吉林大学

本科生毕业论文（设计）

中文题目 半潜式钻井平台的运动数据分析软件的设计与实现

英文题目 Design and Implementation of Movement Data Analysis Software for Semi-submersible Drilling Platform

学生姓名 石家鑫 班级 3 学号 55140322

学　　院 软件学院

专　　业 软件工程

指导教师 姜宇 职称 讲师

半潜式钻井平台运动数据分析软件的设计与实现

摘要

随着时代的进步和互联网以及相关硬件，测量器材,方式的发展。人们积累了大量的数据，数据既是互联网发展的产物，也是驱动计算机科学和互联网向前继续发展的火种。近年来，随着人们意识到大量积累的数据的妙用以及硬件方面的不断发展以支持复杂的模型计算，在个性化推荐，预测运动轨迹，无人驾驶等领域，从占据各大新闻头条的 ALPHAGO 到你身边各种 APP 的推送喜好，人们对数据的运用越来越普及。同时，众所周知，我国海岸线长，海洋领土范围大，在许多地方蕴藏着丰富的海底油气资源。而半潜式钻井平台作为时代演进的产物，是人们进行海上油气勘探与开采的必要的现代化工具。作为一种新型的工具，对于半潜式钻井平台的数据分析由于数据的保密性以及其位于海洋学科和计算机学科的交叉位置导致其一直没有受到业界的重视和研究。但是从对以后半潜式钻井平台的自动化发展，以及数据分析学科的相应应用，对平台的相应运动数据加以利用，根据实际出发，解决实际问题具有一定的价值。

本文试图通过分析半潜式钻井平台的相关运动数据，试图找出数据之间的联系，通过对大量不同种类数据进行分析计算，建立可靠的回归模型，得出数据之间是否具有关联，以及其关联性的结论。尝试通过一部分数据预测平台可能产生的移动，从而达到自动平衡平台运动，维持平台稳定的效果。

关键字：数据分析 ，半潜式钻井平台 ，运动数据

Design and Implementation of Movement Data Analysis Software for Semi-submersible Drilling Platform

Author:Jiaxin Shi

Tutor:Yu Jiang

Abstract

With the progress of the times and the development of the Internet and related hardware, measuring equipment, and methods。People have accumulated a large amount of data. The data is not only the product of the development of the Internet, but also the kind of fire that drives the development of computer science and the Internet. In recent years, as people realize the magical use of accumulated data and the continuous development of hardware to support complex model calculations, in the areas of personalized recommendation, predicted motion trajectory, and driverlessness, etc., from the ALPHAGO that occupies major news headlines. To the push preferences of various APPs around you, the use of data is becoming more and more popular. At the same time, as we all know, China has a long coastline and a large marine territory. It has rich hydrocarbon resources in many places. As a new type of tool, the data analysis of semi-submersible drilling rigs has not received the attention and research of the industry due to the confidentiality of data and its location at the intersection of oceanography and computer science. But from the automation development of the semi-submersible drilling platform in the future, and the corresponding application of the data analysis discipline, the corresponding motion data of the platform will be used. According to the actual departure, solving the actual problem has certain value.

This paper attempts to analyze the relevant motion data of semi-submersible drilling rigs, trying to find out the connection between the data, by analyzing and calculating a large number of different types of data, and establishing a reliable regression model to determine whether there is correlation between the data and its Relevance conclusions. Try to predict the possible movement of the platform through a

part of the data so as to automatically balance the platform movement and maintain the stability of the platform.

Keywords: data analysis，Semi-submersible drilling platform，Sport data

吉林大学学士学位论文（设计）承诺书

本人郑重承诺：所呈交的学士学位毕业论文（设计），是本人在指导教师的指导下，独立进行实验、设计、调研等工作基础上取得的成果。除文中已经注明引用的内容外，本论文（设计）不包含任何其他个人或集体已经发表或撰写的作品成果。对本人实验或设计中做出重要贡献的个人或集体，均已在文中以明确的方式注明。本人完全意识到本承诺书的法律结果由本人承担。

学士学位论文（设计）作者签名：

2017 年 5 月 20 日

目　　录

第 1 章　绪论（黑体 3 号、居中）

1.1　研究内容

基于我国南海某处半潜式钻井平台的多组运动数据，对数据进行处理，分析，尝试以回归问题的思想寻找几组数据之间是否存在联系，并且通过数据训练回归模型，通过建立的模型对相关数据进行预测。尝试找出最适合数据之间的分析模型，优化模型参数，建立数据之间的关系。

1.2　研究意义

大数据隐含着巨大的社会、经济、科研价值,被誉为未来世界的“石油”,已成为企业界、科技界乃至政界 关注的热点,[15]在互联网，云计算，移动互联网等技术的推动下，数据规模正呈爆炸式增长。数据来源应用和结构特征繁多复杂。[14], 海洋平台是认识和开发海洋的一种极具潜力的支撑平台，是实现深海能源开发的最有效工具之一。但是海洋平台造价非常昂贵，以“海洋石油 981”半潜式深水钻井式平台为例，耗资 60 亿元人民币。因此，开展海洋平台安全性评估软件研究对平台的设计具有重要的指导意义，对我国经济建设和社会发展产生积极影响。[1]

由于海洋平台安全性评估预报对用户的专业水平要求高。与其他船舶性能预报一样，也急需将具体专业方向与相关商业、专业软件结合，并融合专家使用经验和知识，固化流程形成集成预报软件。目前国内比较有名的主要是北京索为公司的集成设计平台，以及上海奥蓝托软件技术有限公司的集成平台“Orient. CAE 集成定制平台”，前者以航空领域为主，而后者在船舶领域具有更多技术积累和更大的优势。在集成定制平台上，定制开发具有自主知识产权的评估软件系统，不但可以提高软件的易用性，推广软件的应用，提高我国海洋平台安全性评估水平，还可为打破国外技术垄断做出一定贡献。[2]

具体意义有：

1.合理利用半潜式钻井平台的大量数据，进行数据分析，尝试通过数据预测

钻井平台运动趋势，从而自动开启动力舵，保持钻井平台的平衡。

2.尝试找出多组数据之间的关联性。从而筛选出有用数据，为以后的处理研究提供方向。

1.3 研究现状与发展趋势

1.从数据分析技术的发展与应用来讲，随着硬件设备的不断升级，使在工程界应用大规模的机器学习模型来进行数据分析成为可能，近年来，国内外许多公司均将机器学习回归模型等技术应用于数据分析，例如短视频，新闻等推荐技术，通过建立用户模型来推荐给用户合适的选择。

2.从工程与相关技术实现的方面来讲，Google 等大公司与科研机构开发出了许多已于上手和使用的框架例如 Tensorflow，sklearn 等，足以支撑基础性的实验与数据处理与分析。其中封装了许多适用于数据分析与挖掘的模型，将其与科学计算相关的库结合。可以简便的建立模型完成相关数据分析与挖掘的任务。

3. 从研究的数据方向上，目前业界很少有将数据分析与挖掘技术应用到海洋尤其是钻井平台相关运动的数据上，不过随着时间的推移和技术的发展，必然有更多的工业类相关的项目与计算机技术相结合发展。

4.随着计算机硬件的不断发展以及大量数据的累计，通过建立数据分析系统为人们提供决策也是未来发展的潮流之一。

第 2 章　系统分析

2.1　研究目标

对相关钻井平台数据进行处理，包括从数据源读取数据，对数据进行预处理，对问题进行分类，选择相应的回归/分类问题模型，对选定的多个模型进行训练，熟悉模型的原理以及其公式和数学意义，通过交叉验证以及准确率评分以及自己对模型的了解不断优化模型参数，校验数据是否存在问题，对多组数据之间预计可能存在的关系尝试做出证明，以及对有关系的数据拟合最好的模型以及相应权重参数。

2.2　需求分析

本软件实际上实现的功能是将平台数据从数据源中读取，对其进行预处理，转换为机器学习适合的科学计算的矩阵，确定问题类别后，选定相关的语言，同时选定几个学习模型，将模型实现后，对数据进行训练，最后使用交叉验证以及准确率评分对模型参数不断进行调优，最后得出最优的模型，并且通过画图工具包，详细的展现预测模型的预测结果与实际结果的差距。对每个需求的细分如下：

2.2.1　数据读取与预处理

读取视数据源而定，预处理主要将数据类型统一，解决由于测量仪器问题造成的缺失值问题，同时进行标准化等处理。了解海洋钻井平台相关物理知识，尝试筛选出成组特征。

2.2.2　问题分类与模型选择

确定问题类型为回归，分类，聚类，降维，中的哪一部分，了解相关类型问题的解决方法，通过选择合适的机器学习模型来尝试对数据进行学习和之后的预测。

2.2.3　模型调优与验证

初步选定模型后，了解模型基本原理，通过使用数据例训练模型采取相应的验证法与搜索法寻优调参。

2.3 可能存在的问题与风险

对于以上需求，本次实验存在很多风险：

1. 数据均为原始数据，与许多其他机器学习算法工作环境不同的是这些数据均为传感器数值等原始记录，没有经过特殊处理，在预处理部分可能有一些比较大的问题。甚至由于离群值过多和数据受到其他因素影响较大无法得出结论。
2. 业界对钻井平台相关研究较少，可以参考的实验资料较少，以及对于钻井平台相关物理模型研究比较匮乏，很难从物理模型上选择合适的模型与超参数。
3. 本次实验数据为敏感数据，因此只能对较少的数据（几天）的数据做分析，对于海洋这种收到气候季节影响严重的实验环境来讲，得出的模型具体参数在某些情况下的拟合程度可能较差。

第 3 章　系统设计与实现

3.1　系统设计部分

3.1.1　开发语言及框架选择

语言选择上主要考虑了 python 和 java，鉴于前者在科学计算以及数据分析，机器学习方面有着强大的库，而且具有一些与 matlab 相类似的绘图库，所以选择 python 进行开发，在 python 的基础上，numpy，scipy，matplotlib，等工具包，提供了科学计算于绘图等功能，scikit-learn 框架则作为业界著名的开源数据分析，处理与挖掘的工具包，其中包含了回归，聚类，分类等模型的实现，有助于快速完成实验目标。

3.1.2　具体选择的软件包版本号

Python 2.7.10-release

Numpy 1.14.2

Matplotlib 2.2.2

Pip 9.0.3

Scikit-learn 0.19.1

Scipy 1.0.1

3.1.3　开发环境搭建

先到 https://www.python.org/ 下载相应的 python 库，安装，设置环境变量。之后可以通过命令行或者 pycharm 的仓库管理功能先更新作为 python 库管理工具的 pip，之后可以通过 pip 管理 python 相关依赖，前往 http://scikit-learn.org/等网站下载机器学习建模所需的软件包。用于科学计算与存储的数据结构的 numpy 和 scipy 可以通过 http://www.numpy.org/，https://www.scipy.org/，获取。用于绘图的 matplotlib 可以通过 https://matplotlib.org/获取。一切完成后，可以使用命令行输入 python，验证安装，同时可以使用 python 自带的解释器工具，书写一些简单 python 代码，进行练习。在本次实验中，我主要使用 pycharm IDE 进行代码编辑以及程序测试，pycharm IDE 是 jetbrains 公司开发的一款适用于 python 开发的 ide，学生可以通过

认证学生邮箱的方式获得免费使用权。

3.1.4 问题分类

机器学习主要分四大类，回归（regression），聚类（clustering），分类（classification），降维（dimensionality reduction）。

给定一个样本特征,我们希望预测其对应的属性值,如果是离散的，那么这就是一个分类问题，反之，如果是连续的实数，这就是一个回归问题。

如果给定一组样本特征,我们没有对应的属性值,而是想发掘这组样本在维空间的分布，比如分析哪些样本靠的更近，哪些样本之间离得很远，这就是属于聚类问题。如果我们想用维数更低的子空间来表示原来高维的特征空间，那么这就是降维问题。[7]

从本次实验的数据上来看，输入与输出仅为一段时间内的钻井平台姿态，及其位置以及附近的海洋状况等数据，均为连续型随机变量，我们要完成的任务是对数据进行分析挖掘，找出一组或几组数据之间的关系。所以可以以回归问题的思路来完成这次实验。

3.1.5 回归问题概述

回归问题中最经典的问题就是人口增长问题与房屋面积问题了，譬如：我们手头以一批数据，这些数据包含房屋的面积和对应面积的房价信息，如果我们能得到房屋面积与房屋价格间的关系，那么，给定一个房屋时，我们只要知道其面积，就能大致推测出其价格了。通常，这类预测问题可以用回归模型（regression）进行解决，回归模型定义了输入与输出的关系，输入即现有知识，而输出则为预测。

一个预测问题在回归模型下的解决步骤为：

1. 积累知识： 我们将储备的知识称之为训练集 Training Set，通过度训练集的学习，我们使机器从中获取知识。
2. 学习：学习如何预测，得到输入与输出的关系。在学习阶段，应当有合适的指导方针。在这里，合适的指导方针我们称之为学习算法 Learning Algorithm。
3. 预测：学习完成后，当接受了新的数据（输入）后，我们就能通过学习阶段获得的对应关系来预测输出。

