Machine Learning Assignment 1 Report

110062219

October 5, 2023

1 Regression Equation for Basic Part

I have adopted the simple, straightforward linear regression model:

$$\mathbf{y} = \Phi(x)\mathbf{w}$$

2 Variables for Advanced Part

First, the subjects were grouped by their IDs and the column of `chart time` was dropped. Then I performed the multiple regression based on the remaining 4 columns: `temperature`, `heart rate`, `respiratory rate` & $\texttt{O}_2$ `saturation`.

3 Difficulty Encountered

1. Initially, I didn't preprocess the data of basic part. As a consequence, the validation loss was unstable and varied, even exceeding 10 sometimes.
2. The result of gradient descent was too far from the one of matrix inversion.
3. There are quite plenty of missing data in the training of advanced part.

4 Solutions & Reflections

1. I used to try higher-degree polynomial regression but in vain. After I filtered out the outliers that are out of 3 times the standard deviation, the results looked well.
2. In the beginning, $\mathbf{w}$ was initialised completely randomly and it was hard to tune the hyper-parameter *learning rate*. Therefore, I cheated a bit and initialised $\mathbf{w}$ with the approximate values obtained from matrix inversion. I believe that for real-life challenges, it would be a good idea that we first do matrix inversion of a small batch to initialise $\mathbf{w}$.
3. I learned the **"hot-deck imputation"** which is suitable for time-series data and then put it into practice.

Overall, I built the gradient descent model from scratch by myself without any specific package in this assignment, which was really intriguing.