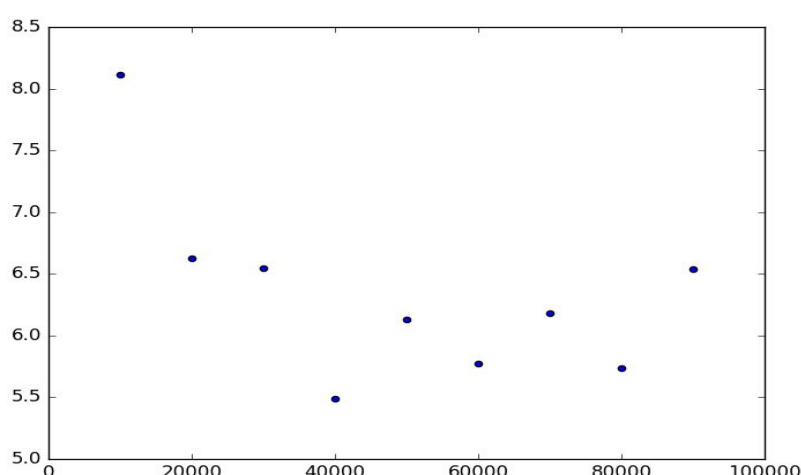


1. 請簡明扼要地闡述你如何抽取模型的輸入特徵 (feature)

答：將每個月連續20天拉成一個時間軸，每9小時取一組18測項x9小時的特徵，另外還加入了pm2.5, pm10, NO₃, NO_x, O₃, SO₂的平方項以及pm2.5xpm10的特徵，再將全部時間點的各測項進行標準化（減掉平均再除以標準差），最後再將和pm2.5相關係數的絕對值小於0.2的特徵移除掉。每一個feature的pm2.5 label並沒有進行標準化。

2. 請作圖比較不同訓練資料量對於PM2.5預測準確率的影響

答：此圖為batch gradient descent中iteration的數目從10000, 20000, ..., 到90000對validation data的損失作比較，validation data為隨機從training data取樣1/10。



從中我們可以發現validation loss呈現小幅度振盪的情況，而選取validation loss最小的iteration數(iter=40000)的模型來測試測資，發現在public score上為5.91501，比hw1_best.sh的模型分數高（5.85519）。

3. 請比較不同複雜度的模型對於PM2.5預測準確率的影響

答：我們以兩種不同複雜度的模型進行比較，這兩種分別為一次和二次的回歸模型：

$$Y = WX$$

$$Y = W_1X + W_2X^2$$

其中Y為預測值，X為輸入特徵，W, W₁, W₂為模型參數。

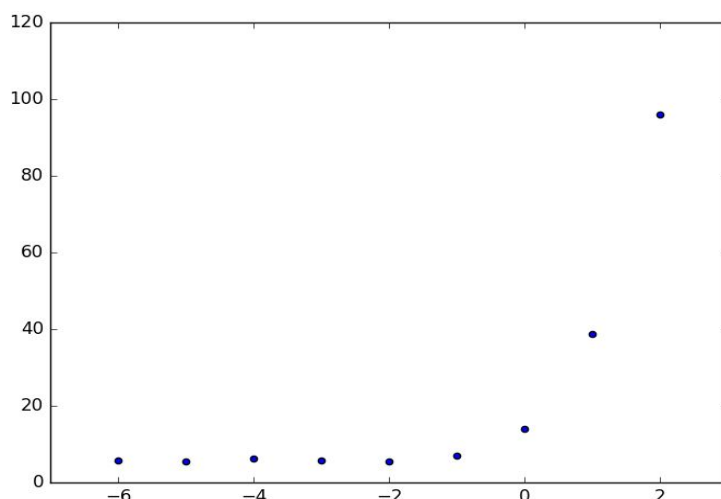
一次的回歸模型抽取feature的方式並沒有加入任何平方項，以stochastic gradient descent加上RMSprop的gradient descent方法，在kaggle public score上為5.89229。

而二次的回歸模型我們將二次的特徵和一次的特徵各自進行標準化（standardization），其餘參數皆和一次回歸模型相同，在kaggle public score上為6.22982。

4. 請討論正規化(regularization)對於PM2.5預測準確率的影響

答：在此加入正規化後的損失函數為 $L = (y^n - w \cdot x^n)^2 + \lambda |w|$ ，此圖為正規化係數（λ）從10⁻⁶, 10⁻⁵, 10⁻⁴, 10⁻³, 10⁻², 10⁻¹, 1, 10, 100分別對validation data的損失作比較，validation

data為隨機從training data取樣1/10, x軸為 $\log_{10}(\lambda)$, y軸為validation loss。其餘參數皆和hw1_best.sh中所執行之模型相同。



從中我們可以發現正規化係數在等於10的時候有較大的validation loss, 而選取 validation loss最小的正規化係數($\lambda=10^{-2}$)的模型來測試測資, 發現在public score上為 5.85175, 比加入正規化之前 (hw1.sh) 的模型分數還低 (5.89229)。

5. 在線性回歸問題中, 假設有 N 筆訓練資料, 每筆訓練資料的特徵 (feature) 為一向量 x^n , 其標註(label)為一存量 y^n , 模型參數為一向量 w (此處忽略偏權值 b), 則線性回歸的損失函數(loss function)為 $\sum_{n=1}^N (y^n - w \cdot x^n)^2$ 。若將所有訓練資料的特徵值以矩陣 $X = [x^1 \ x^2 \ \dots \ x^N]$ 表示, 所有訓練資料的標註以向量 $y = [y^1 \ y^2 \ \dots \ y^N]^T$ 表示, 請以 X 和 y 表示可以最小化損失函數的向量 w 。

答：

令 $L = \sum_{n=1}^N (y^n - w \cdot x^n)^2$, 右式展開後 L 為：

$$L = \sum_{n=1}^N (y^n)^2 - 2w \sum_{n=1}^N x^n y^n + w^2 \sum_{n=1}^N (x^n)^2$$

將總和的式子以矩陣表示, 則 L 會變成：

$$L = y^T y - 2(XW)^T y + W^T X^T X W$$

將 L 對模型參數 w 做偏微分, 令其為0求最小值：

$$\frac{\partial L}{\partial W} = 2X^T X W - 2X^T y = 0$$

由上式得知, 可以最小化 L 的向量 w 為：

$$W = (X^T X)^{-1} X^T y$$