对于学习过程我们有这两点要求：

1. 有手段能评估我们的学习正确性。
2. 当学习效果不佳时，有手段能纠正我们的学习策略。

有了解决回归问题的思路，我们开始着手解决问题，分类问题其实与回归问题很像，分类需要将回归的输出离散化，因此有些分类模型事实上对于回归问题也同样适用。

3.1.6 模型初步选择

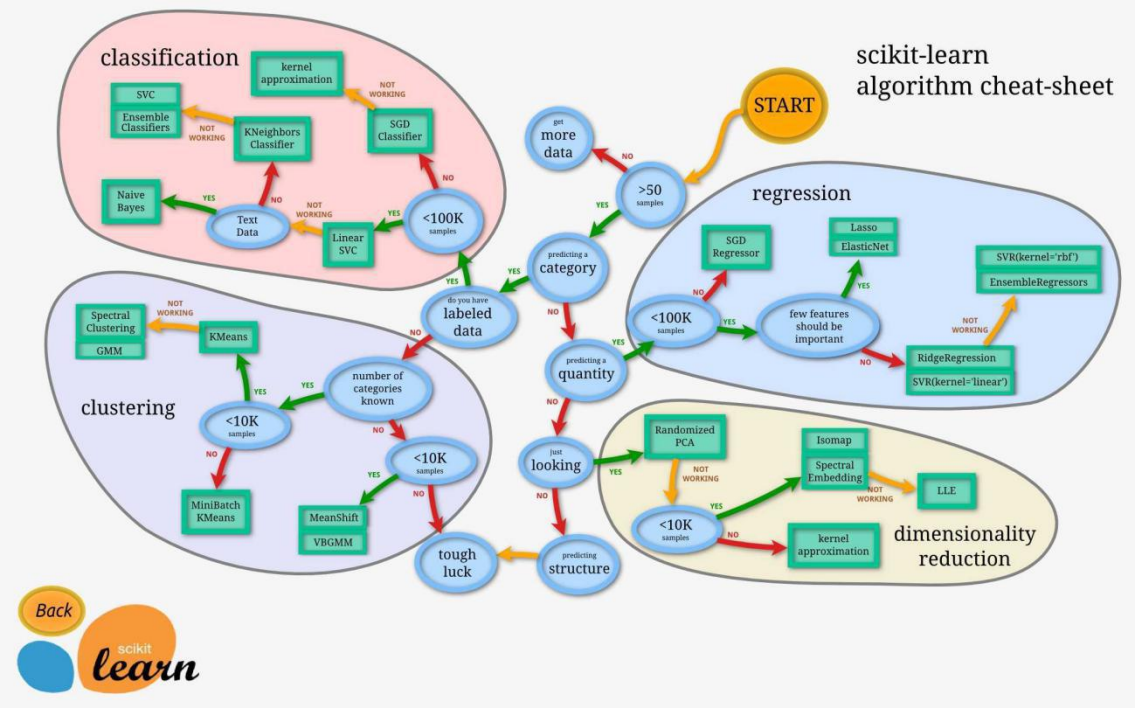


图 3-1

如上图所示，sklearn 官方文档中的一些资料，已经为我们将模型按照问题类型以及适配的数据量分类。根据官方推荐我们基本上选择了普通最小二乘法的线性回归模型，岭回归，还有适合做非线性回归的神经网络模型以及支持向量机。

3.1.7 结果输出方式

本次实验对于每个模型的评价基本打算采取两种方式输出结果，第一种是使用 Python 提供的 matplotlib 包绘图，matplotlib 提供了许多方法函数，用以实现与 matlab 相似的功能，以直观的形式体现出预测值和实际值在结果上的差距，第二种是通过 sklearn 提供的模型选择与验证部分的方法，其中提供了 K 折交叉验证，留一验证等验证方式，以及均方误差，绝对误差，可决系数等方式，以数字的形式表现出来实际至与预测值的距离（差值）。

3.2 系统实现部分

3.2.1 读取与储存数据，缺失值插补与数据筛选转换

本次实验相关数据均以文本文档的形式储存，涉及了钻井平台位置，姿态，倾角等等多种数据。数据可以直接通过文件流的方式读入，使用 numpy 的 array 以及 martix 数据结构，将其保存为浮点数数组和矩阵，以便用于科学计算。在处理的过程中，发现由于部分检测仪器的问题，导致数据处理的时候出现报空的错误，对于缺失值插补，根据经验，我们可以采用同组数据的中位数，平均数或者众数进行插补，sklearn 软件包为我们提供了这些方法，可以直接提取相同特征，二维数组中的一列或者矩阵的一列的中位数，平均数或者众数代替这一列中的某个特定数值，一般为被标注为 null 的空值，来完善我们的数据矩阵，在实现系统是我选择先使用 0 字符串来代替使用平均值来补全数据。

3.2.2 数据标准化

使数据集标准化使大多数机器学习算法的基本要求，通常为了特征数据能够在模型在其中学习的时候获得比较好的结果，我们将其标准化为具有零均值和标准方差的符合正态分布的特征数据，这部分可以通过先去均值来实现中心化，再除以非常量特征的标准差进行缩放来实现，sklearn 中负责预处理的 preprocessing 类其中的 scale 方法可以直接实现这一点，同时 StandardScaler 可以从训练集中学习相应的缩放数值，将其应用于测试集。

图 3-2

如上图所示，经过标准化后的数据具有接近于 0 的均值和标准方差，在数据服从标准正态分布之后，对于多特征的数据，所有数据对于机器学习算法的影响变得相同，不会导致算法从一个方差极大的特征中学习，导致其他特征的影响被缩小。

对于一些具有其他特点的数据我们可能还需要一些其他方法，例如：

1. 归一化：归一化　是缩放单个样本以具有单位范数的过程。如果计划使用二次形式(如点积或任何其他核函数)来量化任何样本间的相似度，那么此过程将非常有用。Sklearn 中的 normalize 方法提供了一个快速简单的方法在类似数组的数据集上执行操作，使用 l1 或 l2 范式。该方法接受的主要参数是 X，即需要转换的数组。以及 norm，一个正则化参数。
2. 二值化：二值化是将特征过滤得到布尔类型变量的过程，在而分类问题中使用很多，例如处理一张黑白图片的时候。虽然在回归问题中可能遇到的比较少，但是通过合适的标签转换，二值化的处理方法可以发挥作用。在 skLearn 中的 binarizer 类实现了一个简单的二值化操作，其主要接受一个在 0 和 1 之间的浮点数，作为阈值，按照阈值二值化数据。

3.2.3 普通最小二乘法的线性回归模型

最小二乘法通过最小化误差的平方和寻找数据的最佳函数匹配，在统计学上把数据点与它在回归直线上的差异叫做残差，把每个残差平方之后加起来称为残差平方和，它表示随机误差的效应。[8]一组数据的残差平方和越小,其拟合程度越好。 最小二乘法通过最小化误差的平方和寻找数据的最佳函数匹配，是一种常用曲线拟合法，[9]因此不光是基本的线性回归模型，还有许多回归模型，他们的损失函数都是用标准差定义的

最小二乘法的数学证明如下：

考虑超定方程组：

$$\sum_{j=1}^{\mathrm{n}} X_{ij}\beta_j = y_i\left(i = 1,2,3...m\right)$$

其中 m 代表有 m 个等式，n 代表有 n 个未知数 β，m>n,将其向量化之后为:

$$X\beta = \mathrm{y}$$

$$X=\begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1n} \\ X_{21} & X_{22} & \cdots & X_{2n} \\ \vdots & \vdots & & \vdots \\ X_{m1} & X_{m2} & \cdots & X_{mn} \end{bmatrix} \qquad \beta=\begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{bmatrix} \qquad y=\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix}$$

图 5-1

显然该方程组一般没有解，为了选取最合适的 β 让该等式"尽量成立"，引入残差平方和函数 S：

$$\mathrm{S}(\beta) = \| X\beta - y \|^2$$

当 $\beta = \hat{\beta}$ 时， $S(\beta)$ 取最小值，记作

$$\hat{\beta} = \mathrm{argmin}(\mathrm{S}(\beta))$$

通过对 $S(\beta)$ 微分求最值，可以得到：

$$X^{\mathrm{T}} X\hat{\beta} = X^T y$$

当矩阵 $X^T X$ 非奇异时，则有唯一解，因为当矩阵为奇异矩阵的时候即行列式为 0 的时候，这个方程是无解或者说有无穷解的，

最小二乘法需要解决的就是如上等式的β取值问题，β在多元线性回归中即为一个 1*n 的矩阵，每一个元素对应一个变量的最佳参数，最终使残差平方和最小。

基于最小二乘法的普通线性回归模型优点在于简单，高效，易于上手和理解，缺点在于从模型本身来讲，无法直接对非线性数据直接进行分析，对于复杂数据的分析能力欠缺，缺乏足够多的调优手段。

在 sklearn 中的普通最小二乘法的线性回归模型是一组用于回归的方法，其中目标值 y 是输入变量 x 的线性组合。 在数学概念中，如果 $\hat{y}$ 是预测值。数学公式表达式如下：

$$\hat{y}(w, x) = w_0 + w_1 x_1 + ... + w_p x_p$$

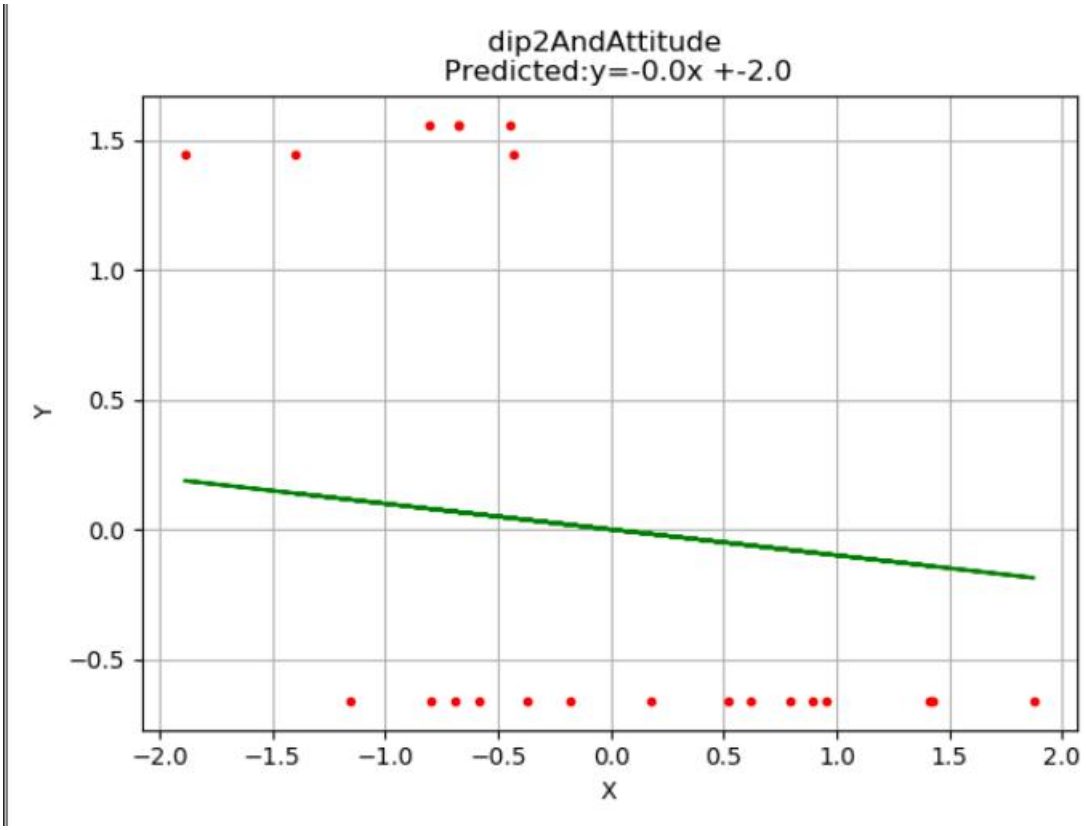
其 LinearRegression 包可以拟合一个带有系数的 $\omega = (\omega_1, ..., \omega_p)$ 的线性模型，同时可以使得数据集实际观测数据与预测数据之间的残差平方和最小，该方法将线性模型的系数储存在成员变量 coef_中。

sklearn 中的普通线性回归模型有一些可选参数：

fit_intercept:布尔值，默认为 True。是否计算此模型的截距，如果设置为 False，则计算中不会使用截距（例如，数据预期已居中）。

normalize:布尔值，默认为 False。fit_intercept 设置为 False 时忽略此参数。如果为真，则回归前的回归系数 X 将通过减去平均值并除以 l2-范数而归一化。

n_jobs:整型，默认为 1，用于计算的核心数量，-1 表示使用全部的 cpu。



在极少数据时，线性模型的 5 折回归模型的R^2评分为[0.43314185 0.76532869 0. -0.64147964 0.09878785]。matplotlib 模型图示如上，当将训练数据增加到一天的

