

**机器学习工程报告**



**课 程：** 机器学习基础

**姓 名：** 王家琪

**学 号：** 2018217918

**完成时间：** 2020年12月22日

一． 工程摘要与每人贡献

摘要：

收集合肥地区过去一段时间（例如过去一年每个月的平均值）的空气质量（例如pm2.5值），然后构建回归模型，能够预测今年某个月的空气质量值。

每人分工：

|  |  |  |  |
| --- | --- | --- | --- |
| 姓名 | 角色  （组长，组员，独自完成） | 工作量  比例 | 负责内容 |
| 王家琪 | 独自完成 | 100% | 收集数据集；算法代码实现；报告撰写 |

二、研究背景与意义

PM是英文particulate matter（颗粒物）的首字母缩写，PM2.5指空气动力学当量直径小于或等于2.5微米（一微米等于百万分之一米）的悬浮颗粒物。它在大气中滞留时间长，传输距离远，含多种由于有害物质，而且与其他空气污染物存在着复杂的转化关系。PM2.5易于滞留在终末细支气管和肺泡中，其中某些还可以穿透肺泡进入血液，也更易于吸附各种有毒的有机物和重金属元素，对健康的危害极大。

PM2.5除来自自然界的风沙尘土、森林火灾、海水喷溅等，更主要来自工业生产、公路扬尘、建筑扬尘以及人类生产生活使用的能源燃烧等。

PM2.5主要对呼吸系统和心血管系统造成伤害，包括呼吸道受刺激、咳嗽、呼吸困难、降低肺功能、加重哮喘、导致慢性支气管炎、心律失常、非致命性的心脏病、心肺病患者的过早死。老人、小孩以及心肺疾病患者是PM2.5污染的敏感人群。如果空气中PM2.5的浓度长期高于10微克/立方米，死亡风险就开始上升。浓度每增加10微克/立方米，总的死亡风险就上升4%，得心肺疾病的死亡风险上升6%，得肺癌的死亡风险上升8%。

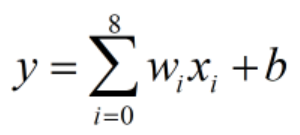
2012年10月11日，中国国家环境保护部副部长吴晓青表示，新的《环境空气质量标准》颁布后，环保部明确提出了新标准实施的“三步走”目标。按照计划，2012年年底前，京津冀、长三角、珠三角等重点区域以及直辖市、计划单列市和省会城市要按新标准开展监测并发布数据。截至目前，全国已有195个站点完成PM2.5仪器安装调试并试运行，有138个站点开始正式PM2.5监测并发布数据。

随着城市对PM2.5进行监测，人们在出行的时候更加的注意保护措施，环保意识也更加的强烈，自己的日常行为也在不断的改善。对空气中悬浮颗粒物、可吸入颗粒物等的关注程度得到了大大的提高，这在一定的程度上也促进国家对PM2.5监测和解决的进行。

