华东师范大学计算机科学技术系实验报告

课程名称： 数据分析实践	**年级：** 2017 级	**实践作业成绩：**
指导教师： 兰曼	**姓名：** 英嘉豪	**作业提交日期：**
实验编号： 2	**学号：** 10175102247	**实践作业编号：** 2

1 实验名称

台风路径预测分析

2 实验目的

运用多种线性回归策略提高台风路径预测的准确性。

实验要求

台风训练数据 typhoon.dat，共计 500 条，共 18 列，前 16 列为预报因子，最后 2 列分别是经纬度。

台风测试数据为 typhoon200Test.dat，共计 200 条，每条有 16 列预报因子，与训练数据的前 16 列数据格式顺序相同，但是没有最后的两列经纬度数据。

测试的任务就是对这 200 条台风数据预测每条记录的经度和纬度值。

1. Lon.V1: 起报时刻经度 lon+00
2. Lon.V2: 起报时刻前 12 小时纬向移速 vlat-12
3. Lon.V3: 起报时刻前 24 小时纬向移速 vlat-24（可能是经向）
4. Lon.V4: 起报时刻前 24 小时所在经度 lon-24

5. Lon.V5: 起报时刻前 12 小时至前 24 小时纬向移速 vlat-24
6. Lon.V6: 起报时刻与前 12 小时的经度差 (lon+00)-(lon-12)
7. Lon.V7: 起报时刻前 6 小时所在经度 lon-06
8. Lon.V8: 起报时刻前 18 小时所在经度 lon-18
9. Lat.V9: 起报时刻纬度 lat+00
10. Lat.V10: 起报时刻前 12 小时经向移速 vlon-12
11. Lat.V11: 起报时刻前 24 小时经向移速的平方 (vlon-24)$\hat{2}$
12. Lat.V12: 起报时刻前 12 小时所在纬度 lat-12
13. Lat.V13: 起报时刻与前 24 小时的经度差 (lon+00)-(lon-24)
14. Lat.V14: 起报时刻前 6 小时所在纬度 lat-06
15. Lat.V15: 起报时刻前 18 小时所在纬度 lat-18
16. Lat.V16: 起报时刻前 6 小时地面附近最大风速 wind-06
17. Lon.t: 要预报的 24 小时后的经度（预报量）lon+24
18. Lat.t: 要预报的 24 小时后的纬度（预报量）lat+24

3　实验内容

3.1　环境准备

3.1.1　基本环境

- Windows 10
- Python 3.6.7

3.1.2 依赖

- pandas 0.25.3
- scikit-learn 0.21.3
- scipy 1.4.1
- numpy 1.17.1
- matplotlib 3.1.2

3.2 训练集选择，以及模型的验证

3.2.1 数据读入

```
# 加载数据
def loadDataSet(fileName):
    return pd.read_csv(fileName, header=None, encoding='utf-8', sep = '\t')
```

我这里将数据直接读入为DATAFrame格式，方便以后做切片。训练数据一共有 500 条，为了进行准确率的验证，将抽取 300 条作为实际的训练数据，而剩下的 200 条作为验证数据集。为了使得抽取的数据具有随机性，我将先将rawdata进行一步随机打乱后再分割得到测试数据和训练数据。

```
import random
rawdata = loadDataSet('./data/typhoon.dat').as_matrix()
random.shuffle(rawdata)
rawtraindata = rawdata[:300]
rawtestdata = rawdata[300:]
```

在开始实验前，为了验证我们的模型是否是具有”线性的“我先手写动实现标准线性回归来测试下模型的线性程度，函数主体部分如下：

```python
def standRegres(xArr,yArr):  # 标准线性回归函数
    xMat = np.mat(xArr); yMat = np.mat(yArr).T  # list 转换成 mat
    xTx = xMat.T * xMat
    if np.linalg.det(xTx) == 0.0:  # 矩阵的逆可能并不存在，要在代码中对此作出判断
        print("This matrix is singular, cannot do inverse")
        return
    ws = xTx.I * (xMat.T * yMat)
    return ws
```

3.2.2　模型的验证

为了验证相关性,我随机选取了两个维度Lon.V1和Lat.V9做出与标签值Lon.t和Lat.t的分布图, 以及利用standRegres预测的标签值的分布图分别如图 1，图 2 所示，其中蓝色点列是原始数据，红色点列是对应的预测结果。可以得出其数据在所展示的维度上具有较好的线性相关性。

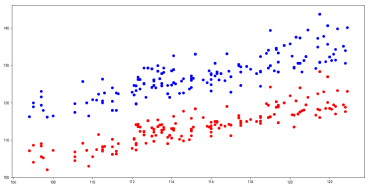

图 1: lon

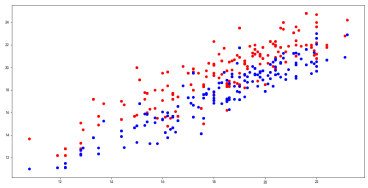

图 2: lat

由于数据是多维度的我有选取了Lon.V2和Lat.V2分别画出 Lon.t 和 Lat.t 的分布图，可以看出明显的线性相关性。

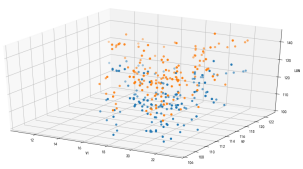

图 3: lon

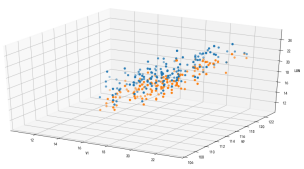

图 4: lat

3.3 误差函数

```
# 计算误差
def absError(yArr, yHatArr): # yArr and yHatArr both need to be array
    return abs(yArr - yHatArr).sum() / len(yArr)
```