数据，大概为1500+组数据时，图示如下（红点为实际数据点，绿线为预测数据）：

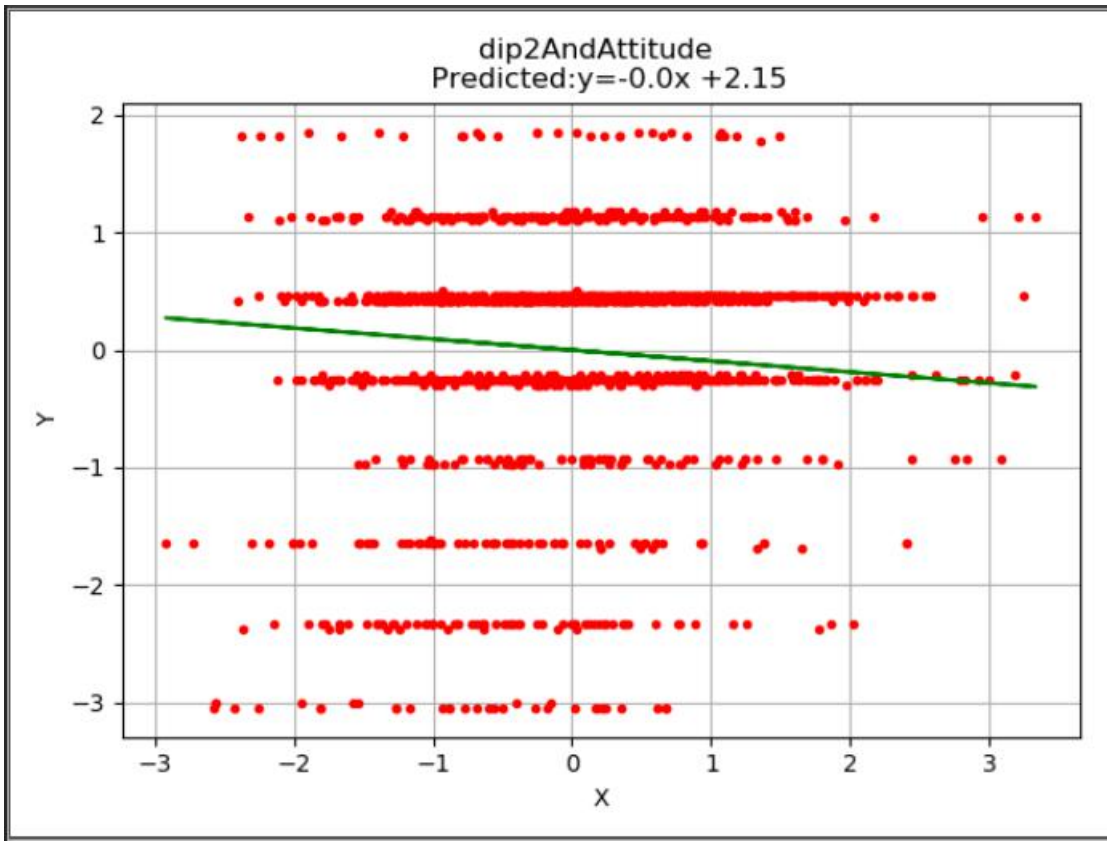


图 3-3

3.2.4 多项式回归：用线性模型训练非线性数据

使用线性模型来训练数据的非线性函数，既可以保持了一般快速的线性方法的性能，同时允许它们适应更广泛的数据范围。可以通过构造系数的 polynomial features 来扩展一个简单的线性回归。在标准线性回归的情况下，你可能有一个类似于二维数据的模型:

$$\hat{y}(w, x) = w_0 + w_1 x_1 + \ldots + w_p x_p$$

当我们创建一组新的自变量：

$$z = [x_1, x_2, x_1 x_2, x_1^2, x_2^2]$$

同时，我们使用 z 代替到第一个公式中的 x 的时候，我们就可以使用一个线性模型来训练具有非线性函数的数据了。在 sklearn 中，通过使 PolynomialFeatures 相关的方法，可以通过指定参数为幂数，可以得到一组数据的多项式展开。从而实现多项式回归。

使用多项式回归有介个需要注意的地方：

1. 能够模拟非线性可分的数据;线性回归不能做到这一点。它总体上更灵活，可以模拟一些相当复杂的关系。

2. 完全控制要素变量的建模（要设置变量的指数）。

3. 需要仔细的设计。需要一些数据的先验知识才能选择最佳指数。

4. 如果指数选择不当，容易过拟合。

3.2.5 岭回归

岭回归在最小二乘法的基础上做了一些改进，通过增加一个正则化罚项，对回归系数加以惩罚来解决最小二乘法可能出现的一些问题，如果多组特征之间的关系具有一个类似于线性的关系，那么可能会造成数据矩阵趋向于奇异矩阵，导致模型对误差非常敏感，通过增加一个正则化罚项，可以使岭系数变成一个代发想的残差平方和。

高共线性的存在可以通过几种不同的方式来确定：

1. 尽管从理论上讲，该变量应该与 Y 高度相关，但回归系数并不显著。

2. 添加或删除 X 特征变量时，回归系数会发生显着变化。

3. X 特征变量具有较高的成对相关性（pairwise correlations）（检查相关矩阵）。

$$\min_{w} ||Xw - y||_2^2 + \alpha||w||_2^2$$

其中， $\alpha >= 0$ 是控制系数收缩量的复杂性参数：α 的值越大，收缩量越大，这样系数对共线性的鲁棒性也更强。

通常正则化化参数具有两种形式即 L1 范数$\alpha \| w \|_1$与 L2 范数$\alpha \| w \|_2^2$，使用了 L1 范数的模型叫做 Lasso 回归，使用了 L2 范数的就是 Ridge 回归。两者的不同点是 L1 正则化是指权值向量w中各个元素的绝对值之和，通常表示为$\| w \|_1$。L1 正则化可以产生一个稀疏权值矩阵，用于特征选择。L2 正则化是指权值向量中各个元素的平方和然后再求平方根（可以看到 Ridge 回归的 L2 正则化项有平方符号），通常表示为$\| w \|_2^2$，通常用于防止过拟合。

岭回归在使用上有几个需要注意的地方：

1. 这种回归的假设与最小平方回归相同，不同点在于最小平方回归的时候，我们假设数据的误差服从高斯分布使用的是极大似然估计（MLE），在岭回归的时候，由于添加了偏差因子，即 w 的先验信息，使用的是极大后验估计（MAP）来得到最终参数的。

2. 它缩小了系数的值，但没有达到零，这表明没有特征选择功能。

在 sklearn 中的 linear_model.Ridge 包中实现了岭回归模型。其中提供了许多参数供我们调整，比较重要的有：

Alpha:浮点型，可以是数组，对应有多个特征的数据。代表着正则化的强度。必须是正数，正则化改善了问题的条件并降低了估计的方差，较大的值指定较强的正则化。 Alpha 对应于其他线性模型（如 LogisticRegression 或 LinearSVC）中的 C ^ -1。如果因变量参数是一个数组，既具有多个特征，那么罚项使针对每个特征的，他们必须在数量上相同。

Solver：求解器，可选值{'auto'，'svd'，'cholesky'，'lsqr'，'sparse_cg'，'sag'，'saga'}。

'auto'根据数据类型自动选择求解器。

'svd'使用 X 的奇异值分解来计算岭系数。对于奇异矩阵比“cholesky”更稳定。

'cholesky'使用标准的 scipy.linalg.solve 函数来获得解决方案。

'sparse_cg'使用 scipy.sparse.linalg.cg 中的共轭梯度解算器。作为一种迭代算法，这种求解器对于大规模数据（可能性设置 tol 和 max_iter）比'cholesky'更合适。

'lsqr'使用专用的正则化最小二乘例程 scipy.sparse.linalg.lsqr。这是最快的，但可能不会在旧的 scipy 版本中可用。它也使用迭代方法。

'sag'使用随机平均渐变下降，'saga'使用其改进的，无偏见的版本 SAGA。两种方法都使用迭代过程，并且在 n_samples 和 n_features 都很大时，它们通常比其他求解器更快。'sag'和'saga'快速收敛只能保证大小相同的特征。可以使用 sklearn.preprocessing 中的缩放器预处理数据。

同时，也可以调用模型的相关属性，例如 coef_来获取权重向量，intercept_获取截距，以及 n_iter_来获取迭代次数。

在本次实验中，主要的 X 特征为 1，所以只需设置一个 alpha 参数即可，同样使用 matplotlib 和R^2评分来评价模型。

参数：alpha=0.5，其他均为默认

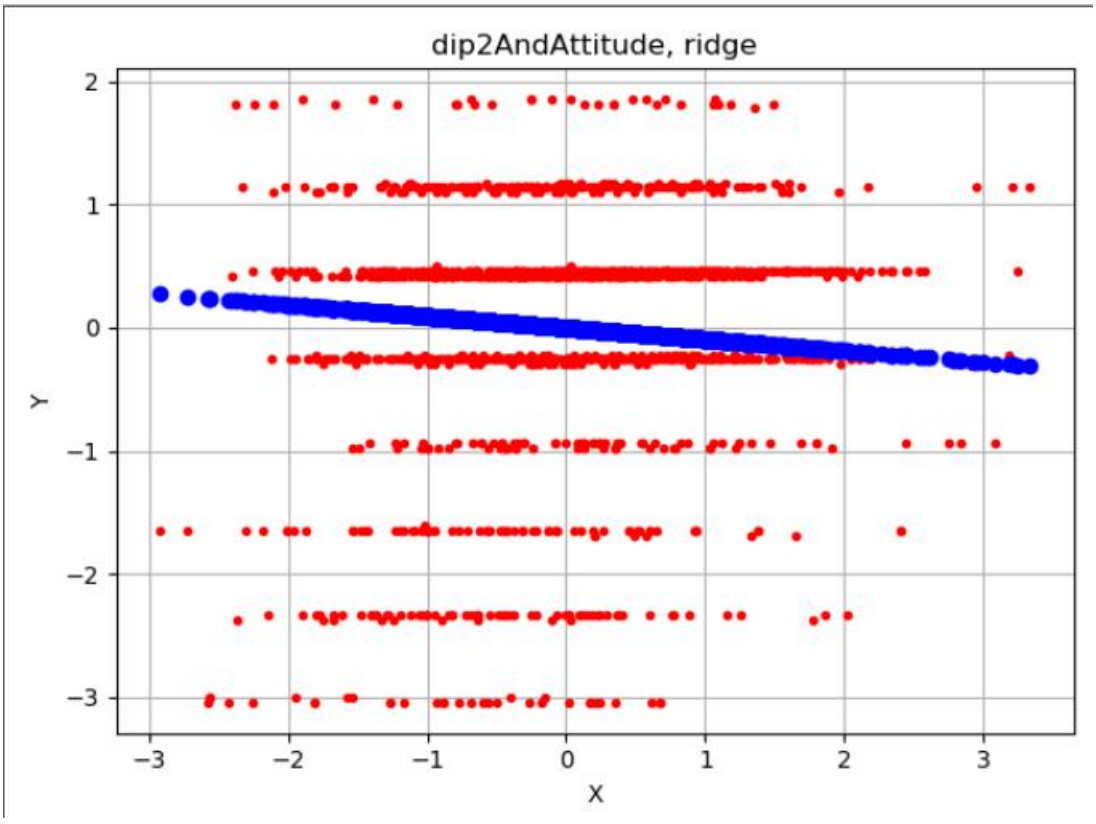


图 3-4

Ridge 的五折 R^2 评分 [-0.62331084 -2.05508438 -0.51603773 -0.13118964 -2.18415972]

3.2.6 svm 支持向量机

支持向量机（support vector ma chine），它是一种二类分类模型，其基本模型定义为特征空间上的间隔最大的线性分类器，其学习策略便是间隔最大化，最终可转化为一个凸二次规划问题的求解，给定一组训练实例，每个训练实例被标记为属于两个类别中的一个或另一个，SVM 训练算法创建一个将新的实例分配给两个类别之一的模型，使其成为非概率二元线性分类器。支持向量机在高维空间中试图找出一个能使支持向量最小的超平面。SVM 模型是将实例表示为空间中的点，这样映射就使得单独类别的实例被尽可能宽的明显的间隔分开。然后，将新的实例映射到同一空间，并基于它们落在间隔的哪一侧来预测所属类别。原本支持向量机属于分类算法，但同样可以运用到回归问题上，在处理分类问题时，由于 svm 的损失函数不在乎边缘之外的训练点，我们要做约束条件是使每个训练集的点尽量的远离自己所属的支持向量，与一般的回归模型不同，svm 不是以标准差作为损失函数的，如同分类一样，因为构建模型的损失函数忽略任何接近于模型预测的训练数据。

支持向量机的优点有：

1. 适合小样本数据，且适合解决高维问题，通过选择合适的核函数以及对支持向量机的调参，可以将低维问题引射到高位解决，而对于支持向量机来说，他最终的决策函数取决于支持向量而非维度，从这一点上避免维数灾难的同时，也让支持向量机适合解决高维问题。
2. 对数据的一定鲁棒性，少数的支持向量决定了最终结果，因此支持向量机可以帮助我们抓住关键样本的同时，通过支持向量“剔除”大量冗余样本,也因此该方法不但算法简单,而且具有较好的“鲁棒”性。这种“鲁棒”性主要体现在: 增、删非支持向量样本对支持向量机本身没有影响;，同时支持向量样本集也具有一定的鲁棒性。

缺点有：

1. SVM 算法对大规模训练样本难以实施，由于 SVM 是借助二次规划来求解支持向量，而求解二次规划将涉及 m 阶矩阵的计算（m 为样本的个数），当 m 数目很大时该矩阵的存储和计算将耗费大量的机器内存和运算时间。针对以上问题的主要改进有有 J.Platt 的 SMO 算法、T.Joachims 的 SVM、C.J.C.Burges 等的 PCGC、张学工的 CSVM 以及 O.L.Mangasarian 等的 SOR 算法。同时，用 SVM 解决多分类问题存在困难 经典的支持向量机算法只给出了二类分类的算法，而在数据挖掘的实际应用中，一般要解决多类的分类问题。可以通过多个二类支持向量机的组合来解决。主要有一对多组合模式、一对一组合模式和 SVM 决策树；再就是通过构造多个分类器的组合来解决。主要原理是克服 SVM 固有的缺点，结合其他算法的优势，解决多类问题的分类精度。如：与粗集理论结合，形成一种优势互补的多类问题的组合分类器。

在 sklearn 中，对于支持向量机的回归实现基于是否为线性核提供了三种形

式，SVR，NuSVR，LinearSVR。在本次实验的代码实现中，我们主要采取了 SVR 实现。

sklearn 中的 SVR 给出了我们许多个可选参数，最重要的几个有：惩罚项 C，规定了在训练损失函数中没有惩罚相关的 epsilon-tube，其中距离实际值在距离 ε 内预测的点数的 epsilon，核函数选项 kernel 等等。

