Content

[1. Searching for Dataset 2](#_Toc487631267)

[2. Dataset processing 2](#_Toc487631268)

[2.1 Retrieve the dataset 2](#_Toc487631269)

[2.2 NA data processing 2](#_Toc487631270)

[2.3 Scaling 3](#_Toc487631271)

[3. Model selection 3](#_Toc487631272)

[3.1 Model of MLP 4](#_Toc487631273)

[3.1.1.Single variate timeseries forecasting PM2.5 4](#_Toc487631274)

[3.1.2 Multivariates timeseries forecasting PM2.5 by MLP 4](#_Toc487631275)

[3.2 Model of RNN with LSTM using timesteps regression 5](#_Toc487631276)

[3.2.1 Single variate timeseries forecasting PM2.5 5](#_Toc487631277)

[3.2.2 Multivariate timeseries forecasting PM2.5 5](#_Toc487631278)

[4. Relationship between the PM2.5 and other variates. 6](#_Toc487631279)

[4.1 Find the relationship between the wind speed and PM2.5 values 6](#_Toc487631280)

[4.2 Relationship between the PM2.5 and dew point 7](#_Toc487631281)

[4.3 Relationship between PM2.5 and Dew point, temperature, pressure and wind speed 7](#_Toc487631282)

# Searching for Dataset

I would like to do something novel and interesting. Based on consideration that I do a short time study about AI, I havenot much time to read a lot of papers. I am considering dataset selection from following aspects:

1. The supervising learning for categorizing issues.
2. Time series for forecasting.
3. Recognize images (classification issues as well).

For the first one, it is a categorizing issue, all I learn may be not enough for me to have some novel ideas for the deep learning. I need to spend a lot time in reading papers first.

For the third one, there may be a lot of ideas in images. I did not consider this one because it is usually taking a long time to train a large quantity of images like autoencodering images we have learned in tutorial 6.

I chose the second one because in our tutorial 4 and 5 for Timeseries forecasting, we only did the single variate forecasting. I would like to try multi variates to forecasting a single value. Multivariates timeseries forcasting may have many actual values such as stock, weather forecasting, business and so on. When I went on UCI website for dataset, the Beijing PM2.5 came up attracting my eyes. We know air quality is impacted by dew point, PM2.5, wind, temperature, humidity and so on. It has practical meaning to forecast PM2.5 especially in China. That’s the reason I choose the Beijing PM2.5 dataset(PRSA\_data\_2010.1.1-2014.12.31.csv).

# Dataset processing

## 2.1 Retrieve the dataset

After I got the dataset, extract the interesting columns I may need for deep learning by following code:

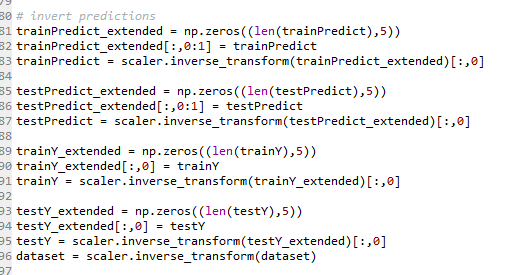


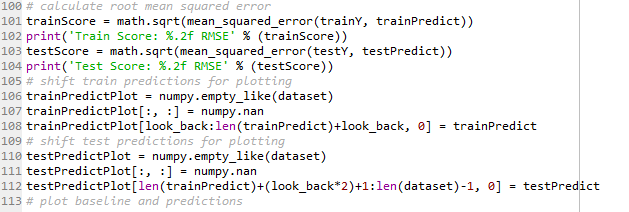
## 2.2 NA data processing

Due to many NA data is consecutive, I rudely throw all NA data directly. But the data is based on hours, it may impact. I don’t have much time to try fill the value of NA which is not a whole day 24 hours consecutive. I guess if I fill the value of some NA which is not consecutive to make a 24 hour whole day data consecutively, it may improve the performance when tune the look back as 24 or 48. I don’t have much time to try.

## 2.3 Scaling

Dataset scaling: The purpose of data scaling is to reduce the data sensitivity of the model in order to get a stable and believable performance. When use the multi-variates to forecast the single output and plot the output, it need some tips for scaling. The whole complete dataset is scaled by the same scaler. Before do inverse transform for the output, it needs to reshape the single output has the same data structure with the whole dataset in order to use the same scaler for inverse transform.





# Model selection

I have learned ANN(MLP), CNN and RNN. For the timeseries forecasting, we have done the practical both on ANN and RNN. I have tried to do convolution as well, but I am failed during input shape issues.

So I do MLP and RNN.

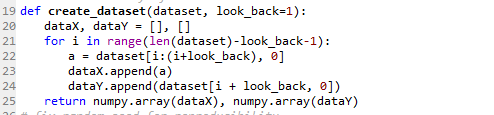
## 3.1 Model of MLP

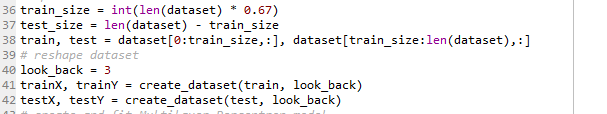
### 3.1.1.Single variate timeseries forecasting PM2.5 with MLP

Code file: Timeseries\_univariate\_forecasting\_pm2.5\_MLP.py

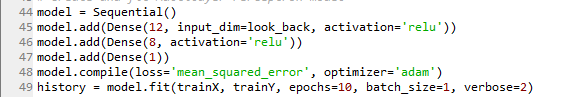
|  |  |  |
| --- | --- | --- |
| Input | Output | Look\_back |
| PM2.5 value | PM2.5 value | 3 and 10 |

Dataset processing:





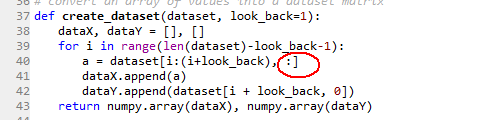
Model:



### 3.1.2 Multivariates timeseries forecasting PM2.5 by MLP

Seems I cannot do that for the reason that dense at the first layer only has 2 dimensions as input, but for multifeatures there is 3 dimensions samples, look\_back, features) as input.

