Math 5750/6880: Mathematics of Data Science
Project #4
project final report due December 4, 2025

A GitHub repo for this assignment is located here:

https://github.com/math-data-science-course/Project4

A LaTeX Project4 report template is available on the Canvas Project4 page. You should modify this template to produce your project final report in pdf format.

1. (Exploratory Analysis) The PulseDB dataset is a large, cleaned dataset based on MIMIC-III and VitalDB for benchmarking cuff-less blood pressure estimation methods[1]. Read the paper to understand the dataset. The full dataset is available here[2] and a subset is here[3]. Use the provided code to import the PulseDB dataset. Conduct an exploratory analysis. Summarize the dataset and report your findings in your final report. You should include, at least, summary statistics and plots for all data, as well as plots of some example arterial blood pressure (ABP), raw electrocardiogram (ECG) and photoplethysmogram (PPG) signals.

2. (Blood Pressure Prediction) Using the PulseDB dataset, we will build models to predict arterial blood pressure (ABP) from raw electrocardiogram (ECG) and photoplethysmogram (PPG) signals.

First, develop regression models that take ECG and PPG siganls and predict diastolic blood pressure (DBP) and systolic blood pressure (SBP). First try a linear model and then try models of increasing complexity, e.g., fully connected NN, RNN, LSTM, transformer etc. Use the provided train/test data split.

In your final report, discuss your findings. For each model, evaluate your models on the test data and report the ME, SDE, MAE, and R^2 (see paper for definitions). Make plots of the true DBP/SBP vs. your estimated DBP/SBP.

Challenge problem: Develop models for the sequence-to-sequence prediction problem of predicting the full ABP waveform from ECG and PPG signals. Again, first try a linear model to establish a baseline and then increase complexity by considering, e.g., fully connected NN, RNN, LSTM, transformer etc.

3. (Generative Modeling) Using the PulseDB dataset, we will develop generative models for arterial blood pressure (ABP) signals.

First, perform a Principle Component Analysis (PCA) to understand the dimensionality of the ABP signals. Perform an "elbow analysis" to determine the intrinsic linear dimension of the data.

Then, train a fully connected or 1D convolutional autoencoder (AE) to minimize reconstruction loss (MSE between reconstructed and true ABP). Again, perform an elbow analysis to choose a latent dimension. Compare this dimension to the dimension obtained via the PCA analysis. Use the decoder to generate some new ABP signals. Include plots with examples of both data and generated signals in your report. In your final report, discuss your findings.

[1] W. Wang, P. Mohseni, K. L. Kilgore, and L. Najafizadeh, PulseDB: A large, cleaned dataset based on MIMIC-III and VitalDB for benchmarking cuff-less blood pressure estimation methods, *Frontiers in Digital Health*, (2023) https://doi.org/10.3389/fdgth.2022.1090854.

[2] https://github.com/pulselabteam/PulseDB

[3] https://www.kaggle.com/code/mineshjethva/eda-pulsedb

Challenge problem: Develop other generative models for ABP signals, e.g., diffusion-based, flow-based, or GAN models. Train the model to generate realistic ABP signals from noise. Are these generated samples better than those generated using an AE?