函数接受标准坐标和预测坐标作为输入，计算差值绝对值的和，并除以样本个数，得到平均误差。

```
# 计算经度预测值与真实值的相关系数
corrcoef(lonHatArr.T, lonArr)
```

计算预测值与真实值的相关系数。

```
# 计算距离公式: 110 * sqrt(lon^2 + lat^2)
110 * sqrt(absErrorLon**2 + absErrorLat**2)
```

运用距离公式来计算经纬度坐标的误差在地面实际距离上的误差。

3.4 模型选择

我们调用 sklearn 的线性回归包括：标准线性回归、岭回归、弹性网络等。

3.4.1 线性回归

```
linear_model.LinearRegression() # 线性回归
```

3.4.2　岭回归

岭回归是一种专用于共线性数据分析的有偏估计回归方法，实质上是一种改良的最小二乘估计法，通过放弃最小二乘法的无偏性，以损失部分信息、降低精度为代价获得回归系数更为符合实际、更可靠的回归方法，对病态数据的拟合要强于最小二乘法。

```
sklearn.linear_model.Ridge(alpha=1.0, fit_intercept=True, normalize=False)
```

- alpha：两项之间的权重
- fit_intercept：默认为 true，数据可以拦截，没有中心化
- normalize：输入的样本特征归一化，默认 false

3.4.3　LASSO 回归

LASSO 回归的特点是在拟合广义线性模型的同时进行变量筛选（variable selection）和复杂度调整（regularization）。因此，不论目标因变量（dependent/response varaible）是连续的（continuous），还是二元或者多元离散的（discrete），都可以用 LASSO 回归建模然后预测。这里的变量筛选是指不把所有的变量都放入模型中进行拟合，而是有选择的把变量放入模型从而得到更好的性能参数。

```
sklearn.linear_model.clf = linear_model.Lasso(alpha = key)
```

3.4.4　弹性网络

ElasticNet 是一种使用 L1 和 L2 先验作为正则化矩阵的线性回归模型，这种组合用于只有很少的权重非零的稀疏模型，比如 :class:Lasso，但是又能保持 :class:Ridge 的正则化属性。

```
1 sklearn.linear_model.ElasticNet(alpha = 1.0)
```

3.4.5 多种回归方法的组合

在这个问题中，需要预测的是经度和纬度。存在一种可能：对经度和纬度分别运用不同的线性回归模型进行预测，会得到更高的准确率。

4 实验结果及分析

4.1 岭回归

4.1.1 参数调整过程

```
for i in range(1,11):
    key = i/10;
    clf = linear_model.Ridge (alpha = key) # 岭回归
    s = s = u' 岭回归"0027+str(i/10)
    # 构建模型
    clf.fit (trainX, lonTrainY) # 训练经度模型
    lonyHat = clf.predict(testX) # 使用经度模型去预测经度
    clf.fit (trainX, latTrainY) # 训练纬度模型
    latyHat = clf.predict(testX) # 使用纬度模型去预测纬度
```

