HW1

Machine Learning, 2016 Fall

R05921075, 電機一, 鄭凱文

Describe your method

1. Read train data

   Ignore “Rainfall”. Use the rest 17 arguments as one vector.

   12 month, 20 day per month, 24 hour per day. Total 12*20*24 = 5760 vectors.
2. Read test data

   240*9 = 2160 vectors.
3. Extract train feature

   Extract N hours ago data to predict PM2.5 in now. N can only set with 1~9 hours, because there are only 9 hours ago test data.

   Merge continuous N hours train data as one feature. However, it should notice that the raw data is not continuous between each month (e.g., 1/20 24:00 and 2/1 1:00 is not continuous) So, it should avoid to merge this data into a feature.

   I set N=3, so each feature has 3*17=51 arguments.
4. Normalize train feature

   Normalize each element to normal Gaussian distribution(μ=0, σ=1). This method can solve problem on adative η. Each element has the same distribution, so η can set a fixed value.
5. Train
6. Test
7. Output test result

Linear regression function by Gradient Descent

Input: train feature X and corresponding Y
Output: regression weight W and bias b
1 W^0={$W|W_i = 0.05$, for each i}
2 $b^0 = 1$
3 $\eta = 0.00001$
4 iteration $i = 0$
5 **while** *true* **do**
6 $\quad i = i + 1$
7 $\quad W^{i+1} = W^i - \eta\nabla L(W^i, b^i)$
8 $\quad b^{i+1} = b^i - \eta\nabla L(W^i, b^i)$
9 $\quad$ **if** $i\%100 == 0$ **then**
10 $\quad\quad$ **if** $|\nabla L(W^i, b^i)| < 0.01$ **then**
11 $\quad\quad\quad W = W^i$
12 $\quad\quad\quad b = b^i$
13 $\quad\quad\quad$ break
14 $\quad\quad$ **end**
15 $\quad$ **end**
16 **end**
17 return W, b;

η=0.00001 is a try-and-error result.

(W, b) initialize with a arbitrary value.

In line 9~15, check gradient norm per 100 iterations to save calculation time. If the norm less than 0.01(I think it is small enough), and return this (W, b).

Discussion on regularization

I try 3 different value of λ to gradient descent.

λ	0.00001	0.0001	0.001
Gradient Norm in first 300 iteration	iteration: 100 4477.99728793 iteration: 200 1932.02819628 iteration: 300 1009.08735248	iteration: 100 4510.79969254 iteration: 200 1984.76014311 iteration: 300 1071.82762237	iteration: 100 4868.55636348 iteration: 200 2598.97386347 iteration: 300 1869.15288337
Score in Kaggle	6.08910	6.10720	6.23315

From gradient norm, we can find that larger λ make descending rate slow. It means error in train set becomes higher. The kaggle score does not improve with larger λ, it should try more continuous value to see the trend between prediction error and λ.

Discussion on learning rate

η=0.00001 is a try-and-error result. Too large η will induce infinite gradient norm, and too small η make the descending rate slow. Because of normalization, it is no need to use adaptive learning rate. We can find that large gradient norm make the descending rate fast, and small gradient norm make the descending rate slow. That is to say, the descending step is proportional to gradient norm. In conclusion, descending rate is always proportional to gradient norm, so, with normalization, it is can do well with fixed η.

Other method

We all know linear regression has a closed form solution. Without lose of generalization, we can always use pseudo-inverse method to calculate linear regression weight. The merit of use pseudo-inverse is it can avoid singular matrix problem.

$$W_{linear\ reg} = YX^{+}$$

In kaggle best, I use this method to predict. I also calculate the gradient at $W_{linear\ reg}$. The result is approach 10^{-8} . In gradient descend method, I believe it can descend to this quautity, but gradient norm descending less than 0.01 take 4 hours more in calculation.

Conclusion

In actually, we should try different hours ago (1~9 hours) models to predict predict PM2.5, and choose model by validation data. However, the training time takes too long. As a result, I only train 3 hours ago model and only descending to gradient norm < 0.01. If we want to make a prediction model better and have enough time, we should try all possible models in the future.