trajectories ($\dot{x}(t)$ vs $x(t)$) for each case and interpret the differences in dynamics.

ii. **Van der Pol Oscillator for Self-Regulating Dynamics:** Next, implement the Van der Pol model (as a nonlinear extension of the 2nd order model above):

$$\ddot{x}(t) - 2\alpha[1 - x(t)^2]\dot{x}(t) + \omega_0^2 x(t) = 0$$

Choice of parameters: A good choice of α and ω_0 would give you action potentials damping in 10 ms to 100 ms (depending on the choice of parameters), with a damped oscillation with peaks appearing with a period of around 1 s.

Examine how the feedback mechanism (the second coefficient becoming a function of $x(t)$) stabilizes the amplitude. Plot phase-plane trajectories and discuss how this model differs from the simple second-order oscillator.

iii. **Comparing FitzHugh-Nagumo and Van der Pol Models:** Implement the FitzHugh-Nagumo model:

$$\begin{cases} \frac{dw}{dt} = \epsilon(v + a - bw) \\ \frac{dv}{dt} = v - \frac{v^3}{3} - w + I \end{cases}$$

Simulate this model and compare its phase-plane and time-domain waveforms with those from the Van der Pol oscillator. Discuss the advantages and limitations of each in capturing the dynamics of action potentials.

B.) **Synthetic ECG Generation Using McSharry-Clifford's Model:**

i. **Generating Synthetic ECG Signals:**

Using the OSET toolbox in MATLAB (available here), implement the original McSharry-Clifford's synthetic ECG model (or other extensions available on OSET):

$$\begin{aligned} \dot{x} &= \rho x - \omega y, \\ \dot{y} &= \rho y + \omega x, \\ \dot{z} &= -\textstyle\sum_i a_i \Delta\theta_i \exp\left(-\frac{\Delta\theta_i^2}{2b_i^2}\right) - (z - z_0) \end{aligned}$$

Use the parameters defined in the original paper (DOI: 10.1109/TBME.2003.808805) or the polar extension of the model available in OSET. Plot the synthetic ECG in the time domain and analyze its morphology in the phase-plane.

ii. **Parameter Variation Analysis:** Explore how changing parameters a_i and b_i (amplitude and width) impacts the ECG morphology. Experiment with different values and discuss the observed effects.

iii. **Incorporating Heart Rate Variability (HRV):** Introduce heart rate variability by setting ω as a function of time. Use a sinusoidal modulation for ω to simulate changing heart rates. Plot the modified ECG and discuss the effects of HRV on the signal.

C.) **Multichannel ECG Modeling and Stochastic Extensions:** Run and report on the performance of multichannel ECG simulation codes in OSET with stochastic variability and maternal-fetal ECG modeling. Introduce random noise and small parameter deviations to mimic beat-to-beat variability. Discuss the significance of stochastic modeling for generating realistic ECG signals.

References and Further Reading:

1. McSharry, P. E., Clifford, G. D., Tarassenko, L., & Smith, L. A. (2003). A dynamical model for generating synthetic electrocardiogram signals. In IEEE Transactions on Biomedical Engineering (Vol. 50, Issue 3, pp. 289–294). DOI: 10.1109/TBME.2003.808805

2. OSET Toolbox: https://github.com/alphanumericslab/OSET

3. FitzHugh, R. "Impulses and Physiological States in Theoretical Models of Nerve Membrane," *Biophys. J.*, 1961.