对于惩罚项 C 的取值默认为 1.0，在分类问题时，这里的参数 C 代表的是在线性不可分的情况下，对分类错误的惩罚程度。C 值越大，分类器就越不愿意允许分类离群点。如果 C 值太大，分类器就会竭尽全力地在训练数据上少犯错误，但是就会造成过拟合。C 越小，分类器会越不在乎分类错误，会欠拟合。[10]在解决回归问题时，与岭回归等的正则化罚项相同，C 的大小则影响着模型预测的准确性以及与分类问题相同的拟合度。

对于 epsilon 设置了在终止时可以容忍的预测值以及实际值偏差距离。

对于核函数，核函数是优化支持向量机的一种重要手段，它包含着一个低维到高维空间的映射，从使用的角度来讲，在分类问题中假设一个二维平面的两类数据是线性不可分的，但是如果我们将它通过一个核函数映射到一个高维空间，这样原本在低维空间线性不可分的数据就有可能变成线性可分的了。

其他除了一些只对特定的单独的核函数生效的参数外，还有：

gamma：float，（默认='auto'）'rbf'，'poly'和'sigmoid'的核系数。如果 gamma 是'auto'，那么将会使用 1 / n_features。

max_iter：int，可选（默认值= -1），对求解器中的迭代进行严格限制，或 -1，意为无限制。

cache_size：float，可选，指定内核缓存的大小（以 MB 为单位）。

tol：float，可选（默认= 1e-3），容许停止标准。

shrinking : boolean, optional (default=True)，是否使用缩小的启发式

同时可以调用模型的 support_vector 等属性，获取模型的支持向量，以及 dual_coef_获取支持向量在决策函数中的系数。coef_ ,intercept_等属性，以获取 SVR 模型相关参数。

在支持向量机中，数据通过核函数(如线性核函数,多项式 函数,高斯核函数等)映射到高维空间,[11]在机器学习中常用的核函数，一般有如下几类，也就是 libSvm（sklearn 中的 svm 是基于 libSvm 实现的）中自带的这几类：

1）线性核函数：$K(V_1,V_2)=<V_1,V_2>$（不投射到高维空间）

2）多项式核函数：$K(V_1,V_2)=(\gamma<V_1,V_2>+C)\mathrm{n}$

3）高斯核函数：$K(V_1,V_2)=\exp(-\gamma\|V_1-V_2\|2)$

4）Sigmoid 核函数：$K(V_1,V_2)=\tanh(\gamma<V_1,V_2>+c)$

在回归问题中，同样可以通过核函数寻找最小的残差，以便寻找最优解。

在本次实验中，主要实验了 C=2.0,核函数分别取四种常用核函数的情况，结果如下（红点为实际值，蓝点为预测值）：

1. 线性核函数

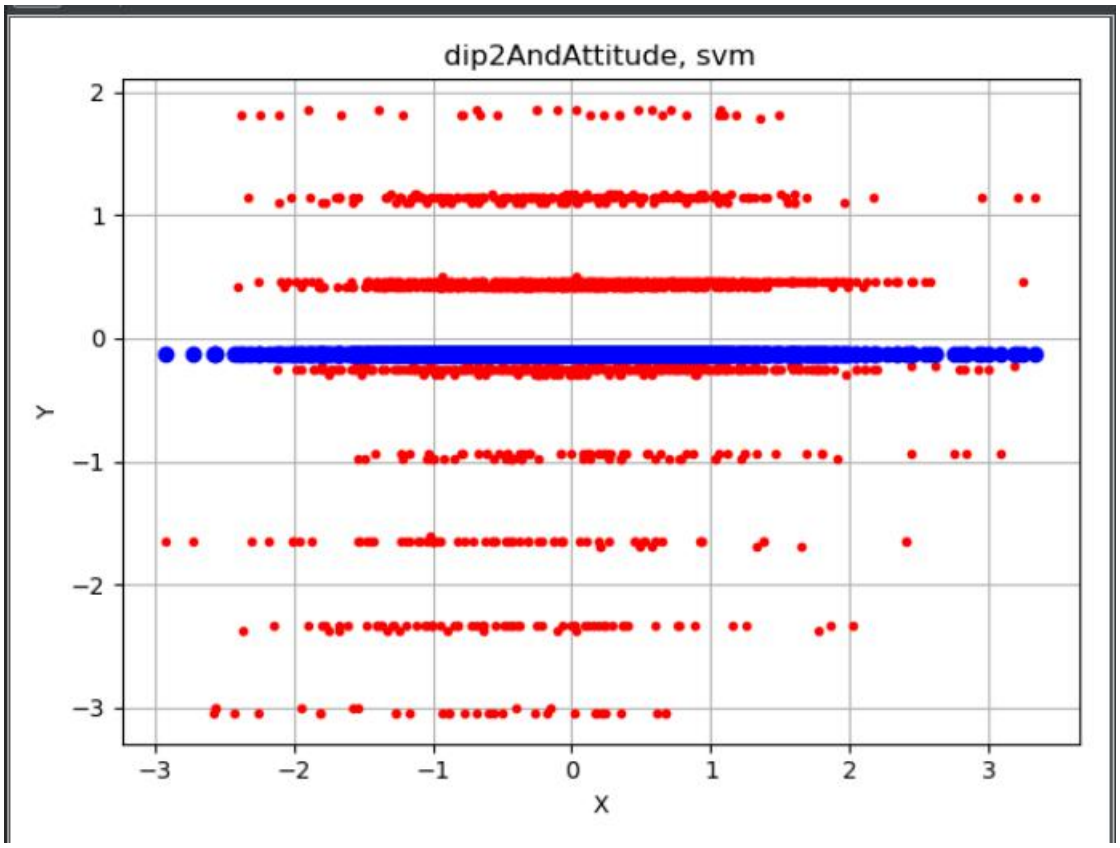


图 3-5

svm 五折R^2评分

[-0.19583767 -1.49859112 -0.12452708 -0.23864549 -2.28438719]

2. 多项式核函数

多项式核函数有额外的可选参数 degree，默认值为 3，代表着多项式函数的幂数，这个参数会被其他的核函数所忽略

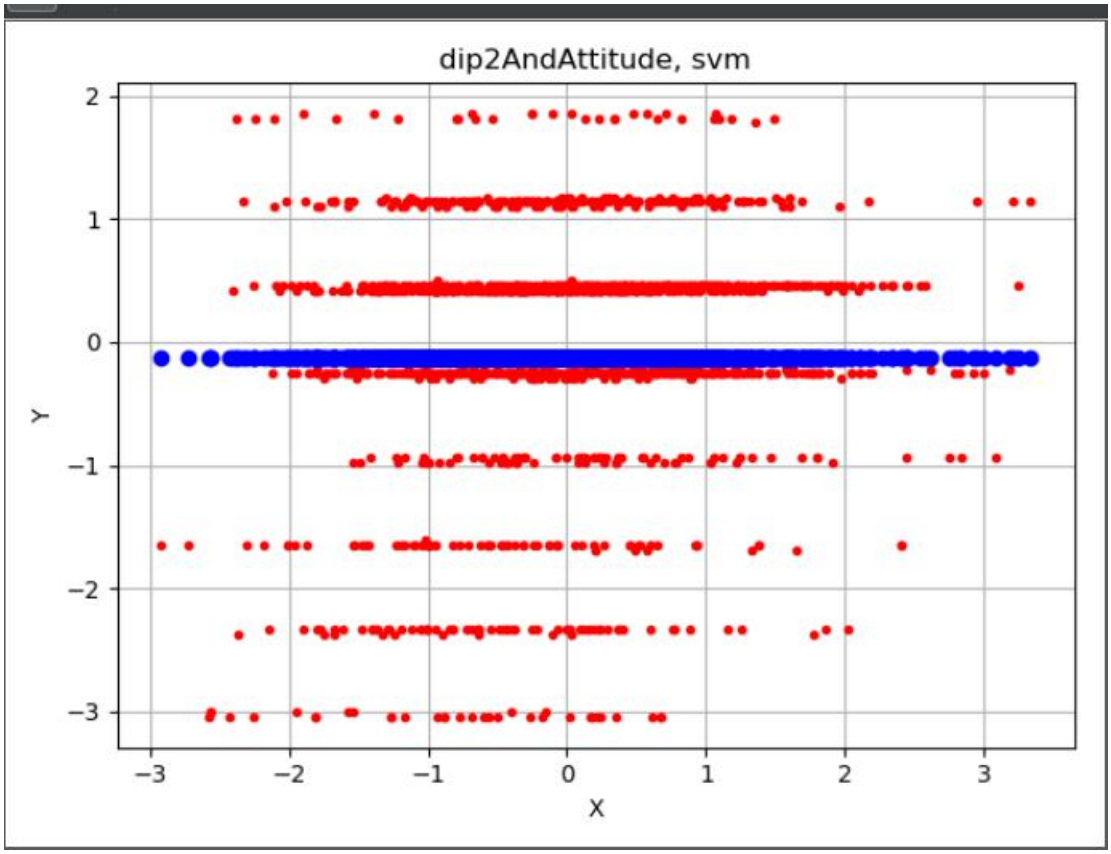


图 3-6

svm 五折R^2评分

[-0.01784884 -1.52324202 -0.04531362 -0.19758432 -2.19849096]

3. 高斯核函数

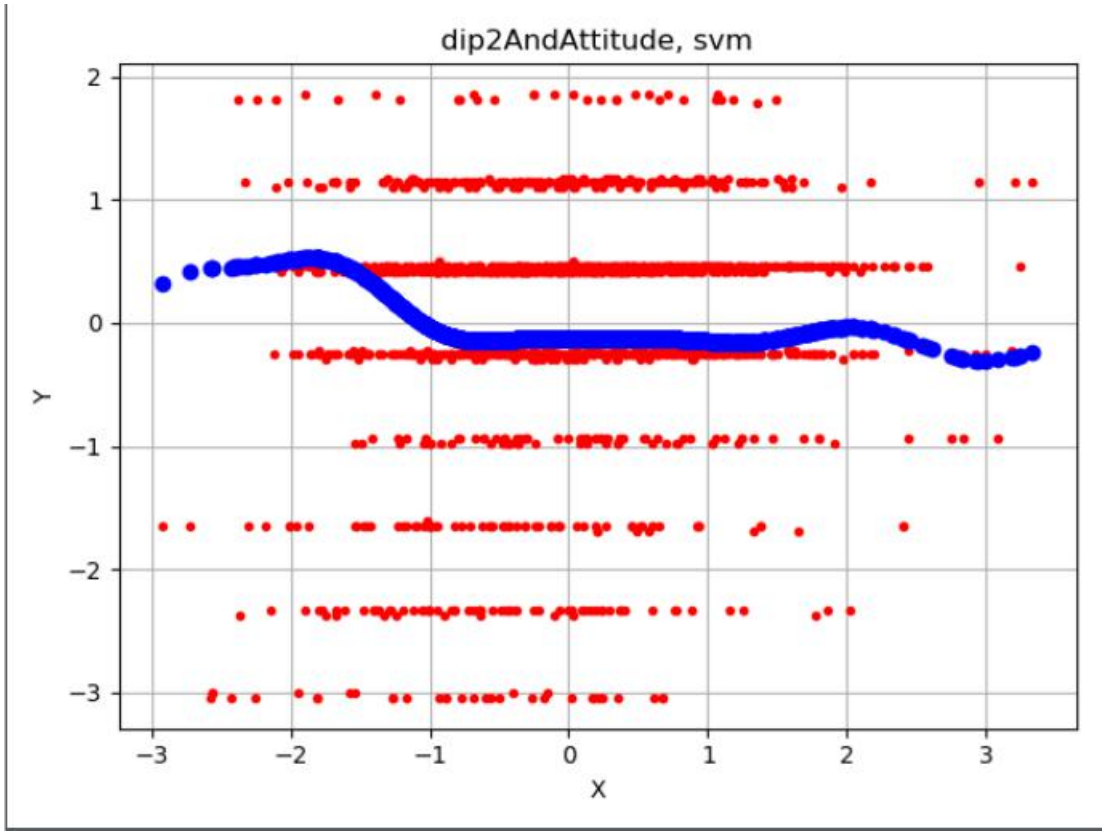


图 3-7

4. Sigmoid 函数

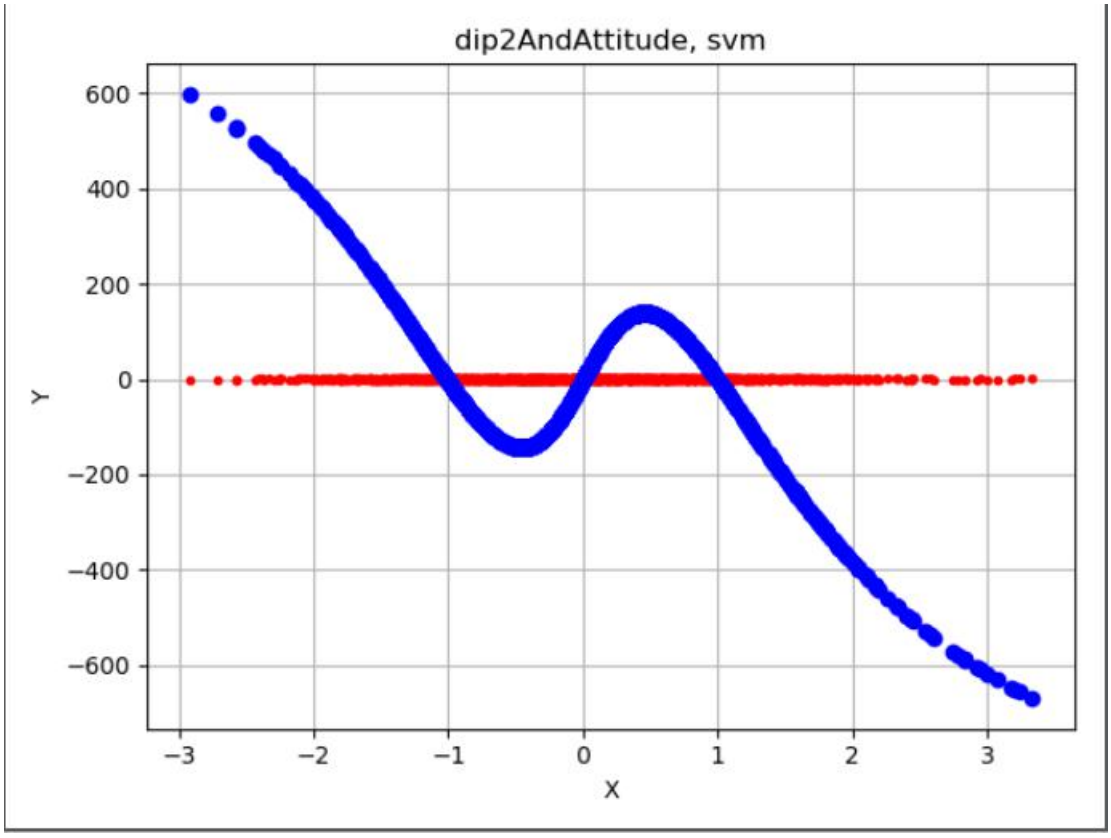


图 3-8

svm 五折R^2评分
[-67072.20278712 -39319.53516942 -57316.44448761 -25342.1704104]

