姓名：夏一凡
学院：信息科学与工程学院
学号：320190942230
日期：2020.4.20

摘要：本次实验主要包含两大部分：第一部分为生成正弦函数的数据集按 8：2 随机划分为数据集和测试集，并使用一元多项式回归模型训练数据集确定回归系数，使用测试集测试，MAE 和 RMSE 评估；第二部分为 uci 下载一个适合回归分析的数据集 D2，标准化后进行 ridge 回归确定回归系数，从正则化路径中获取超参数 λ。划分测试集和测试集，重新确定回归系数获取 MAE 和 RMSE 并重复五次。

1. 引言

   多项式回归：研究一个因变量与一个或多个自变量间多项式的回归分析方法，称为多项式回归（Polynomial Regression）。如果自变量只有一个时，称为一元多项式回归；如果自变量有多个时，称为多元多项式回归。在一元回归分析中，如果依变量 y 与自变量 x 的关系为非线性的，但是又找不到适当的函数曲线来拟合，则可以采用一元多项式回归。岭回归是对不适定问题（ill-posed problem)进行回归分析时最经常使用的一种正则化方法。这两种算法在数据挖掘中都为较常见算法，被广泛应用，因此这项作业具有重要意义

2. 算法：

数据集的处理：下载数据集 Abalone，共'Sex', 'Length', 'Diameter', 'Height', 'Whole weight', 'Shucked weight', 'Viscera weight', 'Shell weight', 'Rings'几个类别标签，其中数值型数据超过三列，样本超过 100，无缺失值，适合作回归.
进行了多种标准化处理，不包含性别数据
Rings 代表年龄

$$f(w)=\sum_{i=1}^{m}(y_i-x_i^Tw)^2+\lambda\sum_{i=1}^{n}w_i^2$$

$$\hat{\mathbf{w}}^{Ridge}=(X^{\mathrm{T}}X+\lambda I)^{-1}X^{\mathrm{T}}\mathbf{y}$$

MAE 和 RMSE 计算公式

$$RMSE(X,h)=\sqrt{\frac{1}{m}\sum_{i=1}^{m}(h(x_i)-y_i)^2}$$

$$MAE(X,h)=\frac{1}{m}\sum_{i=1}^{m}|h(x_i)-y_i|$$

本次实验使用了以下算法：

一元多项式回归：将$x^n$看作一个一元变量，转化为多元回归分析问题

Ridge 回归

数据标准化：将数据按比例缩放，使之落入[-1,1]。目的是去除数据的单位限制，将其转

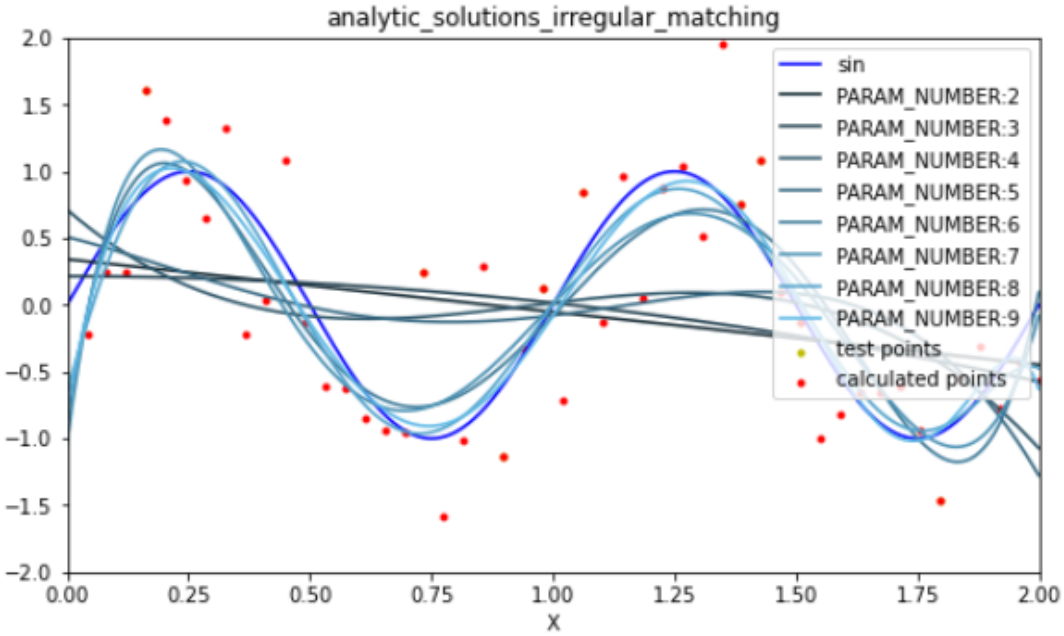
化为无量纲的纯数值，便于不同单位或量级的指标能够进行比较和加权。
MAE（平均绝对误差）：Σ|真实值-预测值|
RMSE（均方根误差）：用于数据更好的描述，MSE 开根号求得。
数据集随机划分