三、模型方法

**1.模型**

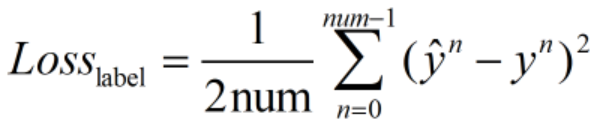
采用普通的线性回归模型，使用每个数据帧样本中的9个PM2.5含量值：



为对应数据帧中第i个PM2.5含量，为其对应的权重值，为偏置项，为该数据帧样本的预测结果。

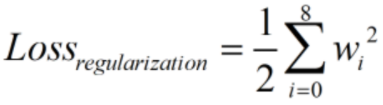
**2.损失函数**

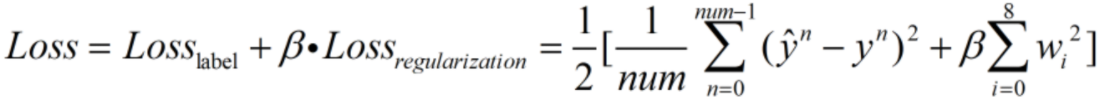
用预测值与label之间的平均欧式距离来衡量预测的准确程度，并充当损失函数：



为第n个label，为第n个数据帧的预测结果，为参加训练的数据帧样本个数。

加入正则项，防止过拟合，：

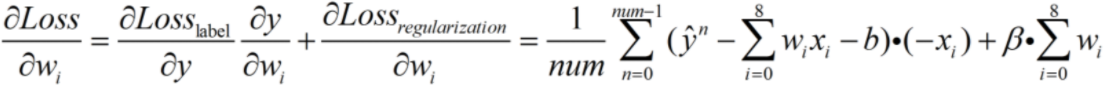


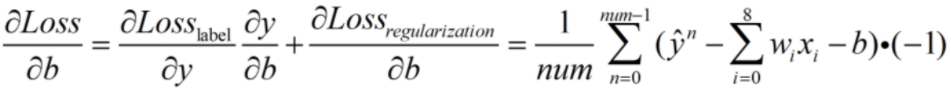


为正则项，为正则项系数。

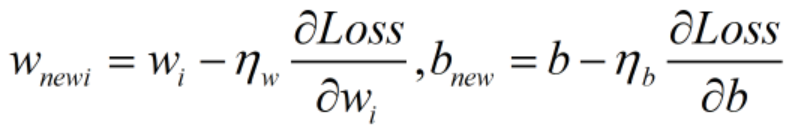
**3.梯度更新**

梯度计算：为使Loss最小，需要求Loss在w上的偏微分和Loss在b上的偏微分。





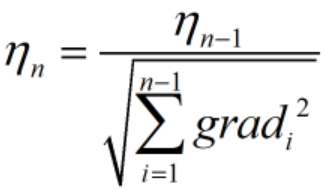
计算出梯度后，通过梯度下降法实现参数更新。

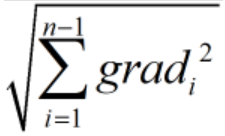


为权重w更新时的学习率，为偏置b更新时的学习率。

**4.学习率更新**

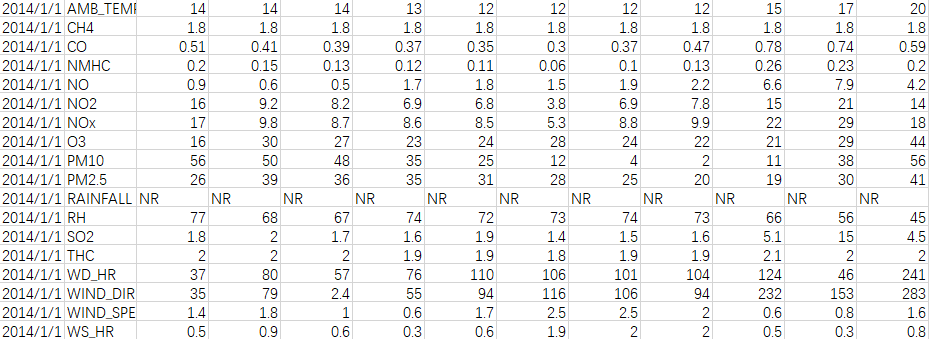
采用adagrad算法来更新学习率。在不影响模型效果的前提下提高学习速度，对学习率进行实时更新，即让学习率的值在学习初期较大，之后逐渐减小。