3.2.7 神经网络模型

神经网络是一个试图模仿自然生物神经网络的学习模式的机器学习框架。 生物神经网络具有相互连接的神经元，神经元带有接受输入信号的树突，然后基于这些输入，它们通过轴突向另一个神经元产生输出信号。而人工神经网络具有生物神经系统的基本特征，在一定程度上反映了人脑功能的若干反映，是对生物系统的某种模拟，具有大规模并行、分布式处理、自组织、自学习等优点，被广泛应用于语音分析、图像识别、数字水印、计算机视觉等很多领域，取得了许多突出的成果。人工神经网络模型（ANN）作为一种优秀的监督学习模型，其优点是明显的：

1. 自学习和自适应能力：BP 神经网络在训练时，能够通过学习自动提取输出、输出数据间的“合理规则”，并自适应的将学习内容记忆于网络的权值中。

2. 非线性映射能力：BP 神经网络实质上实现了一个从输入到输出的映射功能，数学理论证明三层的神经网络就能够以任意精度逼近任何非线性连续函数。这使得其特别适合于求解内部机制复杂的问题。

3. BP 神经网络在其局部的或者部分的神经元受到破坏后对全局的训练结果不会 造成很大的影响，也就是说即使系统在受到局部损伤时还是可以正常工作的。

4. 泛化能力：所谓泛化能力是指在设计模式分类器时，即要考虑网络在保证对所需分类对象进行正确分类，还要关心网络在经过训练后，能否对未见过的模式或有噪声污染的模式，进行正确的分类。也即 BP 神经网络具有将学习成果应用于新知识的能力。

基于以上这些优点，缺点也是显而易见的：

1. 局部极小化问题：从数学角度看，传统的 BP 神经网络为一种局部搜索的优化方法，它要解决的是一个复杂非线性化问题，网络的权值是通过沿局部改善的方向逐渐进行调整的，这样会使算法陷入局部极值，权值收敛到局部极小点，从而导致网络训练失败。加上 BP 神经网络对初始网络权重非常敏感，以不同的权重初始化网络，其往往会收敛于不同的局部极小，这也是很多学者每次训练得到不同结果的根本原因。

2. BP 神经网络算法的收敛速度慢：由于 BP 神经网络算法本质上为梯度下降法，它所要优化的目标函数是非常复杂的，因此，必然会出现“锯齿形现象”，这使得 BP 算法低效；又由于优化的目标函数很复杂，它必然会在神经元输出接近 0 或 1 的情况下，出现一些平坦区，在这些区域内，权值误差改变很小，使训练过程几乎停顿；BP 神经网络模型中，为了使网络执行 BP 算法，不能使用传统的一维搜索法求每次迭代的步长，而必须把步长的更新规则预先赋予网络，这种方法也会引起算法低效。

3. 容易出现过拟合现象，神经网络的算法导致他容易从样本的细节中过多学习。从而忽略了样本特征之间真正存在的联系。造成网络训练能力优秀，预测能力差的状况。

4. 众所周知，神经网络模型理论上的隐藏层是不定的，如果我们针对某个问题研究出了一张极大的神经网络图，在实际应用中，极大的神经网络能否商业化落地，是否对整个系统的实时性与速度造成负担，是一个极大的问题。

在国内外许多方面，对于神经网络模型的可借鉴的成功应用经验有很多，例如：GRUDNITSKI, OSBURN 应用神经网络对 S&P 指数和黄金的期价进行了预测[3]。DEMATOS 等基于前馈网络模型对日元汇率进行了预测[4]。SHAIKH 等用神经网络预测标普 500 指数期货价格[5]。[6]

构建神经网络模型从构建最基本形式单层感知器开始，感知器具有一个或多个输入，偏置，激活函数和单个输出。感知器接收输入，将它们乘以一些权重，然后将他们传递到激活函数用以产生输出有许多激活函数可供选择，例如逻辑函数，三角函数，阶跃函数等。我们还确保向感知器添加偏差，这避免了所有输入可能等于零的问题（意味着没有乘权重会有影响）。

下图表示了感知器的工作原理：

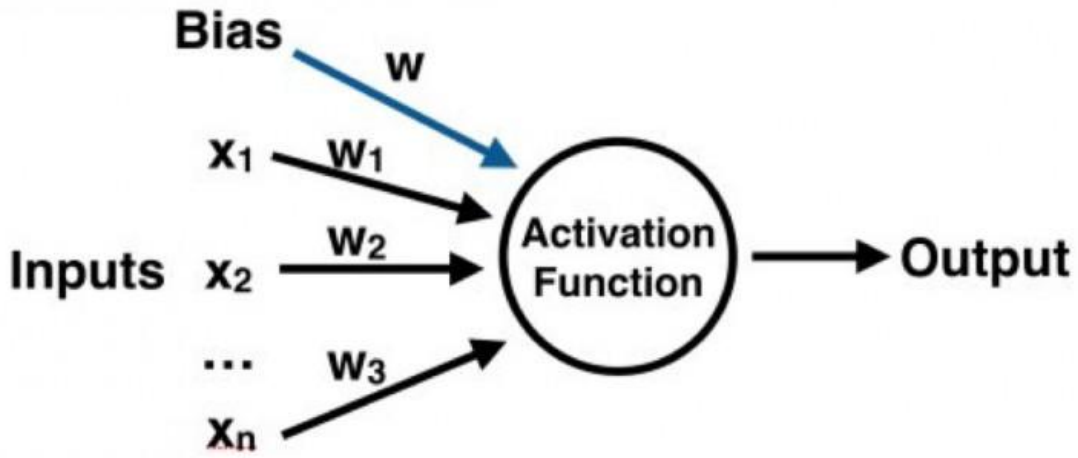


图 3-9

一旦我们有输出，我们可以将其与已知标签进行比较，并相应地调整权重（权重通常以随机初始化值开始）。 我们继续重复此过程，直到我们达到允许迭代的最大数量或可接受的错误率。

创建神经网络时，我们从叠加感知器层开始创建多层感知器模型（MLP），我们需要一个接收输入的输入层，一个输出结果的输出层，中间有数目不定的感知层，这些感知层被叫做隐藏层。我们不需要知道中间的感知层，因为我们并不关心他们的输入输出。下图表现了一个基于多层感知器的神经网络模型。

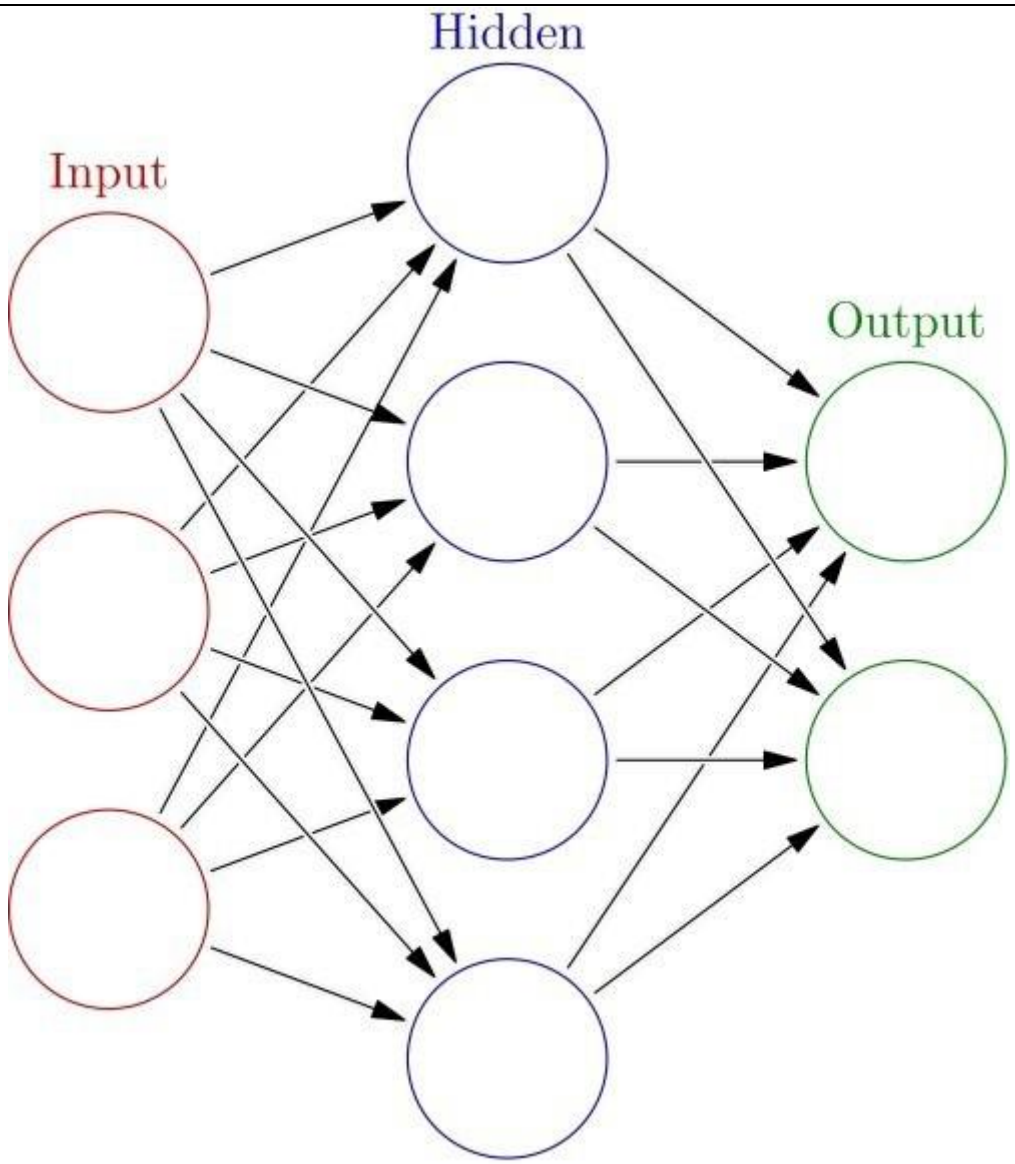


图 3-10

神经网络模型中的反向传播法，反向传播法适用于更新神经网络。在使用神经网络进行预测时，首先进行前向传播，我们假设一个具有一个输入层，两个隐藏层，的三层神经网络模型，除了输出层的每一层具有两个神经元，步骤为：计算神经元 h1 的输入加权和（输入层的输入加权加截距）→通过神经元 h1 的激活函数得出神经元的输出 o1→计算神经元 h2 的输入加权和→通过神经元 h2 的激活函数得出神经元的输出 o2（隐藏层计算完成）→输出层（第二个隐藏层同理）→前向传播过程结束，神经网络具体结构如图所示：