As multi-features, I create dataset as below, it will generate the 3 dimensions train dataset with multi-inputs.



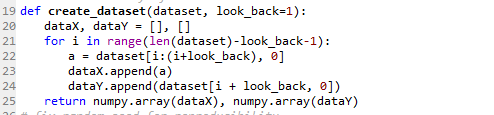
## 3.2 Model of RNN with LSTM using time steps regression

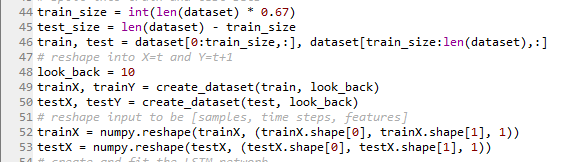
### 3.2.1 Single variate time series forecasting PM2.5 using time steps regression

Code file: Timeseries\_univariate\_forecasting\_pm2.5\_LSTM.py

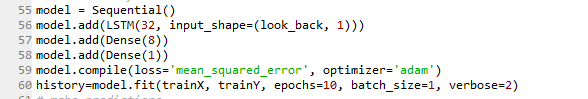
|  |  |  |
| --- | --- | --- |
| Input | Output | Look\_back |
| PM2.5 value | PM2.5 value | 3 and 10 |

Dataset Processing:





Model:



### 3.2.2 Multivariate timeseries forecasting PM2.5 using time steps regression

Code files: Multifeatures\_TimeSeriesLSTM\_forecastingPM2.5.py

This is most interesting part I want to do, but seems RMSEs don’t have obvious advantages compared with single variate forecasting 2.5 by LSTM or MLP. I expect to try more but it takes so long time to run a round.

The input:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| PM2.5 | Dew point | Temperature | Pressure | Wind speed |

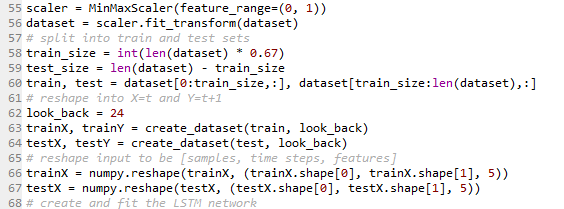
The output

|  |
| --- |
| PM2.5 |

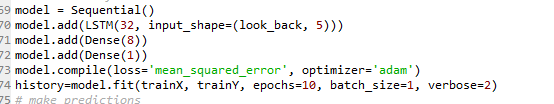
Look\_back: 3 and 10, 24(the data is hour-based).

Dataset Processing:





Model:



# Relationship between the PM2.5 and other variates.

I would like to find more interesting things in these parts, but the RMSEs are not so good.

## 4.1 Find the relationship between the wind speed and PM2.5 values

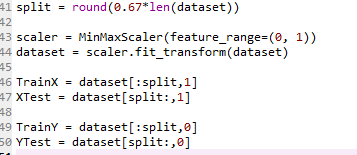
Code file: PM2.5\_Relationshipwith\_WindSpeed\_MLP.py

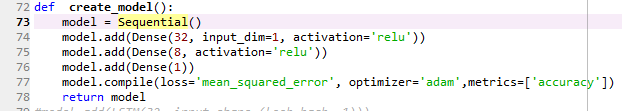
The model of MLP

|  |  |
| --- | --- |
| Input | output |
| WindSpeed | PM2.5 |

Code as below:





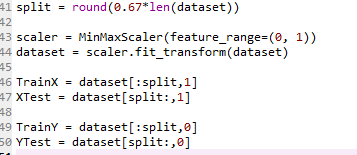


## 4.2 Relationship between the PM2.5 and dew point

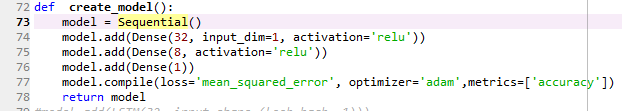
Code file: PM2.5\_Relationshipwith\_dewpoint\_MLP.py

|  |  |
| --- | --- |
| Input | output |
| Dew point | PM2.5 |

Dataset 



Model:



## 4.3 Relationship between PM2.5 and Dew point, temperature, pressure and wind speed

Code file: PM2.5\_Relationshipwith\_other4\_variates\_MLP.py

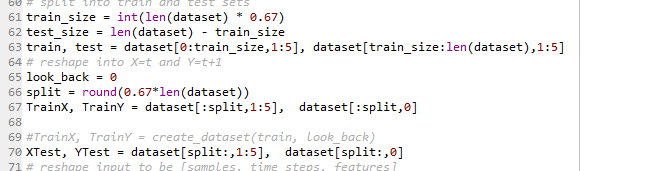
The input:

|  |  |  |  |
| --- | --- | --- | --- |
| Dew point | Temperature | Pressure | Wind speed |

The output

|  |
| --- |
| PM2.5 |

Dataset Processing:



Model:

