ML hw1 Report B03202047 物理三 王昱翔

1 Linear regression function by Gradient Descent

I use Adam as my optimizer.

Code:

‘’’

loss <- dot(y - dot(w, x) + b)^2 + λw^2

grad_w <- -2*dot(y – dot(w, x)) * (-x) + 2λw

grad_b <- -2*dot(y – dot(w, x))

#Init:

m_0 <- 0

v_0 <- 0

t <- 0

#update

t <- t + 1

lr_t <- learning_rate * sqrt (1 – beta2^t) / (1-beta1^t)

m_t <- beta1 * m_{t-1} + (1 - beta1) * grad_w

v_t <- beta2 * v_{t-1} + (1 - beta2) * grad_w * grad_w

m_t <- beta1 * m_{t-1} + (1 - beta1) * grad_b

v_t <- beta2 * v_{t-1} + (1 - beta2) * grad_b * grad_b

w <- w - lr_t * m_t / (sqrt(v_t) + epsilon)

’’’

2 Method

(1) Training set and Validation set

I divided train.csv into 5652 size training set. It is because there are 24 hrs * 240 days = 5760 size, every hour’s PM2.5 and the previous 9 hours parameters form a training size. The previous 9 hours of the first 9 hours in every month don’t exist, therefore the training size is 5760-9*12= 5652. The validation size is 1000 when doing cross validation.

(2) Features

I take 9 kinds of parameters in 9 hours, therefore my weight dimension is (9*9,1) = (81, 1). I chose my features of pm2.5_model by the following steps.

- Feed only one parameters in 9 hours(1*9) as features of the model, train the models in 10000 iterations.

- Calculate the average loss in validation set, pick 5 least validation loss parameters as pm2.5_model features
- Add the least loss parameters which are not in the pm2.5_model features until the validation loss start to increase.
- Also I have ask my Atmospheric department roommate for feature selection suggestion.

Features
AMB_TEMP
CO
NMHC
NOx
O3
PM10
PM2.5
SO2
WIND_SPEED

(3) Hyper parameters

Learning rate	1e-3
Epsilon*	1e-8
Beta1*	0.9
Beta2*	0.999
iteration	30000

Table 1: Hyper parameters

*: Hyper parameters in Adam optimizer

(4) Eliminate noise with larger error in training set

After finish training with 30000 iteration, I eliminate the training data whose abs(y – (np.dot(w,x)+b)) > 11, and then train the model again.

3 Discussion on regularization

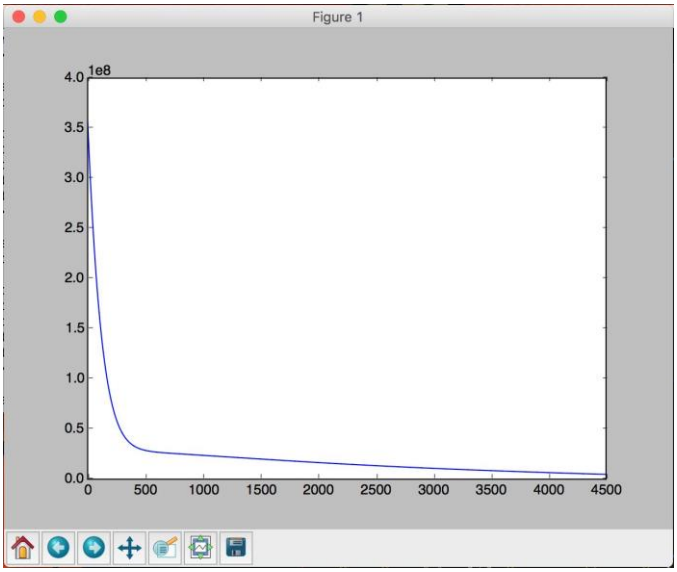
λ value	Cross validation loss
0.6	6.31385826
0.8	6.31870264
1.0	6.04509282

1.2	7.04331021
1.4	6.46920103

Table 2 Different λ

In order not to be overfitting, I did regularization on my model. I choose the range of λ from 0.6-1.4. According to table 2, when λ = 1.0 has significant better validation score than others. We can see that if the λ is less than 1.0, our training model is seemed to be overfitting to the training data. If the λ is larger than 1.0, the validation loss will begin to increase. Therefore, I choose λ = 1.0 in my pm2.5_model.

4 Discussion on learning rate



I use normal gradient descent first. I choose learning rate=1e-6, and the loss exploded. Then I choose learning rate=1e-8, then the loss started to decrease. If I choose adam as my optimizer, Somehow I have checked the loss of learning rate = 1 , 1e-1, 1e-2,1e-3,1e-4. Not like normal gradient descent, loss wouldn't exploded even if learning rate =1, it just kept jumping up and down. If I want the loss to decrease in a steady and adequate rate, the loss should be 1e-3 according to my experiment

5 Extra Discussion

I would like to have some extra discussion below.

- I think shuffle the training set may be better, because the pm2.5 may depend on weather.

- Also I think the validation set should randomly choose from the training set.
- I have tried feature scaling, but it didn’t improve on the test set. I think maybe feature scaling may help when the value of features are much bigger than others.
- I have added the nonlinear term of pm2.5 features, but it didn’t improve either. I really wonder how to select nonlinear features.