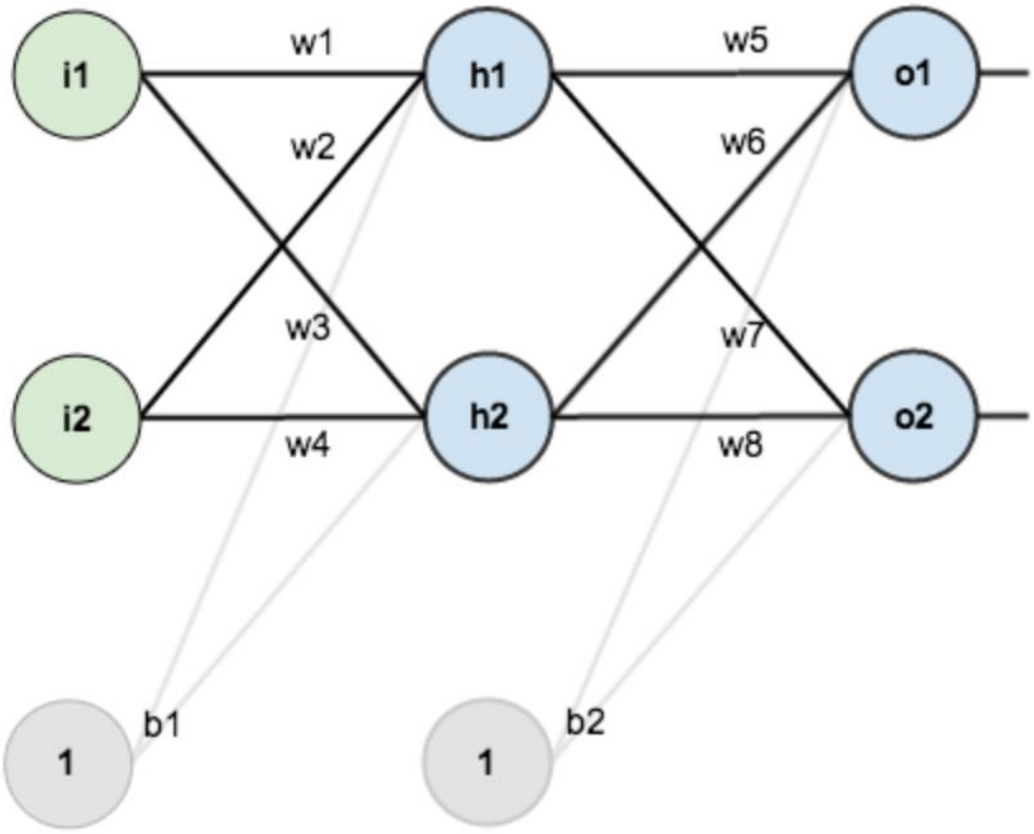


图 3-11

之后进入反向传播过程：

1. 计算总误差：$E_{total}=\sum\frac{1}{2}(target-output)^2$，根据这个公式，分别计算 o1 和 o2 的误差，总误差为两者之和。
2. 隐藏层→输出层的权值更新：通过计算整体误差对某一个权值的偏导计算，误差的值（通过链式求导法则）。

$$\frac{\partial E_{total}}{\partial w_5}=\frac{\partial E_{total}}{\partial out_{o1}}*\frac{\partial out_{o1}}{\partial net_{o1}}*\frac{\partial net_{o1}}{\partial w_5}$$

3. 最后根据学习速率更新权值 $w_5(new)=w_5(old)-\eta*\frac{\partial E_{total}}{\partial w_5}$，其中$\eta$为学习速率。

神经网络模型中的梯度下降法，梯度下降法与最小二乘法一样，是求解无约束最优问题的最常用方法之一，梯度下降法也应用于神经网络模型中，最为最常用的优化方法。

在微积分中，对一个多元函数的参数求偏导数，并把这些偏导数用向量的方式表现出来，这个向量就是梯度。从几何意义上讲，梯度向量所指的方向，就是函数增加/变化最快的地方，更容易找到函数的最大值。与梯度向量相反的方向就是

变化最慢的地方，更容易找到函数的最小值。

在机器学习算法中，在最小化损失函数时，可以通过梯度下降法来一步步的迭代求解，得到最小化的损失函数，和模型参数值。反过来，如果我们需要求解损失函数的最大值，这时就需要用梯度上升法来迭代了。

梯度下降法和梯度上升法是可以互相转化的。比如我们需要求解损失函数 f(θ)的最小值，这时我们需要用梯度下降法来迭代求解。但是实际上，我们可以反过来求解损失函数 -f(θ)的最大值，这时梯度上升法就派上用场了。

梯度下降法同样存在问题，即梯度下降寻找的最优解，是在起始点的最优解，如果起始点并不好，通过梯度下降得到的最优解可能是局部最优解，不过如果寻优函数是凸函数的话，由于其增长性质则不会存在这个问题。

梯度下降法在应用之前需要确定模型的假设函数和损失函数，以及设定步长a以及算法终止距离ε，对假设函数而言它的损失函数可以设定为它的残差平方和，即 Mean Square Loss 损失函数,此时以他的损失函数作为梯度下降法的目标函数，对其每个参数求偏导计算梯度向量，引用步长参数a 乘以梯度向量，如果最后所有参数的梯度向量乘以步长结果均小于ε，算法结束，返回结果。否则则更新所有的模型参数（使用当前参数减去步长与梯度值之乘积）返回最初的步骤继续进行下降。

从随机梯度下降延伸的还有批量梯度下降法，于随机梯度下降不同的是批量梯度下降所求的是最小化所有训练样本的损失函数，使得最终求解的是全局的最优解，即求解的参数是使得风险函数最小。而随机梯度下降所做的是最小化每条样本的损失函数，虽然不是每次迭代得到的损失函数都向着全局最优方向， 但是大的整体的方向是向全局最优解的，最终的结果往往是在全局最优解附近。[13]

在 sklearn 中 neural_network 包的 MLPRegressor 类实现了用于回归的神经网络，同时也是一个基于反向传播算法的前馈人工神经网络。在输出层去掉了激活函数，或者可以认为输出层的激活函数是一个恒等函数，从而输出一组连续的值作为结果。

同时 MLPRegressor 类为我们提供了许多参数，包括：

hidden_layer_sizes：(接收元组输入，元组长度为总感知器层数减 2，第 i 个元素表示第 i 个隐藏层中的神经元数量)。

Activation：隐藏层的激活函数，有线性函数，logistic sigistic 函数，tanh 双曲线 tan 函数，relu 整数线性单位函数等等。

Solver：求解器，包括 lbfs 准牛顿方法族的优化器，sgd 随机梯度下降法以及 adam 基于随机梯度的优化器三种。

Alpha：L2 正则化项惩罚参数

Batch_size:批量随机梯度下降的批量，不可大于特征数目。

Learning_rate：用于更新权重的学习率策略，可选值 {‘constant’, ‘invscaling’, ‘adaptive’}, 默认值为‘constant’

constant：等同于初始学习速率的学习速率。

Invscaling：根据初始学习速率，使用’power_t’的反比例指数逐步减少每个时间步’t’的学习率（effective_learning_rate = learning_rate_init / pow(t, power_t)）。

adaptive：只要训练损失持续下降，“适应性”就会将学习速率保持不变至“learning_rate_init”。每次连续两个时期至少将训练损失降至 tol，或者如果

'early_stopping'开启时未能将验证分数至少提高到 tol,则当前学习速率除以 5。仅当求解器=“sgd”生效。
learning_rate_init:默认值 0.001，使用的初始学习率，控制权重的步长，仅当求解器=“sgd”或者“adam”时生效
power_t:上文提到的反比例学习参数，仅当求解器='sgd'时有效，默认值为 0.5

Max_iter: 整型，默认值为 200，最大迭代次数，求解器迭代直到收敛或者到最大迭代次数 Max_iter。
Random_state:整型，RandomState 实例或者 none，表示随机数生成器使用的种子，如果是定义的 RandomState 实例，如果没有，随机数生成器则使用 numpy 中的 random 方法。
Tol：优化的容差，当损失在连续两次迭代中没有改善的时候，除非学习率被设定为‘adaptive’，否则认为函数达到收敛并且训练停止。
Verbose:布尔型，默认为 False，是否将进度消息打印到标准输出。
warm_start:布尔型，默认为 False，是否重新调用先前的解决方案初始化。
Momentum：浮点型，默认 0.9，梯度下降更新的动量。应该在 0 和 1 之间。仅当求解器='sgd'时使用。
nesterovs_momentum：布尔型，默认为 True，是否使用 Nesterov’s 动量。仅当求解器='sgd'且动量> 0 时使用。
early_stopping：布尔型，默认为 False，当验证分数没有提高时是否使用提早停止来终止训练。如果设置为 true，则当验证分数未改善至少两个连续迭代的时间时，它将自动预留 10％的训练数据作为验证并终止训练。只有在求解器='sgd'或'adam'时才有效。
validation_fraction：float，可选，默认为 0.1，训练数据的比例作为验证集提前停止。必须在 0 和 1 之间。仅当 early_stopping 为 True 时才使用。
beta_1：float，可选，默认为 0.9，估计 adam 中第一个矩矢量的指数衰减率应该在[0,1）。仅当求解器='adam'时才使用。
beta_2：float，可选，默认值为 0.999，adam 中二阶矩矢量估计的指数衰减率应该在[0,1）。仅当求解器='adam'时才使用。
epsilon：float，可选，adam 中数值稳定性的值。仅当求解器='adam'时才使用

sklearn 中的 natural_network 同时提供了一些属性来输出模型相关信息。
loss_：float，用损失函数计算当前损失。
coefs_：列表，长度 n_layers - 1，列表中的第 i 个元素表示对应于第 i 层的权重矩阵。
intercepts_：列表，长度 n_layers - 1，列表中的第 i 个元素表示与层 i + 1 对应的偏向量。
n_iter_：int，求解器运行的迭代次数。
n_layers_：int，层数。
n_outputs_：int，输出数量。
out_activation_：string，输出激活函数的名称。

在本次实验中，主要对损失函数和相关参数进行了调整，求解器均使用 ada 或 sgd，同样通过 R^2 评分以及 matplotlib 画图的方式展现结果。

本次实验大多是寻找单个自变量 x 与单个因变量 y 的之间的关系，而且按照文本文件分类的每天的数据例为 1500 组左右，所以将模型设置为 3 层，每层 200 神经元。