3.实验结果及分析：
实验一数据集产生时，先随机生成 50 个点作为 x，y = sin（x）加一个取值在 0~1 中间的随机数作为噪声。使用公式进行
实验一拟合过程中，发现并不是维度越高 MAE 和 RMSE 效果越好，以$X^6$和$X^7$为最高项时结果最好，（$X^6$：MAE: 0.6915544169532408，RMSE: 0.8381880037924861，$X^7$：MAE: 0.6915544209820349，RMSE: 0.8381880153650814）（见图一），拟合图像如图二所示，实际上当维度足够高，且无干扰项时，一元多项回归可以看作函数（sin x）的泰勒公式展开，但干扰项产生噪声。试图增加正则项，MAE 和 RMSE 不变。但由于样本数据太少，每次计算的差距都太大，没有较大参考价值
实验二：uci 下载 abalone.data 数据集，自定义标准化函数，并将性别属性使用 one-hot 编码处理后标准化，得到数据集 D2.csv。自定义 Ridge 函数，画出岭迹图，通过图像得到 λ 在 0.00015 附件（图三），运用库函数 sklearn，计算 λ，最佳 λ = 0.0001448118227674533。（图四），较符合。

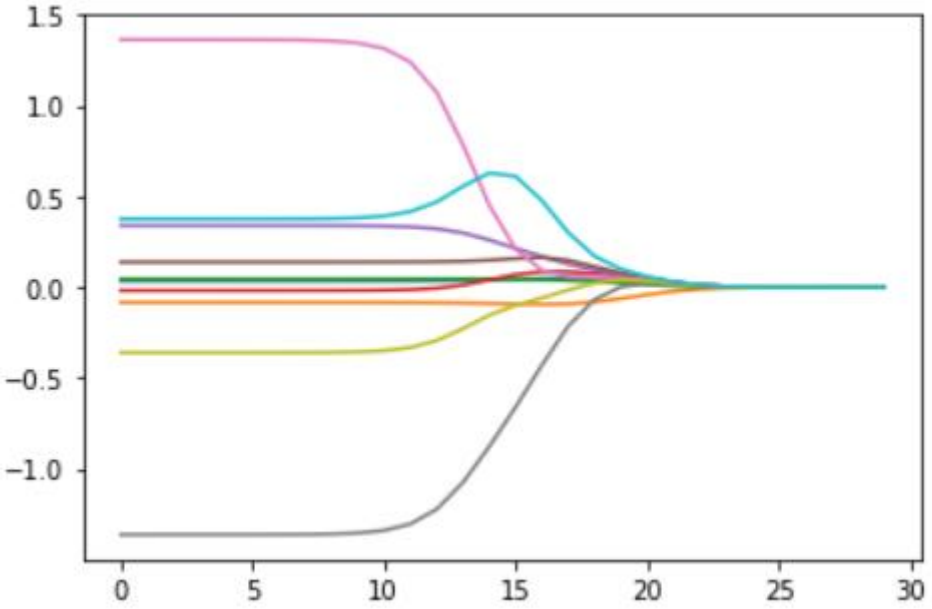
```
analytic_solutions_irregular_matching
MAE: 0.7328124645591136
RMSE: 0.9550018753775852

-0.395 x + 0.3426
MAE: 0.6931027745102577
RMSE: 0.8258991664697287
          2
-0.1952 x - 0.00458 x + 0.2151
MAE: 0.693102774510246
RMSE: 0.8258991664697576
          3          2
-1.326 x + 3.782 x - 3.154 x + 0.7134
MAE: 0.7718498888092616
RMSE: 1.0246525708782415
