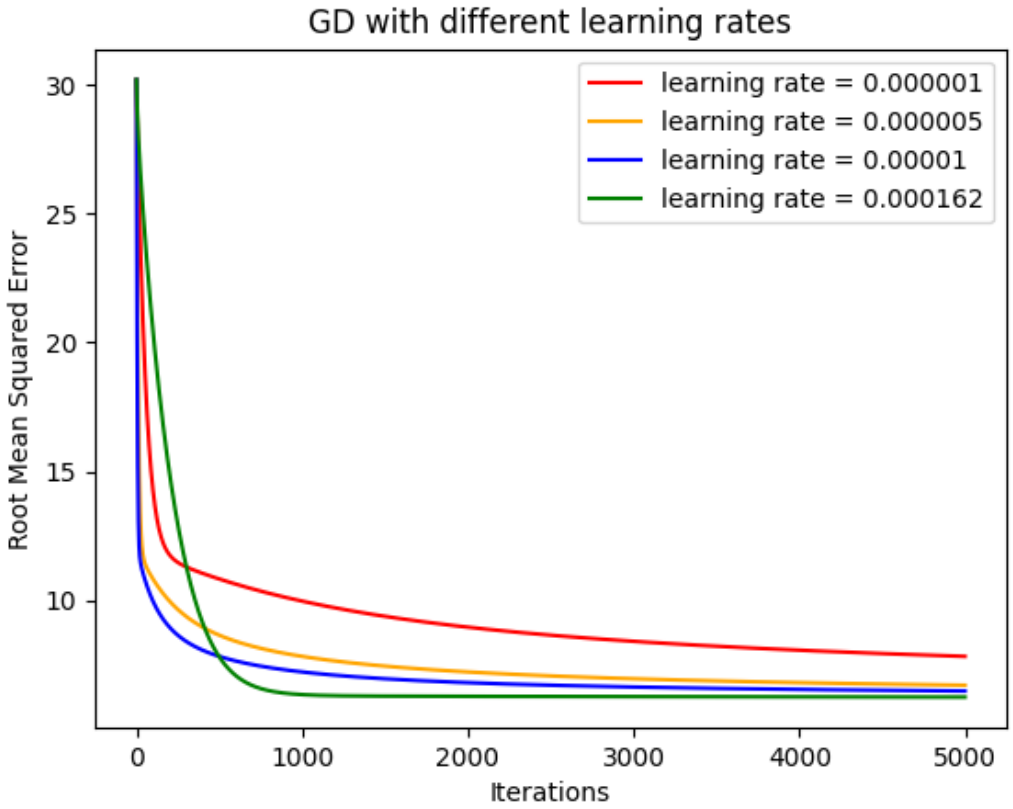
Machine Learning 2020 Spring - HW1 Report

學號:B07902064 系級:資工二 姓名:蔡銘軒

1. (2%) 使用四種不同的 learning rate 進行 training (其他參數需一致)，作圖並討論其收斂過程（橫軸為 iteration 次數，縱軸為 loss 的大小，四種 learning rate 的收斂線請以不同顏色呈現在一張圖裡做比較）



圖中可以觀察到當learning rate是0.00001或是0.000005時，表現都穩定而且差距不大。而當learning rate較大時，例如圖中的0.000162，在前幾輪的表現是比較差的。可能的原因是較大的learning rate使下降不穩定，可能跨得太大步反而跨過最小值的點到了另一端。在learning rate較小時，例如圖中的0.000001，則收斂的相較其他三者都緩慢許多。

2. (1%) 比較取前 5 hrs 和前 9 hrs 的資料（5·18 + 1 v.s 9·18 + 1）在 validation set 上預測的結果，並說明造成的可能原因。

	5hrs	**9hrs**
Training set	5.699826467085332	5.5384664796340255
Validation set(20%)	5.6977157778702825	5.621467762358471

取5hrs所形成的model是取9hrs所形成的model的子集，因此9hrs在training data上表現的比較好是可以預期的，但同時也因為model比較複雜，我們也可以觀察到他的variance比較大。另一方面，5hrs的model較簡單，因此variance較小，但可能會發生underfitting。從這次的結果來看，5hrs在validation set上的表現較9hrs差，可能是model不夠強大造成的underfitting。

3. 比較只取前 9 hrs 的 PM2.5 和取所有前 9 hrs 的 features（5·18 + 1 v.s 9·18 + 1）在 validation set上預測的結果，並說明造成的可能原因。

	PM2.5 only	**All features**
Training set	5.94311320645387	5.376935636228499
Validation set(20%)	5.893646481375374	5.811878246361573

所有features的model雖然在training data上會有非常好的表現，但也因為他的model complexity很高，很容易出現overfitting的現象。反之只有PM2.5的model則因為model過於簡單，容易出現underfitting。在這次的實驗中，從validation set的成果來看，PM2.5的model表現較差，可以推測是model過於簡單產生的underfitting。但我們也觀察到所有features的model在validation set上的表現比在training set上差了不少，可能說明他也發生了overfitting。

4. (2%) 請說明你超越 baseline 的 model(最後選擇在Kaggle上提交的) 是如何實作的（例如：怎麼進行 feature selection, 有沒有做 pre-processing、learning rate 的調整、advanced gradient descent 技術、不同的 model 等等）。

   feature selection: 我對每一項feature分別對PM2.5作相關係數的分析，並查詢了一些PM2.5的形成/來源的資料，最後選擇['CO', 'NO', 'NO2', 'NOx', 'O3', 'PM2.5', 'RAINFALL', 'SO2', 'THC', 'WIND_SPEED']作為model的feature。

   pre-processing: 我觀察到train.csv裡面有一些數值會是負的，我認為這並不是合理的數據，因此我將含有負數數值的資料刪除，不加入training的過程。

   gradient descent: 我使用了Nadam的方式，進行了5000次的training。

   不同model: 根據feature的相關係數分析以及我查到關於PM2.5的資料，我挑選出一些我認為較有可能性的feature組合。我將training data隨機切出1000筆資料作為validation set，進行約五次的validation，最後選擇平均起來表現最好的model。

Collaborator: b07902047 羅啟帆