调整 alpha 和训练数据量的大小，得到误差结果的变化关系。

4.1.2　结果分析

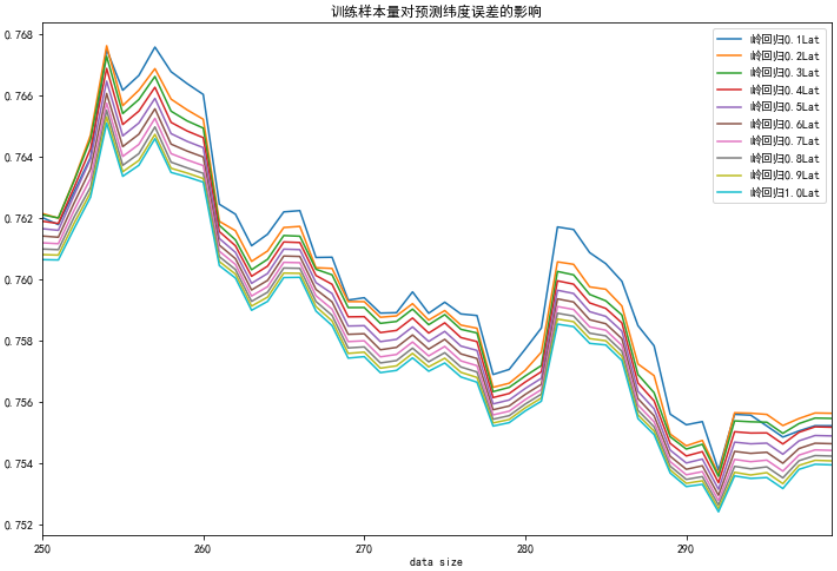

图 5: 岭回归训练样本量对预测纬度误差的影响

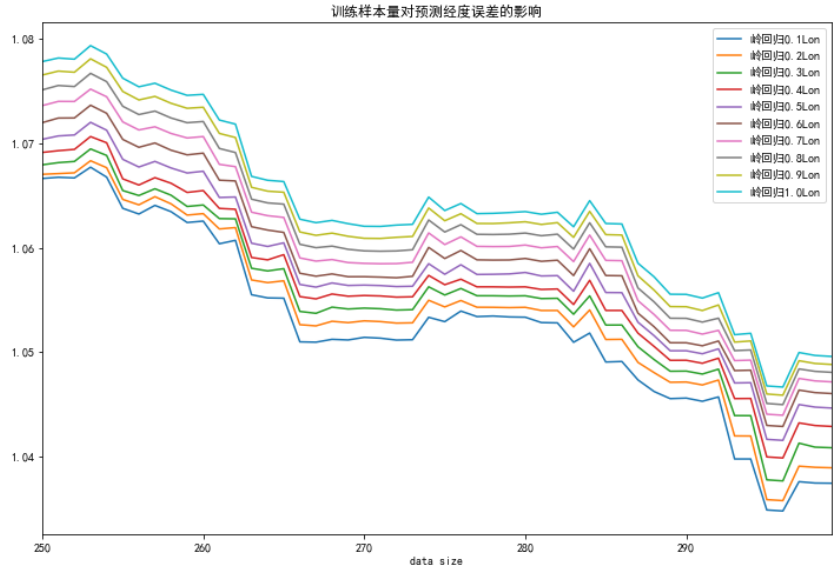

图 6: 岭回归训练样本量对预测经度误差的影响

我们可以观察到随着训练数据集中数据数量的增多，无论是经度预测误差还是维度预测误差都是是逐渐减小的，同样其对距离的误差也是逐渐减小的。并且在图中没有看到明显的收敛趋势，这是因为 500 个训练数据的数据量偏小。

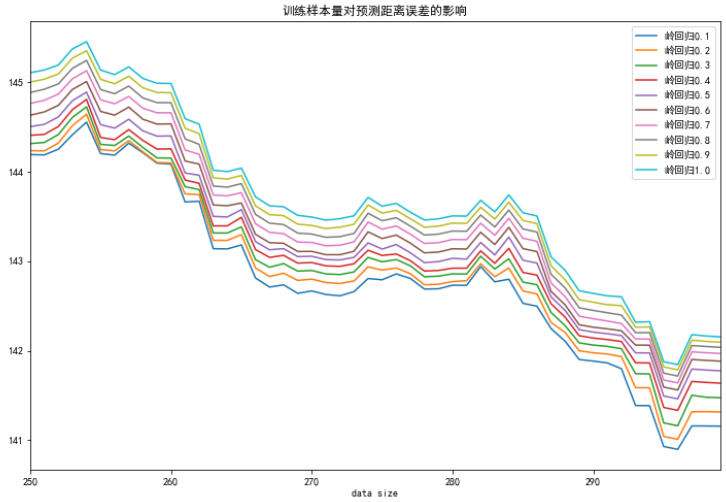

图 7: 岭回归训练样本量对预测距离误差的影响

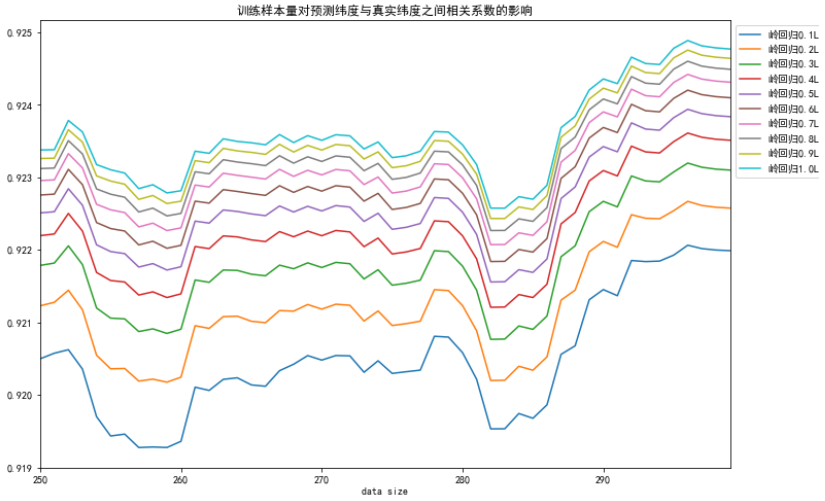

图 8: 岭回归训练样本量对预测纬度与真实纬度之间相关系数的影响

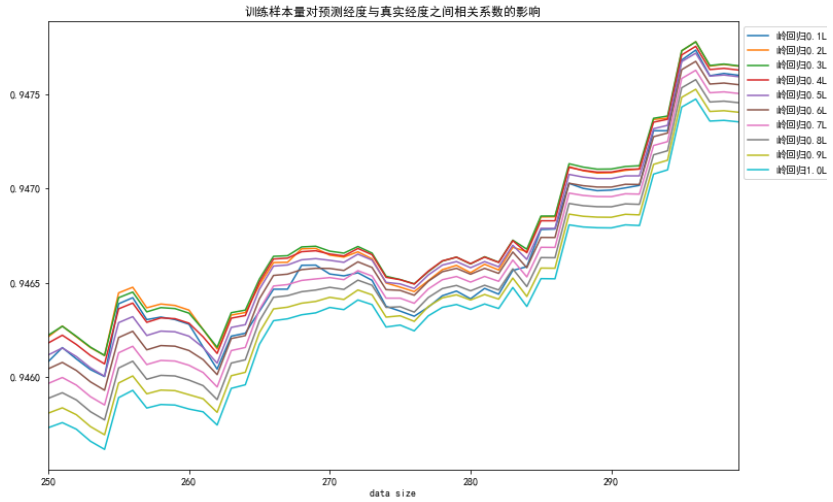

图 9: 岭回归训练样本量对预测经度与真实经度之间相关系数的影响

由图 9 可知lamda=0.3时，对经度预测表现较好。由图 8 可知当lamda=0.1时，对纬度预测表现较好。应该将lamda=0.1/0.3的岭回归预测同时加入最后的结果比较。