          4          3          2
-1.013 x + 2.728 x - 1.396 x - 0.9064 x + 0.5091
MAE: 0.7718498888190666
RMSE: 1.024652570896822
         5         4         3         2
13.34 x - 67.7 x + 120.7 x - 88.6 x + 23.04 x - 0.8628
MAE: 0.6915544169532408
RMSE: 0.8381880037924861
         6         5         4         3         2
-3.51 x + 34.4 x - 115.4 x + 171.1 x - 113.2 x + 27.68 x - 1.04
MAE: 0.6915544209820349
RMSE: 0.8381880153650814
         7         6         5         4         3         2
-16.5 x + 112 x - 284.4 x + 324 x - 143.4 x - 3.722 x + 12.71 x - 0.6469
MAE: 0.7760375314598107
RMSE: 1.0326341015681217
          8         7         6         5         4         3         2
-14.14 x + 96.6 x - 256.6 x + 343.6 x - 272.8 x + 167.9 x - 84.96 x + 21.08 x - 0.8008
analytic_solutions_regular_matching with PARAM_NUMBER:9
```

(图一)




（图二）



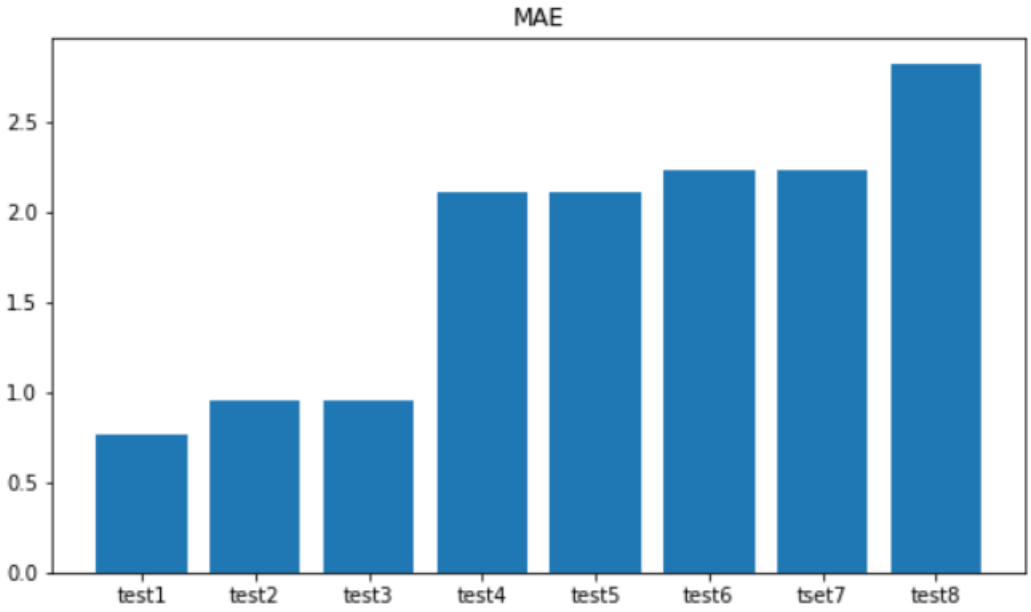
（图三）

```
best lambda 0.0001448118227674533
```

