空气污染预报

在这个小例子中，我们将使用空气质量数据集。这是一个数据集，在美国驻北京大使馆五年内报告天气和污染水平。数据类型包括日期时间、称为PM2.5浓度的污染物、以及天气信息、包括露点（露点温度）、温度、压力、风向、风速和累积的降雪小时数。原始数据中的完整功能列表如下：

      1.**NO**：行号。

      2.**year**：年。

      3.**month**：月。

      4.**day**：日。

      5.**hour**：小时。

      6.**pm2.5**：PM2.5浓度。

      7.**DEWP**：露点温度。

      8.**TEMP**：温度。

      9.**PRES**：压力。

      10.**cbwd**：风向。

      11.**Iws**：风速。

      12.**ls**：积雪的时间

      13.**Ir**：累积的下雨时数

我们可以使用这些数据并构建一个预测问题，基于天气条件和前几个小时的污染，我们预测在下一个小时的污染。你可以从UCI Machine Learning Repository下载数据集—[北京PM2.5数据集](https://archive.ics.uci.edu/ml/datasets/Beijing+PM2.5+Data)。下载数据集并将其放在你当前的工作目录中，文件名为“raw.csv”。

基本数据准备

第一步，我们必须清洗数据。

以下是原始数据集的前几行。

No,year,month,day,hour,pm2.5,DEWP,TEMP,PRES,cbwd,Iws,Is,Ir

1,2010,1,1,0,NA,-21,-11,1021,NW,1.79,0,0

2,2010,1,1,1,NA,-21,-12,1020,NW,4.92,0,0

3,2010,1,1,2,NA,-21,-11,1019,NW,6.71,0,0

4,2010,1,1,3,NA,-21,-14,1019,NW,9.84,0,0

5,2010,1,1,4,NA,-20,-12,1018,NW,12.97,0,0

第一步是将日期时间信息整合为一个日期时间，以便我们可以将其用作Pandas的索引。我们需要快速显示前24小时的pm2.5的NA值。因此，我们需要删除第一行数据。在数据集中还有几个分散的“NA”值；我们现在可以用0值标记它们。

以下脚本加载原始数据集，并将日期时间信息解析为Pandas Data Frame索引。“No”列被删除，然后为每列指定更清晰的名称。最后，将NA值替换为“0”值，并删除前24小时。

from pandas import read\_csv

from datetime import datetime

# load data

def parse(x):

return datetime.strptime(x, '%Y %m %d %H')

dataset = read\_csv('raw.csv', parse\_dates = [['year', 'month', 'day', 'hour']], index\_col=0, date\_parser=parse)

dataset.drop('No', axis=1, inplace=True)

# manually specify column names

dataset.columns = ['pollution', 'dew', 'temp', 'press', 'wnd\_dir', 'wnd\_spd', 'snow', 'rain']

dataset.index.name = 'date'

# mark all NA values with 0

dataset['pollution'].fillna(0, inplace=True)

# drop the first 24 hours

dataset = dataset[24:]

# summarize first 5 rows

print(dataset.head(5))

# save to file

dataset.to\_csv('pollution.csv')

运行该示例打印转换的数据集的前5行，并将数据集保存到“pollution.csv”。

Date pollution dew temp press wnd\_dir wnd\_spd snow rain

2010-01-02 00:00:00 129.0 -16 -4.0 1020.0 SE 1.79 0 0

2010-01-02 01:00:00 148.0 -15 -4.0 1020.0 SE 2.68 0 0

2010-01-02 02:00:00 159.0 -11 -5.0 1021.0 SE 3.57 0 0

2010-01-02 03:00:00 181.0 -7 -5.0 1022.0 SE 5.36 1 0

2010-01-02 04:00:00 138.0 -7 -5.0 1022.0 SE 6.25 2 0

下面的代码加载了“ pollution.csv ”文件，并将每个系列作为单独的子图绘制，除了风速dir，这是分类的。

from pandas import read\_csv

from matplotlib import pyplot

# load dataset

dataset = read\_csv('pollution.csv', header=0, index\_col=0)

values = dataset.values

# specify columns to plot

groups = [0, 1, 2, 3, 5, 6, 7]

i = 1

# plot each column

pyplot.figure()

for group in groups:

pyplot.subplot(len(groups), 1, i)

pyplot.plot(values[:, group])

pyplot.title(dataset.columns[group], y=0.5, loc='right')

i += 1

pyplot.show()

运行示例创建一个具有7个子图的程序，显示每个变量的5年数据。

