PM2.5 prediction report

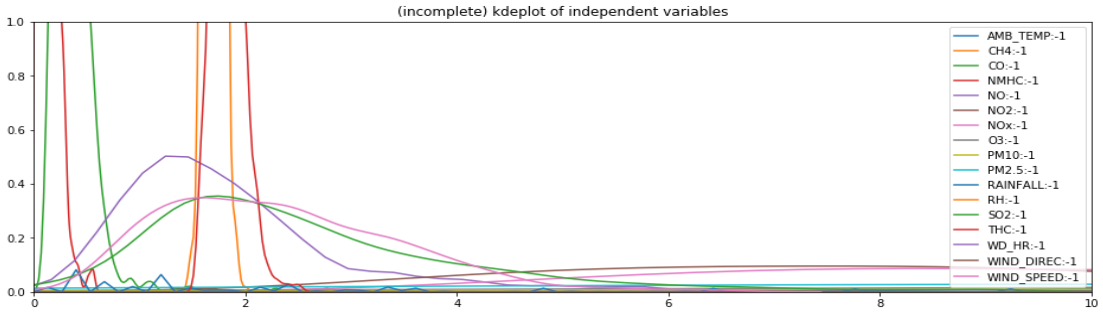
何数学

（一） 数据处理

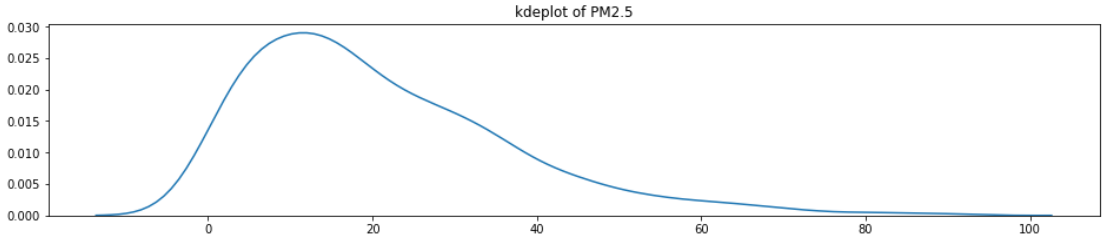
1. 经过检查，缺失值仅存在于 RAINFALL 字段，且 RAINFALL 字段有大量缺失。由实际经验，降雨量对空气质量应有显著影响，并猜想这里的缺失值较可能是因当时未下雨、所以未记录，因此对 RAINFALL 字段不采取删除处理，而是用 0 填充。
2. 对所有的数字转为 float 格式。
3. 检测的空气指标变量共 18 个。认为各变量的时序影响有限，不超过 8 小时。因此对于每日 24 小时，每八小时的数据作为一个样本，可认为样本之间是独立的。

   因变量为：每日第 8 次观测、第 16 次观测、第 24 次观测的 PM 2.5 的值（对应的 column 名为 7、15、23）。

   自变量为：所有 18 个变量在每日第 8、16、24 次观测前的 4 个时间点的观测值（对应的 column 名为['3', '4', '5', '6', '11', '12', '13', '14', '19', '20', '21', '22']），相当于 18*4=72 个自变量。
4. 挑选每一日的某 3 个时刻中所有空气指标的观测值画出核密度估计图，无法展示完全，截取部分如下所示：



   作为因变量的 PM2.5 的核密度估计图如下：



   可以看出因变量与许多自变量都是偏态的正态分布，且分布范围差异较大，因此对自变量的数据列表采取了去均值、除以标准差的标准化处理。
5. train.csv 分为训练集与测试集，比例为 3:1。

（二）求解回归模型

由于自变量之间具有时序性，即时序上连续的自变量存在强相关，导致X^TX存在若干接近于 0 的特征值，损害其可逆性，最终使参数估计值不稳定。因此采用岭回归，使得X^TX特征值远离 0，也即加入 L2 正则化，此时 loss function 为（其中$h_\theta(x)$为预测值）：

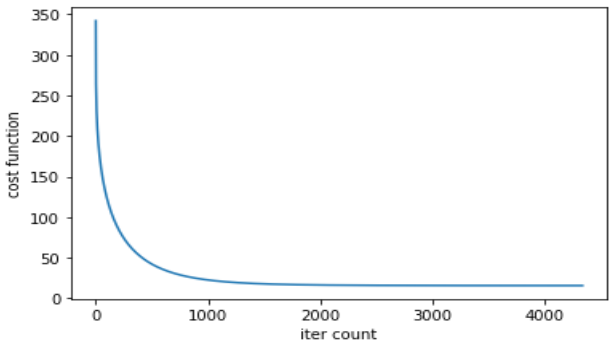
$$J(\theta) = \frac{1}{2n}\sum_{i=1}^{n}\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)^2 + \frac{\lambda}{2}\sum_{j=1}^{m}\theta_j^2$$

使用 Adagrad 求解，初始学习率设定为 0.4，终止迭代标准为：1. 某次迭代前后的两个参数向量的平方差小于阈值，即认为收敛；2. 迭代次数超过最大限定次数。

对于不同的正则系数λ分别求解模型，最终结果如下：

lambda	training_loss	testing_loss	testing_R2	testing_RMSE
0.0	14.9113	20.0830	0.8675	6.3377
1.0	14.9294	19.6714	0.8703	6.2724
5.0	15.1445	18.8256	0.8758	6.1361
10.0	15.5667	18.6706	0.8769	6.1107
20.0	16.5931	19.2879	0.8728	6.2109
30.0	17.7090	20.2771	0.8663	6.3682
40.0	18.8776	21.4039	0.8588	6.5428
50.0	20.0905	22.6098	0.8509	6.7246
100.0	26.6787	29.3422	0.8065	7.6606
200.0	41.2364	44.4829	0.7066	9.4322

基于上述结果，选择正则系数$\lambda = 10$，此时在测试集上的 R^2 达到最大 0.8769，RMSE 达到最小 6.1107。对于正则系数$\lambda = 10$，损失函数走势图如下（不包括正则项，只考虑预测值与实际值的均方差）：



具体回归模型的系数见代码，已与调用 sklearn.linear_model 中的岭回归模型 Ridge 进行了比对，系数大致相同。

（三）预测 Test.csv

将 RAINFALL 字段的缺失值填充为 0，选取每日所有空气指标变量的末尾 4 列观测值进行同 train.csv 一样的标准化，使用$\lambda = 10$时对应的回归模型进行预测。