4.2　Lasso 回归

4.2.1　参数调整过程

```
for i in range(1,10):
    key = i/20;
    clf = linear_model.Lasso(alpha = key)
    s = s = u'Lasso 回归'+str(i/20)
    # 构建模型
    clf.fit (trainX, lonTrainY) # 训练经度模型
    lonyHat = clf.predict(testX) # 使用经度模型去预测经度
    clf.fit (trainX, latTrainY) # 训练纬度模型
    latyHat = clf.predict(testX) # 使用纬度模型去预测纬度
```

4.2.2 结果分析

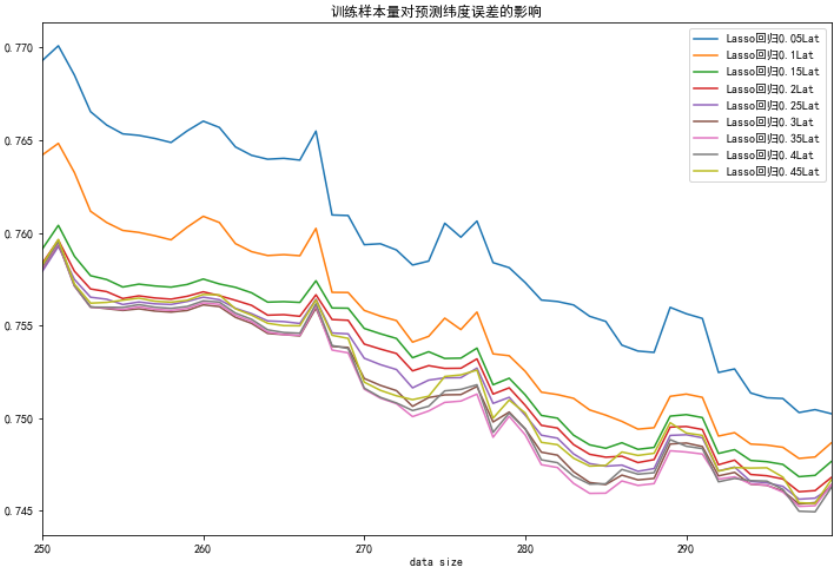

图 10: Lasso 回归训练样本量对预测纬度误差的影响

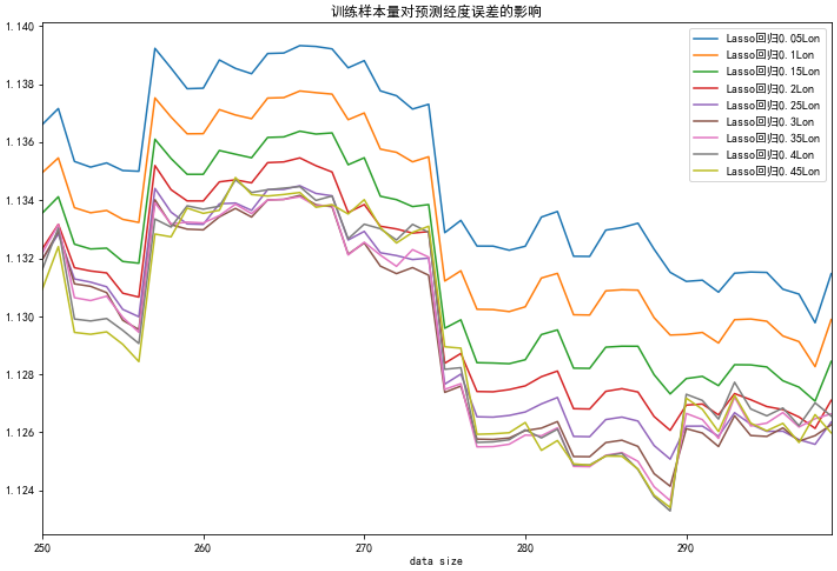

图 11: Lasso 回归训练样本量对预测经度误差的影响

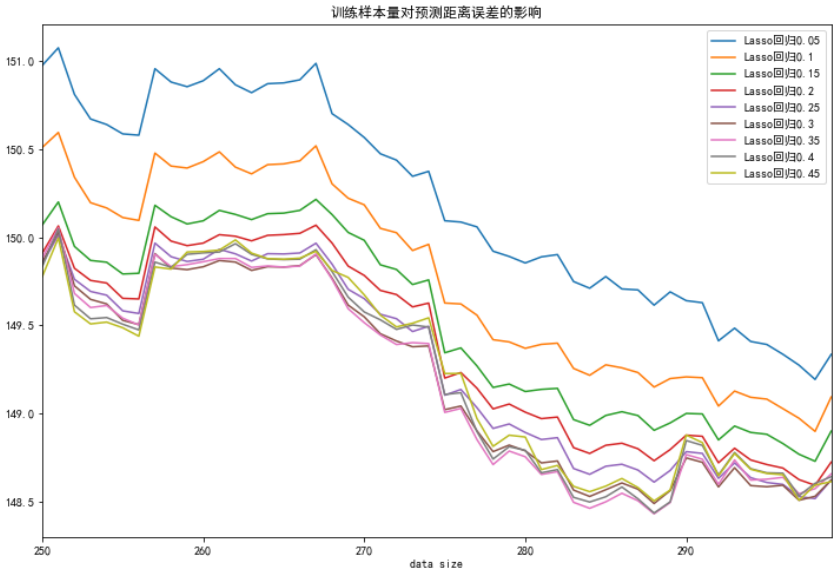

图 12: Lasso 回归训练样本量对预测距离误差的影响