为更新后的学习率，为更新前的学习率，为在此之前所有梯度平方和的二次根。

四、系统设计

**1.数据处理**

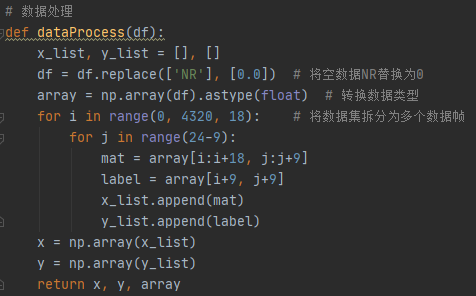


部分数据截图

数据中存在空数据NR，RAINFALL表示当天对应时间点是否降雨，有降雨值为1，无降雨值为NR，则将空数据NR替换为0。

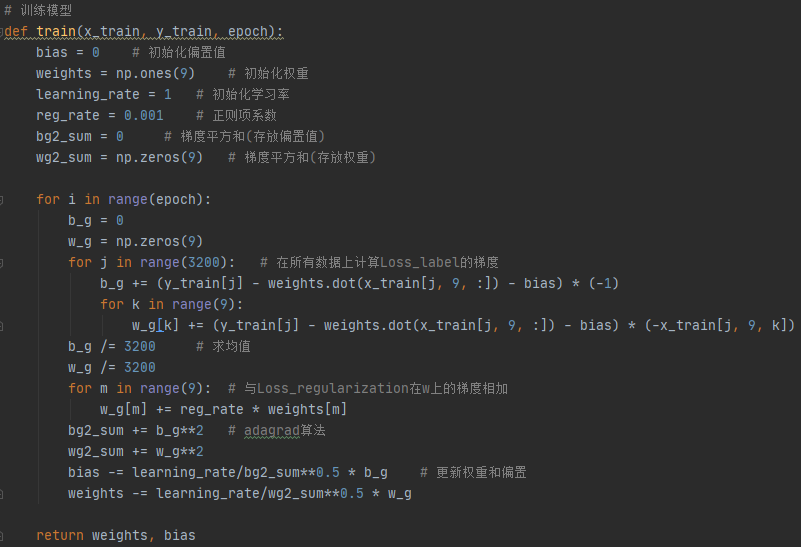
每一天包含的信息维度为(18,24)(18项指标，24个时间节点)。将0到8时的数据截取出来，形成一个维度为(18,9)的数据帧，作为训练数据，将9时的PM2.5含量取出来，作为该训练数据对应的label；同理可取1到9时的数据作为训练用的数据帧，10时的PM2.5含量作为label。以此分割，可将每天的信息分割为15个shape为(18,9)的数据帧和15个label。

训练集中共包含240天的数据，共可获得3600个数据帧和对应的label。



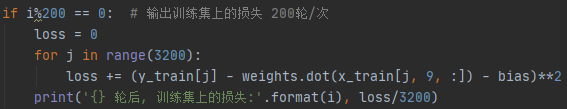
**2.训练模型**

参照上文模型方法中所述：

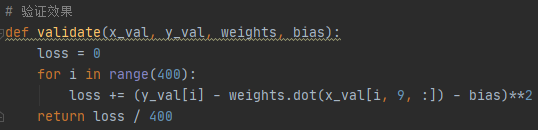


**3.验证效果**

在训练过程中，每训练200轮输出一次在训练集上的损失：

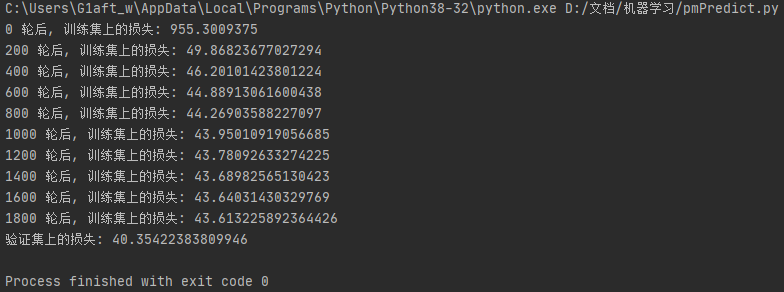


结束后，输出在验证集上的损失：



五．实验结果分析、对比和讨论

**1.实验结果**



模型在验证集上的损失为40左右，即预测值与label之间的平均差异在6到7之间，模型的整体效果较差。

**2.改进思路**

(1)更换模型，不使用线性回归模型，从数据中发现规律并以此为依据建立模型。

(2)数据集中除PM2.5的值外还有许多可能会影响或与PM2.5有关联的数据，在构建模型时，可充分考虑PM2.5与其他大气成分之间的关系，构建更合理的模型。

(3)分割训练集和验证集时，可以按照比例随机抽取数据帧作为训练集和验证集。

六．对本门课的感想、意见和建议

建议在本门课程前开设python语言相关课程，由于对python比较不熟悉，在进行实验的过程中遇到了很多问题并需要花费大量时间解决，在本门课程结束后我对此语言的使用更加熟练了。

个人感觉本门课程与之前所学专业课难度跨度较大，仅通过上课所学无法理解掌握，需要自己在课余时间查阅资料并进行实践。

随着大数据时代的发展和AI领域的延伸，我认为开设本门课程很有必要，可以让我们了解专业的前沿发展并培养相关领域的兴趣。