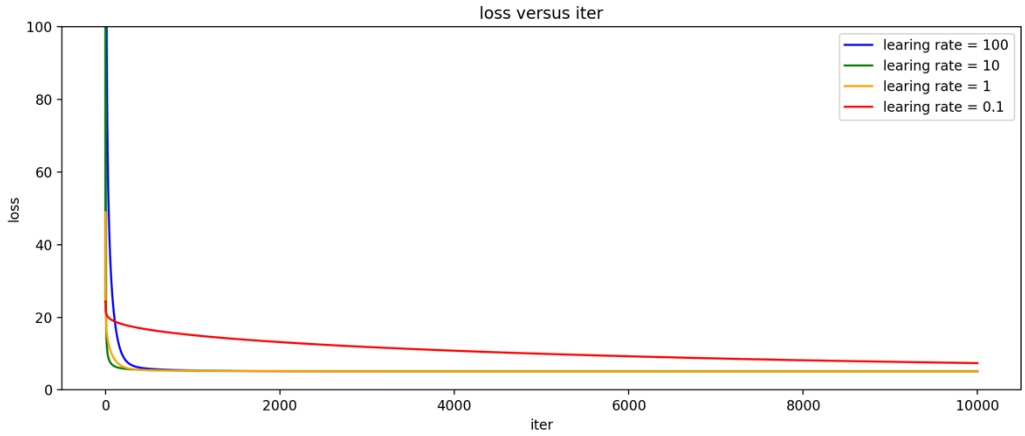
學號：B06902042 系級：資工二 姓名：劉愷為

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）。



從上圖來看，learning rate 為 0.1 時，每次更新的幅度太小，會導致收斂速度極慢。而 learning rate 為 1 跟 10 為比較好的結果，收斂速度也比較快。最後，learning rate 為 100 時，更新幅度會太大，造成參數容易走偏。由於我們使用 Adagrad，會使 learning rate 隨時間降低，最後也會收斂到合理 loss。

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。

Validation (20%), learning rate = 100, itr = 1000

前 5 hrs	前 9 hrs
5.7537144915068925	5.912205466286509

在這邊取前 5 小時會比取前 9 小時好，我認為可能的原因是因為 9 小時這個區間太長了。有可能只有近 3~5 個小時的資料才會真正影響 pm2.5 的值，取太長反而會造成預測的精準度下降。

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

Validation (20%), learning rate = 100, itr = 1000

前 9 hrs 的 PM2.5	取所有前 9 hrs 的 features
5.86461175212293	5.912205466286509

在這邊只取 pm2.5 會比全部都取好一點點，我認為可能的原因是 features 當中可能會有與 pm2.5 完全不相干的值。這會造成預測的精準度下降。而兩者只差一點點的原因，我猜應該是有相關的資料與不相關的資料的影響相互抵銷。如果透過分析，把不相干的資料全部剔除，預測的結果應該會好很多。

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）。

- Test.csv 和 train.csv 中，pm2.5 會有一些資料為-1。將其數值改為前後的平均
- 把 AMB_TEMP, NO, WD_HR, WIND_DIREC, WIND_SPEED 移除
- 使用 gradient descent with adagrad, Learning rate = 7, Itr = 50000