1. 参数 hidden_layer_sizes=（200，200，200），Activation=‘relu’其余参数均为默认。

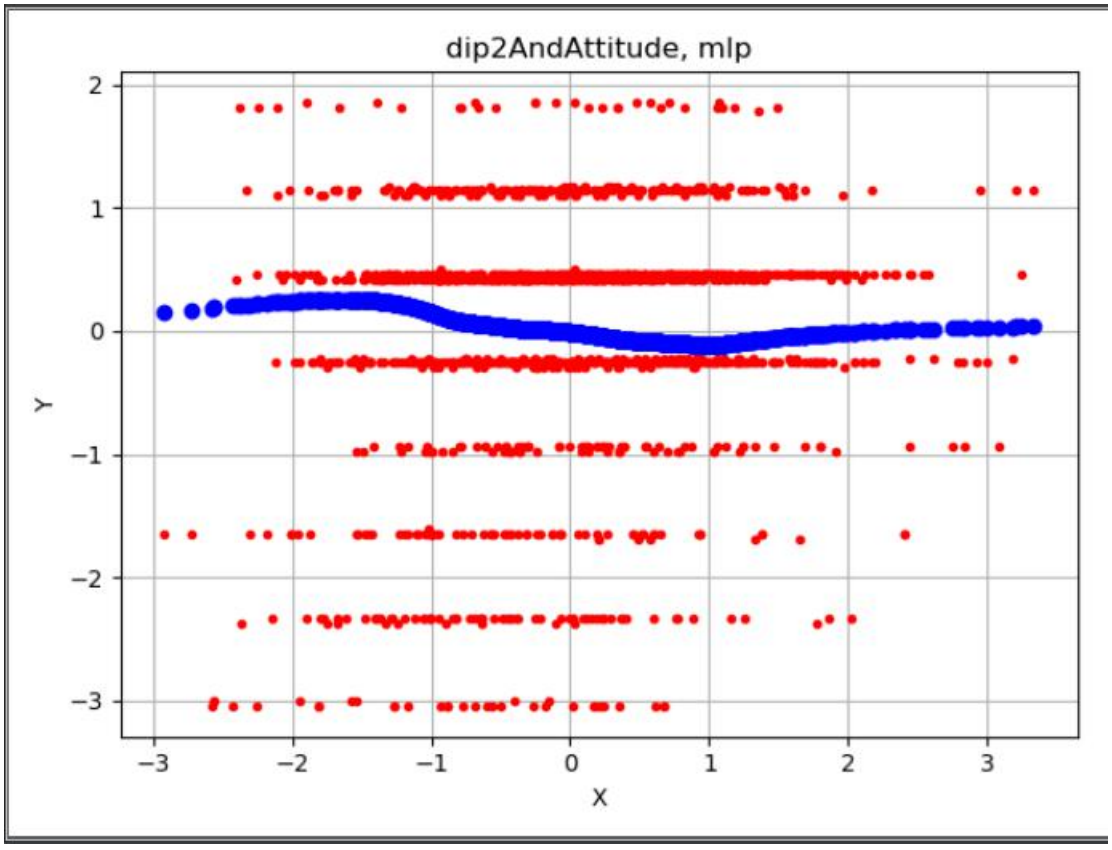


图 3-12

MLP 的五折 R^2 评分为[-0.58794938 -2.39491289 -0.48171342 -0.197272 -2.29071547]。

2. 参数 hidden_layer_sizes=（200，200，200），Activation=‘tanh’其他参数均为默认。

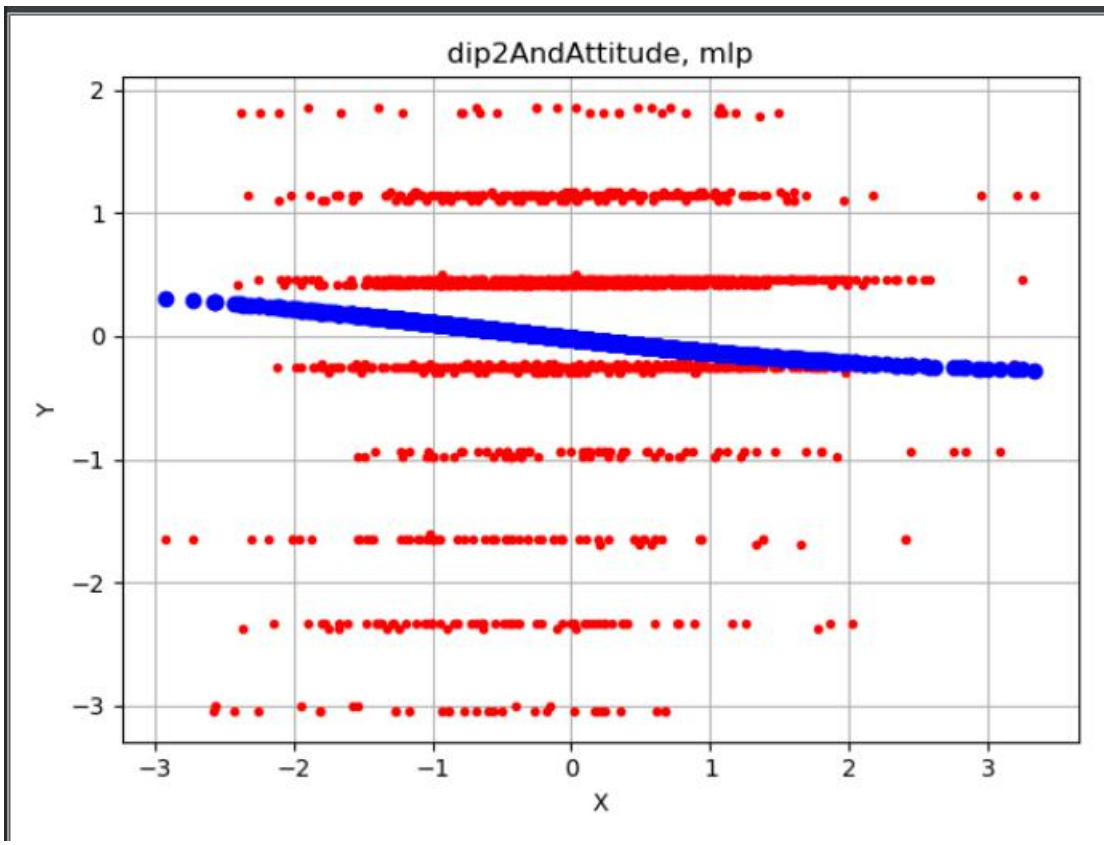


图 3-13

MLP 的五折 R^2 评分为[-0.61720031 -2.07269075 -0.37524578 -0.15034237 -1.8940353]。

3. 参数 hidden_layer_sizes=（200，200，200），Activation='logistic' 其他参数均为默认。

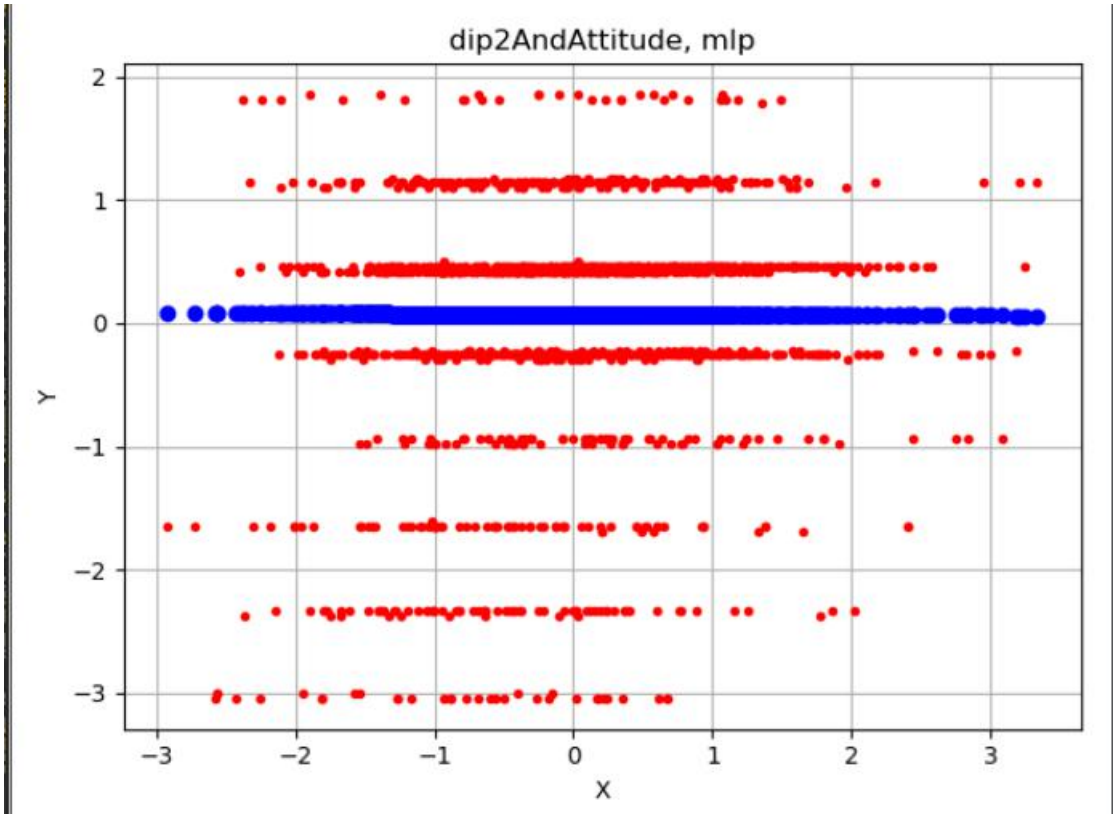


图 3-14

MLP 的五折 R^2 评分为[-0.72740066 -2.0734581 -0.71547158 -0.01157113 -2.11727837]。

4. 参数 hidden_layer_sizes=（200，200，200），Activation=‘identity’其他参数均为默认。

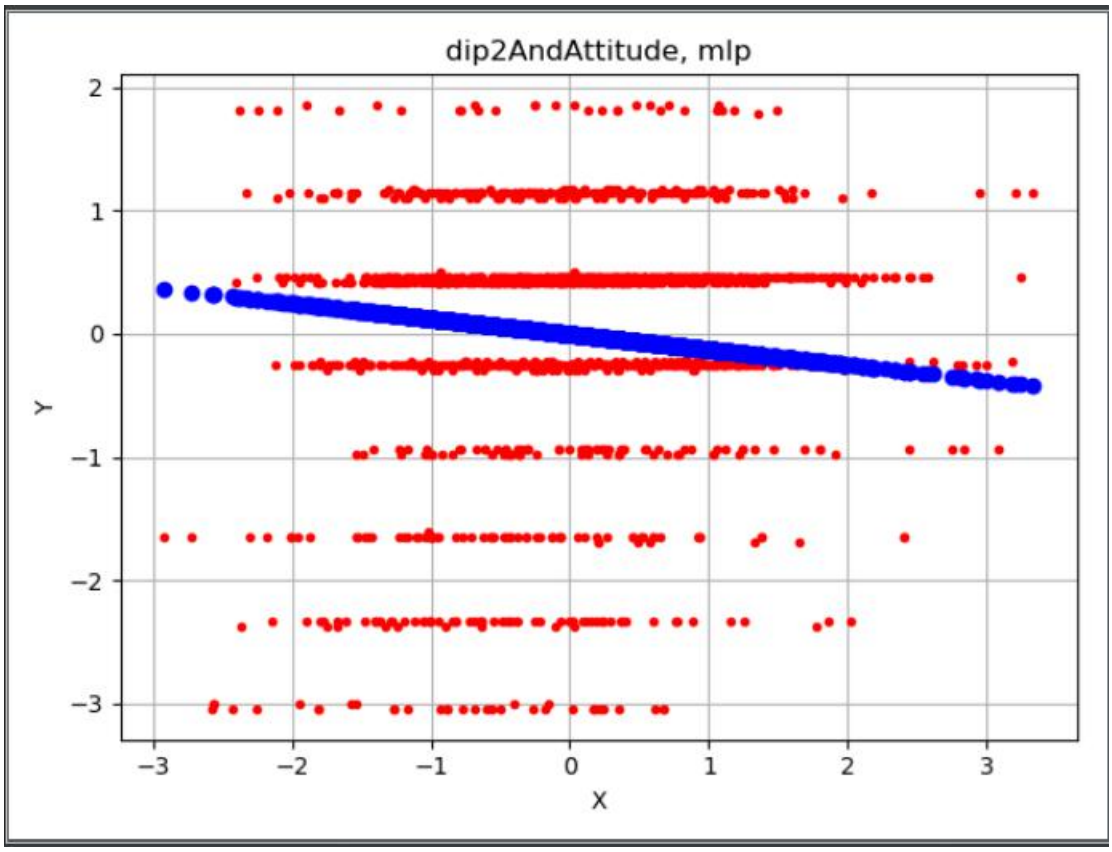


图 3-15

MLP 的五折 R^2 评分为[-0.5733593 -1.71896078 -0.51353679 -0.17871959 -2.27628277]。

3.2.8 模型选择与验证

对于评估一个模型估算器的表现时，测试集与训练集使用同一组数据时容易造成该模型获得极高的分数的情况，因为他对同一组数据可能预测出较好的结果，但是对没有出现过的数据，他则可能无法预测出较为准确的结果，出现过拟合的情况，对于这种情况，在进行实验的时候，可以将一部分数据集按照比例分为测试集和训练集。但是在划分测试集与训练集的时候，仍然无法保证同一份数据集测试集与训练集不会受到其他因素的影响，鉴于此，我们可以使用交叉验证的方法解决问题。

交叉验证主要通过将数据集分成若干份的方法来进行验证，加入我们将数据集分成 5 份，那么便被称为 5 折交叉验证，交叉验证通过每次改变训练集与测试集，对交叉训练进行重复达到每次训练集和测试集都不相同，达到最终整个训练集都做过测试集的目标。

在 sklearn 中的 model_selection 模块中包含了包括 RepeatedKFold（重复 k 折交叉验证），留一交叉验证（一种特殊的 k 折交叉验证）等方法，以重复 k 折交叉验证为例，它接受的参数包括：

n_splits：折叠数量，即将数据集划分为几份，应大于 2。

n_repeats：交叉验证器需要重复的次数。

Random_state:整型，RandomState 实例或 None，可选，默认值为 None 如果是 int，random_state 是随机数发生器使用的种子； 如果 RandomState 实例，random_state 是随机数生成器； 如果没有，随机数生成器将会使用 numpy 的 np.random 方法。

在本次实验中，类似与神经网络模型的模型需要调整大量的超参数以取得模型的最优拟合。网格追踪法是一种优秀的适合寻找最优超参数的方法。网格搜索法的原理是通过限定一个或几个超参数的范围通过不断的尝试找出最优超参数组合。

在 sklearn 中，提供了 GridSearchCV 相关的包，通过指定 param_grid 相关参数的值，可以对所需要寻优的模型，超参数以及范围中做出限定。同时 sklearn 也为我们提供了随机参数优化的方法，与网格搜索不同的是，随机搜索需要指出参数可能的分布范围，通常通过 scipy 相关的包实现。

3.2.9 量化评估模型

量化评估模型是指通过一些可以反映模型预测准确度的数学计算方法，得出确切的数值，来评估模型是否适合某组数据，数据之间是否存在关系，以及超参数的设置等问题的。

本次实验讨论的主要是回归问题，对于回归问题来讲主要有解释方差得分，平均绝对误差，均方误差，均方误差对数，中位绝对误差，R^2 score 可决系数等评价方法，这也是 sklearn 框架为我们提供了实现的主要的六种用于评价回归模型的评分方法。

1. 解释方差得分

如果 $\hat{y}$ 是预估的目标输出，y 是相应（正确的）目标输出，并且 var 是方差，标准差的平方，那么解释的方差预估如下:

$$\exp lained\_\text{var}iance(y,\hat{y})=1-\frac{\text{var}\{y-\hat{y}\}}{\text{var}\{y\}}$$

最好的得分是 1.0，值越低越差。在 sklearn 中 metrics 中的 explained_variance_score 中实现，接收目标值与预测值两个主要参数，同时可以通过参数定义样例权重以及输出方式即是否对多个输出分数进行汇总。

2. 平均绝对误差

平均绝对误差是一个对应绝对误差损失预期值或者 $l1-norm$ 损失的风险度量.

如果 $\hat{y}_i$ 是第 i 个样本的预测值，并且 y_i 是对应的真实值，则平均绝对误差 (MAE) 预估的 $n_{samples}$ 定义如下：

$$MAE(y,\hat{y})=\frac{1}{n_{samples}}\sum_{i=0}^{n_{samples}-1}|y_i-\hat{y}_i|$$

在 sklearn 中 mean_absolute_error 函数给出了计算平均绝对误差的方法实现，同样接收目标值与预测值两个主要参数，同时可以通过参数定义样例权重以及输出方式即是否对多个输出分数进行汇总。

3. 均方误差

均方误差是一个对应平方（二次）误差或损失的预期值的风险度量，如果 $\hat{y}_i$ 是第 i 个样本的预测值，并且 y_i 是对应的真实值，则平均绝对误差（MAE）预估的 $n_{samples}$ 定义如下：

$$MAE(y,\hat{y})=\frac{1}{n_{samples}}\sum_{i=0}^{n_{samples}-1}(y_i-\hat{y}_i)^2$$

在 sklearn 中的 mean_squard_error 包中实现了均方误差的方法。同样接收目标值与预测值两个主要参数，同时可以通过参数定义样例权重以及输出方式即是否对多个输出分数进行汇总。

4. 中位绝对误差

中位绝对误差的离群值很强，它通过取目标和预测之间的所有绝对差值的终止来计算损失，如果 $\hat{y}_i$ 是第 i 个样本的预测值，并且 y_i 是对应的真实值,则中位绝对误差（MedAE）预估的 $n_{samples}$ 定义如下：

$$MedAE(y,\hat{y})=median(|y_1-\hat{y}_1|,...,|y_n-\hat{y}_n|)$$

在 sklearn 中的 median_absolute_error 包中实现了均方误差的方法。同样接收目标值与预测值两个主要参数,sklearn 中的中位绝对误差不支持对输出做平均值等计算。

5. 均方误差对数

均方误差对数是一个对应平方（二次）误差或损失的预期值的风险度量，如果 $\hat{y}_i$ 是第 i 个样本的预测值，并且 y_i 是对应的真实值，则平均绝对误差（MAE）预估的 $n_{samples}$ 定义如下：

$$MSLE(y,\hat{y})=\frac{1}{n_{samples}}\sum_{i=0}^{n_{samples}-1}(\log_e(1+y_i)-\log_e(1+\hat{y}_i))^2$$

其中 $\log_e x$ 表示 x 的自然对数. 当目标具有指数增长的趋势时，该指标最适合使用，例如人口数量，跨年度商品的平均销售额等。 请注意，该指标会对低于预测的估计值进行估计。在 sklearn 中的 mean_squaerd_log_error 函数实现了均方误差对数。对于均方误差对数同样接收目标值与预测值两个主要参数，同时可以通过参数定义样例权重以及输出方式即是否对多个输出分数进行汇总。

6. R^2score 可决系数

R^2score 计算了 computes R²，即可决系数。它提供了将来样本如何可能被模型预测的估量. 最佳分数为 1.0，可以为负数（因为模型可能会更糟）。[12]总是预测 y 的预期值，不考虑输入特征的常数模型将得到 R^2 得分为 0.0。如果 $\hat{y}_i$ 是第 i 个样本的预测值，并且 y_i 是对应的真实值，则 R^2 得分预估的 $n_{samples}$ 定义如下：

$$R^2(y,\hat{y})=1-\frac{\sum_{i=0}^{n_{samples}-1}(y_i-\hat{y}_i)^2}{\sum_{i=0}^{n_{samples}-1}(y_i-\bar{y})^2}$$

其中 $\bar{y}=\frac{1}{n_{samples}}\sum_{i=0}^{n_{samples}-1}y_i$

在 sklearn 中的 r2_score 实现了 R^2score，对于 R^2score 同样接收目标值与预测值两个主要参数，同时可以通过参数定义样例权重以及输出方式即是否对多个输出分数进行汇总。

这些评分手段的作用是可以在 sklearn 中的模型选择模块中的 cross_val_score 函数中作为参数使用。cross_val_score 接受的参数有：

X:python list 或者 numpy array，模型学习的参数，即因变量。

Y:在监督学习的情况下试图预测的目标变量。

Groups：在将数据集分解成训练/测试集时使用的样本的组标签。

Scoring:字符串，可调用或无，可选，默认值：无。一个字符串，或者一个可调用的评分模型。上文提到的六种评分模型便应用于此。

Cv:整形，交叉验证生成器，即确定交叉验证的 k 折的折数 k。cv 可能的输入是：无，要使用默认的 3 折交叉验证，整数，以指定（分层）K 折叠中的折叠次数，一个要用作交叉验证生成器的对象。一个训练测试集的分割的迭代器。对于整数/无输入,如果估计器是分类器并且 y 是二进制或多分类的,则使用 StratifiedKFold 作为交叉验证法。在所有其他情况下，使用 KFOLD 作为交叉验证法。该函数返回一个浮点型数组，代表每次交叉验证的评分。

同在模型选择模块的 cross_validate 函数与 cross_val_score 函数在两个方面有些不同：

1. 它允许指定多个指标进行评估。
2. 除了测试得分之外，它还会返回一个包含训练得分，拟合次数， score-times（得分次数）的一个字典。

cross_validate 函数与 cross_val_score 函数接受的参数基本相同，除了对于单个度量评估，其中 scoring 参数是一个字符串，可以调用或 None，keys 将是 -['test_score','fit_time','score_time']。

而对于多度量评估，返回值是一个带有以下的 keys 的字典
-['test_<scorer1_name>','test_<scorer2_name>','test_<scorer...>','fit_time','score_time']

3.2.10 绘图工具包可视化展示结果

Python 中的 matplotlib 包提供了许多简单的方法用于绘制数据统计图表，引入 matplotlib 包之后，可以先调用其 figure()方法，创建一个新的图像，figure 方法接收很多参数，包括：

1. Num:整数或字符串，可选，默认值：无。如果没有提供，将会创建一个新图，图号将会增加，图号将会被保存在一个 object 中。
2. Figsize：整数元组，可选，默认值：无 。宽度，高度以英寸为单位。如果未提供，则默认为 figure.figsize。
3. dpi:整数，可选，默认值：无,图的分辨率。如果未提供，则默认为 figure.dpi。
4. Facecolor:背景颜色
5. Edgecolor：边颜色
6. frameon：bool，可选，默认值：True，如果是 False，则禁止绘制数字框。
7. 图类：派生自 matplotlib.figure.Figure 的类，可以选择使用自定义图形实例。

Matplotlib 包中的 title 方法可以设定标题。同时可以使用其中的 xlabel，ylabel 方法设置 x 轴与 y 轴的上的字表示下 x 与 y 轴的字。通过 grid 方法展示网格。最后使用 plot 方法，plot 方法是主要的画图（点）方法。可以设置具体的数值与数值点的颜色。Plot 方法接受可变参数，例如：plot(x, y)意为 plot x, y 使用默认的线条样式和颜色,plot(x, y, 'bo')意为 plot x,y 用蓝色圆圈标记,plot(y)意为 plot y 用 x 作为自变量,plot(y, 'r+')意为 y 用 x 作为自变量，但是是使用红色作为标记。

最后通过 show 方法展示图表。

3.2.11 模型保存与持久化

本次实验的数据以小时为单位，按照每天的顺序，分成多个文本文档文件，每个文本文档文件以文件流的形式导入。为了方便系统构建以及如果以后增加新的数据系统的可用性，我使用每次读取一组数据的文本文档的数据进行模型训练，对每次训练之后的模型进行保存。之后通过判断对应路径下是否有保存的模型，如果有，对之前的模型进行导入，重新训练，否则，重新初始化一个模型进行训练。这样可以在训练模型的同时，观察在控制台输出的回归系数评分，来判断模型设置的超参数和激活函数以及核函数等参数是否需要中断训练进行更改。

从模型持久化上，sklearn 为我们提供了相应的方法，包括 joblib 与 pickle 两种方式，将模型保存为.m 等格式的文件，使用 load 和 dump 等方法载入和保存。

第 4 章　系统测试与结果

4.1　系统测试方案及结果

本次实验测试所有五天的数据，通过测试试图寻找最优的超参数与模型，以及寻找数据之间的关系。同时验证系统对于此类文本文档文件的处理的方式的鲁棒性，以及系统的健壮性。前文通过图形和可决系数评分的方式已经展示了一些数据之间的关系。本章将对所有数据处理情况和模型做总结。

4.1.1 数据预处理结果

由于本次数据属于敏感数据，在仅对少量数据例之中进行分析可能存在问题。对多组数据分析的结果如下：

在位置数据中存在部分存在缺失值的问题，在这个问题上采取前文说到的 sklearn 中数据预处理部分的缺失值插补部分，在读入数据的过程中，我将空值数据设为 0，在对 Imputer 类进行实例化的时候，设置参数 missing_values = 0, strateg='mean',确保将 0 设置为缺失值代数，插补方式为使用特征平均值进行插补。最后对于位置数据的处理结果是位置经纬度以及高度的平均距离是 115.700530785， 20.8232339595， 32.2414707354。位置数据的极值分别为 32.2414707354， 115.700543333， 20.8232433333， 33.9， 115.700516667， 20.8232216667，30.7。

在对平台姿态相关的数据分析中，将角度相关数据与姿态相关数据成对分析，在本次实验中，由于不同组数据的采集时间有微小差异但是都是按照时间顺序排列的，而且实验为了保证所分析的数据来自于同一时间，需要对成对的数据取相同时间过滤，最终大概从以小时为单位的总数 3500+条数据的每份样本中，提取出 1500 条左右的合适数据用以进行分析，同时对通过文件流方式获取的数据进行转换，由字符串转换为浮点型。

经过处理后的角度相关数据的平均值为：-0.819892307692，-0.606758333333。极值为-1.798，-1.319，-0.031，-0.036。

经过处理后的姿态相关数据的平均值为：-2.62034045394，-3.97251001335，-0.823965287049。极值为：-4.02，-4.78，-2.4，-1.42，-3.48，0.4。

在经过对平台相关物理信息调研以及一部分数据模型的试验后，位置数据与时间之间的关系并不紧密明显。倾角和 姿态则从物理上可能具有一定关系。所以选择这对数据做为主要实验数据。

4.1.2 模型评估结果

前文提到，本次实验主要采取，简单线性回归，岭回归，支持向量机，以及人工神经网络进行训练，经过对可决系数评分的对比之后，我认为人工神经网络比较适合作为这对数据的模型，而且人工神经网络相对于来讲比较易于理解，优化的方法和方面比较多，有可能在我的模型上持续改进。

4.1.3 系统整体稳定性

本次实验的系统在训练模型的测试过程中没有任何问题，由于 svm 等模型以及对于数据预处理读取部分使用了较大内存，因此对内存溢出可能产生的问题做了测试，在测试过程中在 32 位 python 的内存空间下，对 4000 左右数据一次性进入内存处理也没有出现问题。

第 5 章　综述

本论文从多个角度阐明了基于半潜式钻井平台平台的数据分析系统设计与实现，我们总结了半潜式钻井平台数据可能存在的问题，并按照一般数据分析系统的设计与开发模式,对数据从读取到处理到结果输出设计出了一套完整的流程。对于最重要的的机器学习模型训练过程，分别使用线性回归，岭回归，神经网络模型，支持向量机模型等来进行训练，通过人工经验结合结果输出尝试获得几组数据之间的最优参数以及最优模型。通过程序完成了对初始数据的预处理，数据结构转换。模型训练及参数寻优，结果输出与可视化等等。

近年来，随着硬件设备的不断提升，以及互联网应用普及，大量数据可供分析训练，使大规模机器学习应用成为可能，但是机器学习，人工智能真正应用实践=尚短。数据挖掘与分析作为机器学习中的重要一环，在许多方面亟待开发。与海洋等相关学科的交叉研究更是发展的主要方向之一。但是在国内海洋科学与计算机科学尤其是人工智能相关领域碰撞出的火花较少，希望我的论文能为其他的有志于从事此类研究的同学同事提供一定的经验。同时本次实验由于数据均为敏感数据，在模型拟合度上有所欠缺，希望能有老师和同学提出宝贵的意见，是模型趋于实用化。

参考文献

[1] 朱丰,张群,冯有前,等. 基于压缩感知的逆合成孔径激光雷达鸟类目标识别方法[J]. 红外与激光工程，2013，42(1) : 256-261.

[2] 姜海娇，来建成，王春勇，等. 激光雷达的测距特性及其测距精度研究［J］. 中国激光，2011，38(5) : 0514001-0514007.

[3] GRUDNITSKI G, OSBURN L.Forecasting SPA and Gold Futures Prices:an Application of Neural Networks[J].Futures Markets，1993，13 (2) : 633 — 643.

[4] DEMATOS G. ，BOYD M S. ，KERMANSHAHI B. Feed Forward Versus Recurrent Neural Networks forForecasting Monthly Japanese Yen Exchange Rates ［J］ . Finance Engineer，1996，2 (1) : 59 — 75.

[5] SHAIKH A，HAMID Z I. Using Neural Networks for Forecasting Volatility of S&P 500 Index Futures Prices［J］ . Journal of Business Research，2004，11 (7) : 16 — 25.

[6] 曾星月.基于 BP 神经网络模型的玉米价格基差预测[J].粮食经济研究,2015,1(01):47-57.

[7] 马志强. 蛋白质功能预测的非同源性计算方法研究[D].吉林大学,2009.

[8] 熊国民. 基于 SVM 的商业银行客户流失预测[D].郑州大学,2014.

[9] 李强,杜煜.基于 3D 激光雷达道路边缘实时检测算法的研究与实现[J].计算机应用与软件,2017,34(10):219-222.

[10]胡沐晗.基于 PCA 和 SVM 的人脸识别系统[J].计算机时代,2017(12):60-63+67.

[11]王增茂,杜博,张良培,张乐飞.基于纹理特征和形态学特征融合的高光谱影像分类法[J].光子学报,2014,43(08):122-129.

[12]陈敬武,朱建伟,孙平昌.采用支持向量回归从测井曲线定量计算油页岩含油率[J].地质与资源,2017,26(02):157-160+183.

[13]胡林林. 基于数据挖掘技术的股价指数分析与预测研究[D].西南财经大学,2013.

[14]许长福. 日志数据分析系统的设计与实现[D].北京交通大学,2017.

[15]李德仁，张良培，夏桂松.遥感大数据自动分析与数据挖掘[J].测绘学报,2014,43(12):1211-1216.

致　　谢

在本课题的设计实验以及论文编写过程中，我的指导老师姜宇老师给予了我很大的帮助，包括从实验数据集，框架选型，以及模型初步选择上的建议和意见，提供了大量的论文资料和非计算机专业的知识指导，在完成实验与论文的总体方向上也给予了我很多时间经验。拓宽了我的知识和思路，也因此我能够按时且顺利的完成该课题的研究实验于论文辨析额工作，再次我对姜宇老师的知道表示由衷的感谢。

同时，我也感谢吉林大学软件学院四年来对我的悉心栽培，浓厚的学术氛围与活泼开放的校园气氛使我在这四年中受益匪浅，优秀的师资队伍给予了我们丰富的知识，大学给了我人生中最难以忘怀的一段经历。