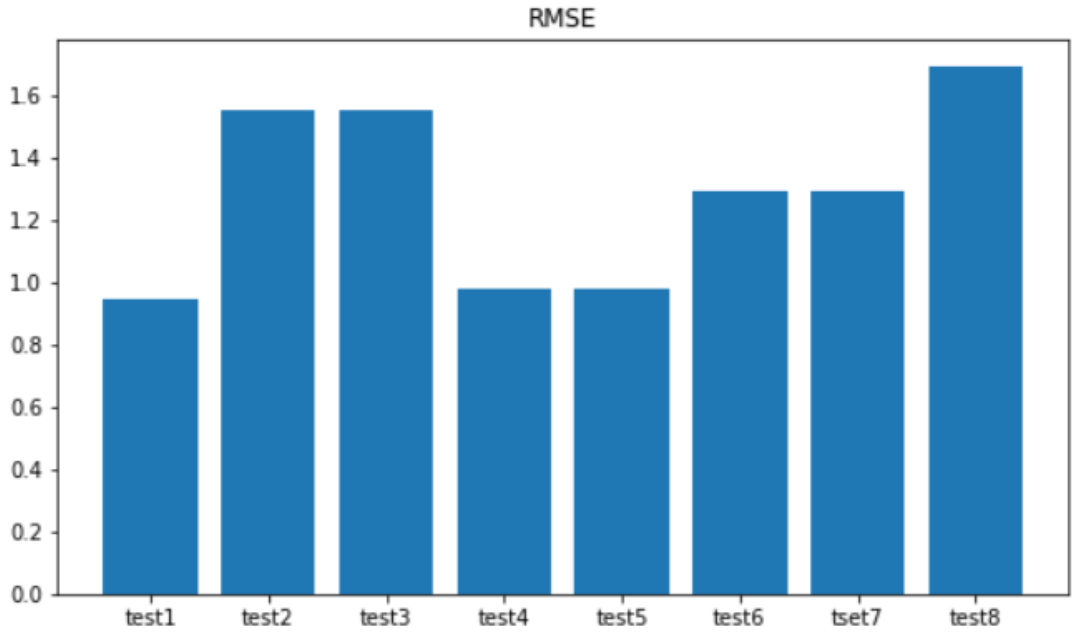
（图四）

```
MAE: 0.9359262263598562
RMSE: 1.231663679035447
MAE: 0.9655850908427875
RMSE: 1.2568868314804962
MAE: 0.9687117087043419
RMSE: 1.2533178055409484
MAE: 0.9279242279606396
RMSE: 1.2149989260288154
MAE: 0.9445847511447805
RMSE: 1.2523546404505275
```

（图五）




（图六）




（图七）

4.结论

随机扰动：在做算法时，有时数据量太少，训练得到的结果不准确，此时可以对数据进行随机扰动进行扩容，即对数据进行一个范围的上下浮动，以增加数据量，来提高算法的鲁棒性。使用数据集预测鲍鱼年龄，数据标准化处理比较重要，此外手动实现 ridge regression。对于有些矩阵，矩阵中某个元素的一个很小的变动，会引起最后计算结果误差很大，这种矩阵称为“病态矩阵”。有些时候不正确的计算方法也会使一个正常的矩阵在运算中表现出病态。岭回归(英文名：ridge regression, Tikhonov regularization)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。通常岭回归方程的 R 平方值会稍低于普通回归分析，但回归系数的显著性往往明显高于普通回归，在存在共线性问题和病态数据偏多的研究中有较大的实用价值。

参考文献：

1.参考教材：Python 机器学习预测分析核心算法-[美]M·鲍尔-沙嬴＆李鹏(译)-人民邮电出版社-2017

2.Machine Learning in Python：Essential Techniques for Predictive Analysis（Python 机器学习：预测分析核心算法）, Michael Bowles, Wiley, 2015.pdf

文献参考：