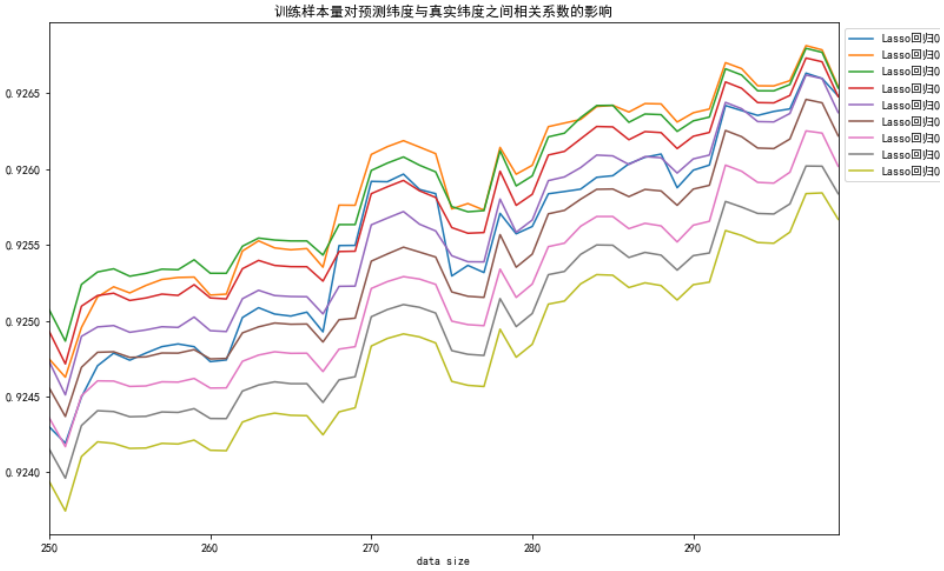

图 13: Lasso 回归训练样本量对预测纬度与真实纬度之间相关系数的影响

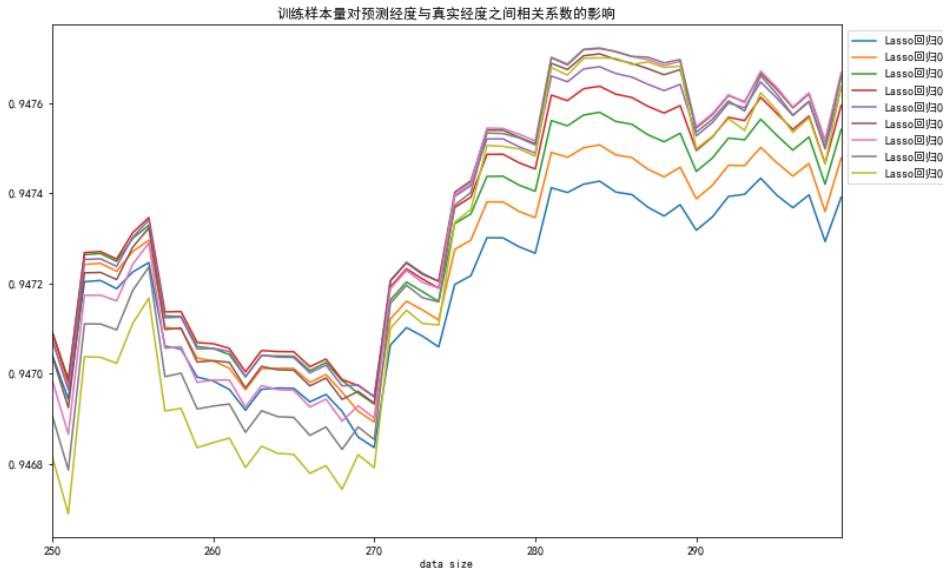

图 14: Lasso 回归训练样本量对预测经度与真实经度之间相关系数的影响

随着训练数据集中数据数量的增多，预测误差是逐渐减小的；由图易得，预测误差均随着 lamda 的值减小而减小。

4.3 ElasticNet 回归

4.3.1 参数调整过程

```
for i in range(1,20):
    key = i/20;
    s = s = u'ElasticNet 回归'+str(i/20)
    # 构建模型
    clf = linear_model.ElasticNet(alpha = key,max_iter=1000) # 岭回归
    clf.fit (trainX, lonTrainY) # 训练经度模型
    lonyHat = clf.predict(testX) # 使用经度模型去预测经度
    clf.fit (trainX, latTrainY) # 训练纬度模型
    latyHat = clf.predict(testX) # 使用纬度模型去预测纬度
```

4.3.2　结果分析

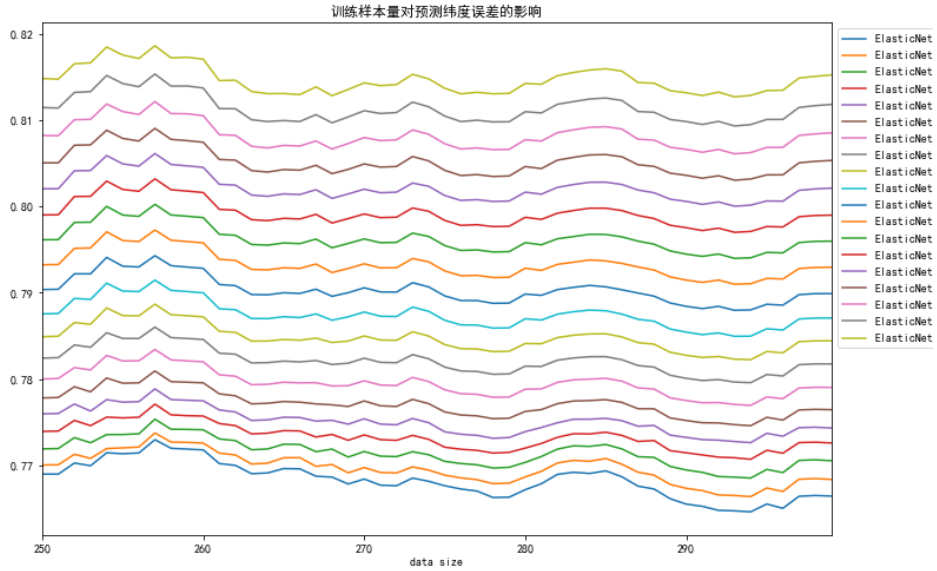

图 15: ElasticNet 回归训练样本量对预测纬度误差的影响

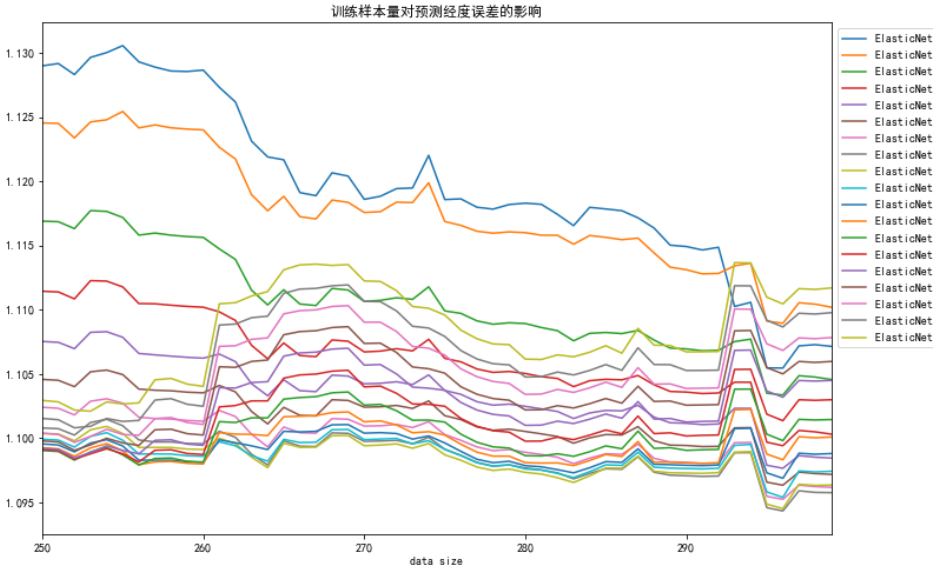

图 16: ElasticNet 回归训练样本量对预测经度误差的影响

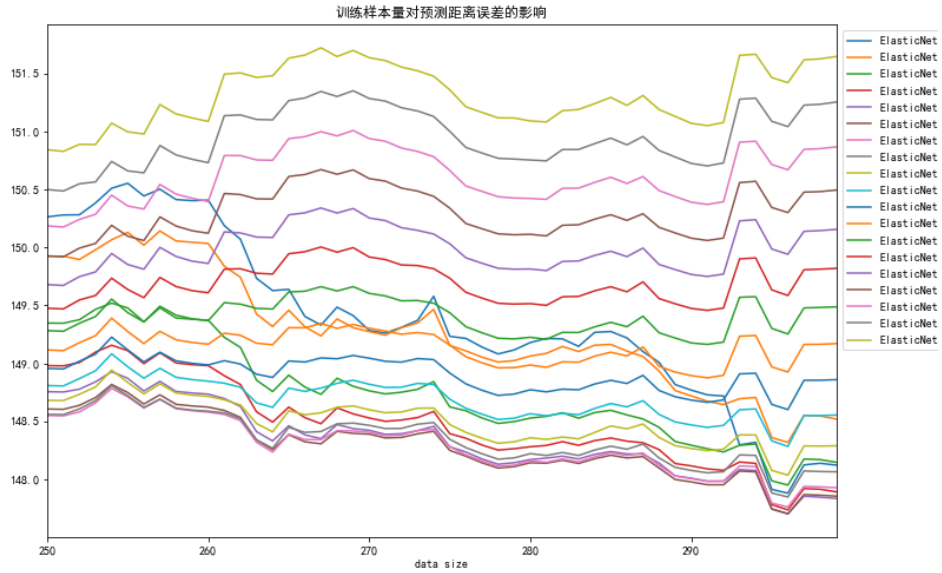

图 17: ElasticNet 回归训练样本量对预测距离误差的影响

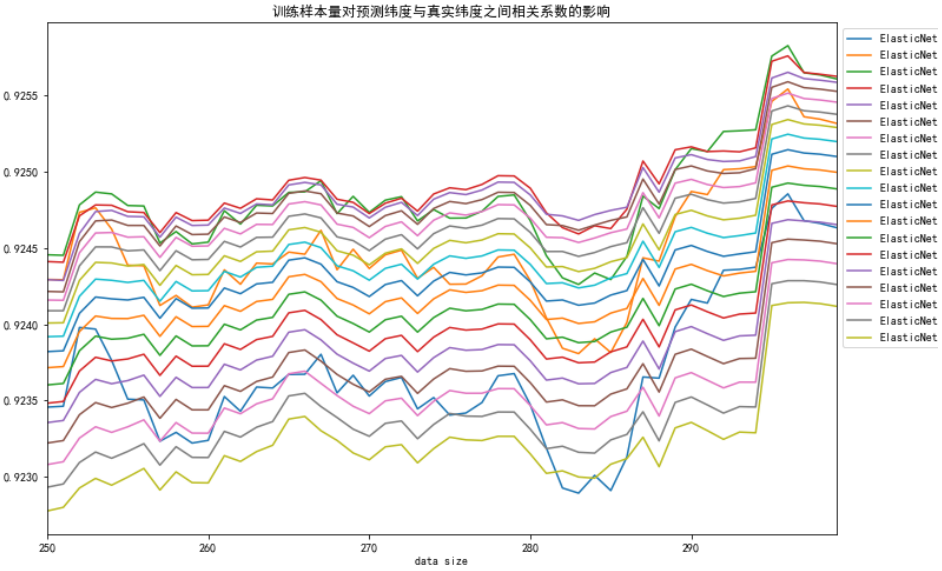

图 18: ElasticNet 回归训练样本量对预测纬度与真实纬度之间相关系数的影响

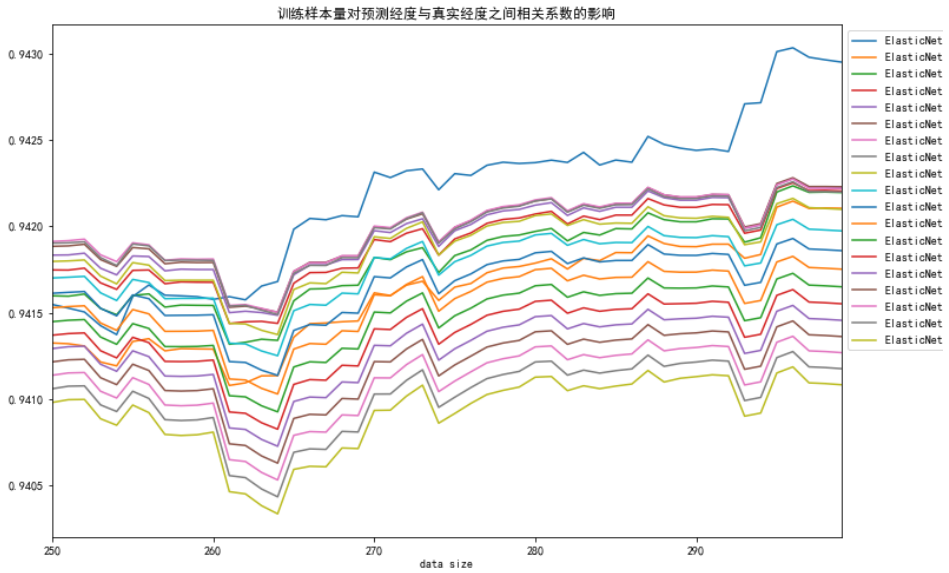

图 19: ElasticNet 回归训练样本量对预测经度与真实经度之间相关系数的影响

随着训练数据集中数据数量的增多，预测误差是逐渐减小的；由图易得，ElasticNet 结果与 Lasso 相同，预测误差均随着 lamda 的值减小而减小。

4.4　模型比较

我们根据上面的实验结果讲各个模型中参数最优的模型作为最后比较的选择，其中 ElasticNet 回 alapha 选择为 0.05，Lasso 回归选择为 0.01，岭回归分别选择了 0.1 和 0.3 作为参数取值。

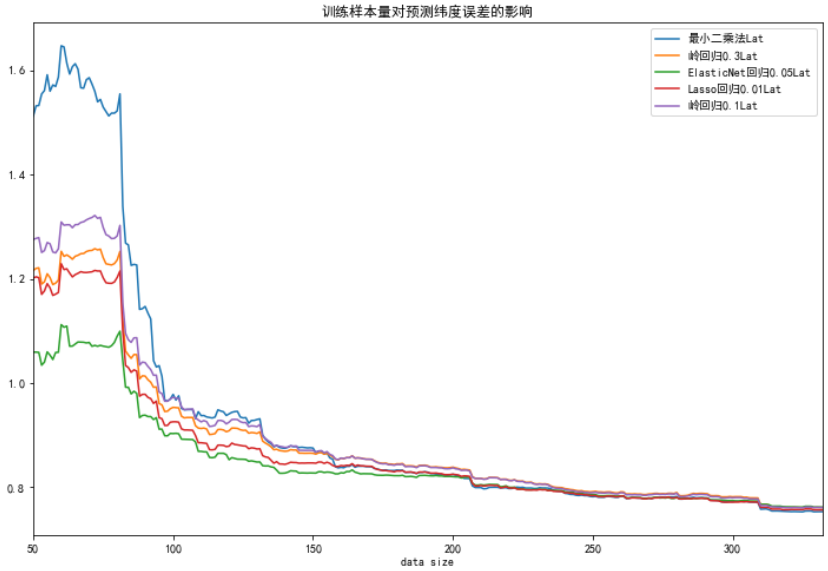

图 20: total 训练样本量对预测纬度误差的影响

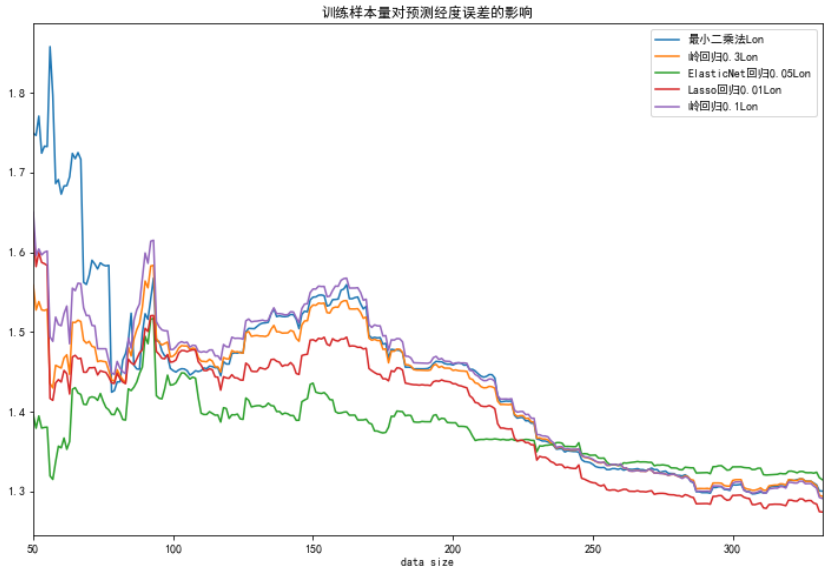

图 21: total 训练样本量对预测经度误差的影响

图 22: total 训练样本量对预测距离误差的影响

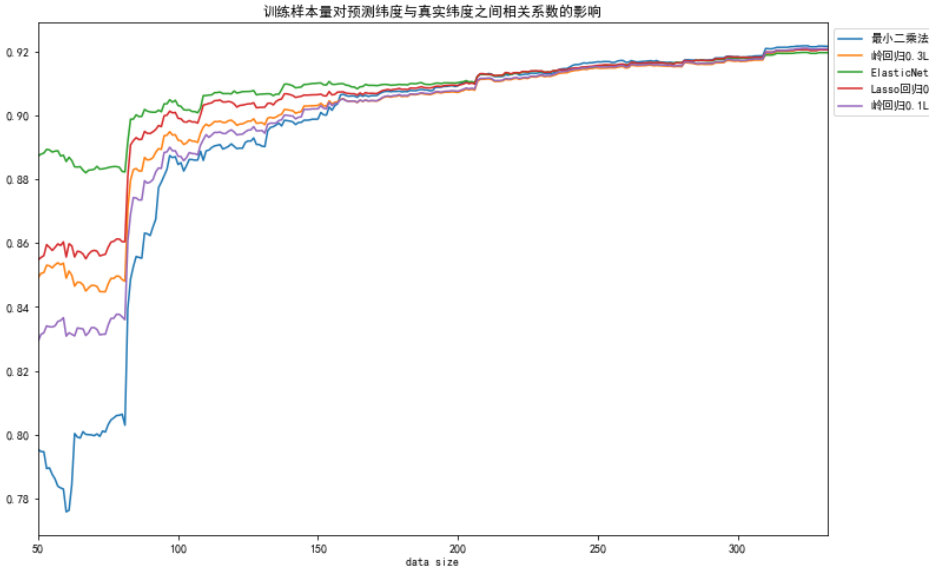

图 23: total 训练样本量对预测纬度与真实纬度之间相关系数的影响

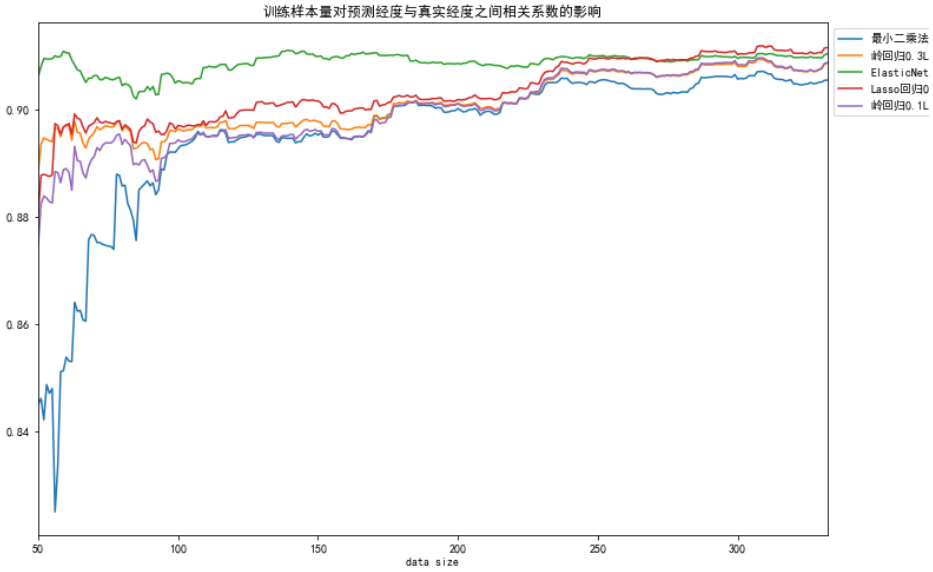

图 24: total 训练样本量对预测经度与真实经度之间相关系数的影响

随着训练数据集中数据数量的增多，预测误差是逐渐减小的；关于经度的预测，

Lasso(0.01) 取得较好结果，预测误差最小，预测经度与真实经度相关系数最大；关于纬度的预测，最小二乘法取得较好结果，预测误差最小，预测纬度与真实纬度相关系数最大。故关于距离的预测，选择 Lasso 回归算法进行经度预测，最小二乘法进行纬度预测。

5 问题讨论

- 在后面进行模型分析因为数据高维度的关系并没有挑选维度进行数据可视化，只有在前面进行模型初步分析的时候进行了可视化。
- 感觉实际中台风的模型不一定是线性模型，就算是的话预测得到的结果应该也只是作为趋势的分析。
- 调用 Sklearn 的 SGDRegressor 回归模型相较于其他模型得到的误差极大，故排除此模型。

6 结论

- 随着训练数据集中数据数量的增多，预测误差是逐渐减小的.
- 关于经度的预测，Lasso(0.01) 取得较好结果，预测误差最小，预测经度与真实经度相关系数最大。
- 关于纬度的预测，最小二乘法取得较好结果，预测误差最小，预测纬度与真实纬度相关系数最大。