**深 圳 大 学 实 验 报 告**

|  |
| --- |
| **课程名称 机器学习**  **项目名称 实验二：线性回归算法**  **学 院 计算机与软件学院**  **专 业 软件工程**  **指导教师 赖志辉**  **报 告 人 唐健龙 学号 2018132100**  **实验时间 2020年4月7日至2020年4月19日**  **实验报告提交时间 2020年4月19日** |

**教务处制**

# 一、实验目的与要求

**实现基本的线性回归算法，对一个简单的数据进行预测； 参考相关论文与文献，实现3-4个现有论文中的回归算法，并比较其在人脸识别中的性能**

# 二、实验内容与方法

**线性回归算法：**

**Ⅰ. 多元线性回归**

**原理：**

对一个多特征（即多维）的向量，通过对每个特征对应的数值进行加权，从而进行求和，得到一个标量，通常通过该标量能够进行预测或分类等操作

**推导：**

假设存在样本集合 ，包含 个样本，每个样本有 个维度，现在假设存在一个 维的列向量 ，我们能够通过

进行样本的某个指标或某个类别的预测，由于需要单独处理偏移量对计算容易造成不变，我们将样本集的每个样本的末尾添加常数1，即令

处理过后，我们只需要为 于末尾也添加一个维度用以作为偏移量权重，原本的线性预测公式就化为了

接下来的操作都将对增加了常数的样本集 操作

假设我们已有真实的预测值 ，则我们希望的结果是以下误差

能够尽可能小，上式对 求偏导，并令其为0，就有

于是就有

则求得的 就是满足我们要求的向量

容易知道，若预测值不再为标量，而是向量，那么真实值序列就会从向量 进化为矩阵 ，权重向量 也会进化为权重矩阵 ，上述公式则相应改为

**Ⅱ. Logistics回归**

**原理：**

在上述多元线性回归的基础上，有人提出，通过某种映射，将实数域的预测值映射至区间，通过最大似然法，进行梯度下降算法，不仅实现了回归，也能做到分类

**推导：**

首先介绍一个 sigmoid 函数

容易发现，它能够将实数域上的任何实数映射至 区间

通常，对于二分类问题，类标签的每个数字总可以定义为0或1，因此很适合使用sigmoid函数进行模型构建

为了使用梯度下降，首先需要先随机初始化权重向量 ，随后我们利用sigmoid函数进行映射，对于某个样本 ，其预测值为 的概率定义为

则其预测值为 的概率为

综合，就得到

构造对数似然函数，得到

将该似然函数对 求导，得到

上式即为 对 的梯度，记为

于是，通过

进行迭代（ 称为学习率，控制迭代速度和结果精度），直至预测误差小于某个给定误差，或者进行给定足够次数的迭代，就能得到我们需要的

**Ⅲ. 岭回归**

**原理：**

在上述多元线性回归的基础上，我们考虑到了一个问题，那就是 属于 的方阵，而当 时，我们将无法求得它的逆矩阵，也就是说，只要样本集的样本数 小于其维度数 ，，此时便不可逆，我们通过在优化模型中很巧妙地引入一个 正则化项，从而既避免了 的奇异性，又防止了过拟合的发生

**推导：**

将多元线性回归的优化模型添加一个 正则化项，我们得到

于是就有

**其中** 是一个 的单位矩阵，我们容易知道此时是能够求得 的逆矩阵的，于是我们通过增加正则化项，不仅将有可能不可进行的计算转为了能够进行的计算，还防止了过拟合 的发生

**Ⅳ. Lasso回归**

**原理**：类似岭回归，但我们考虑的是另一个正则化项，即 正则化项，该正则化项主要针对稀疏化，而不是最小化，它可以使得不必要的特征为0，从而达到防止过拟合的目的

**推导**：

将多元线性回归的优化模型添加一个 正则化项，我们得到优化模型

于是有

为了方便，我们记

于是，当 ，我们令偏导数为0，得到

由于需要使用迭代算法， 同样需要满足 ，于是知道

类似地，当 ，我们得到

而当

便可以直接令

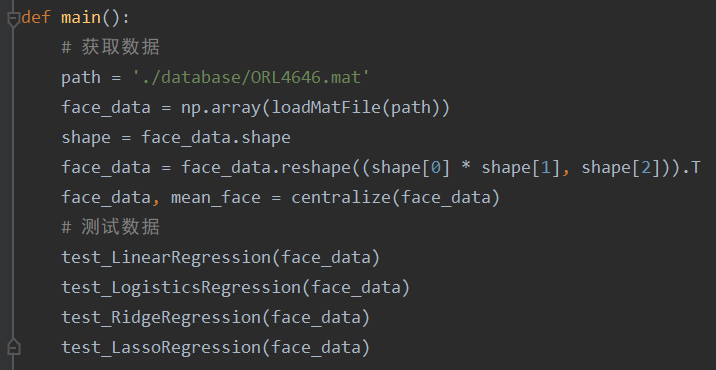
通过以上迭代算法，在实验中，通常较快就能够收敛，设定一个阈值，当收敛误差小于阈值即可停止，对于不同的数据集，我们需要进行人工调参使得正则化项系数能够达到较为理想的效果

# 三、实验步骤与过程

**回归算法**

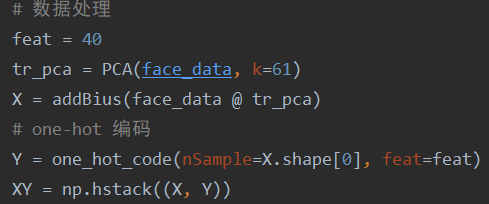
**1. 数据准备**

加载ORL人脸数据，共400个样本，每个样本为46×46的图片，每10张为同一个人的人脸图片，将每个样本拉成行向量，接着中心化，得到新的数据集，随后对上述涉及的多个回归算法进行识别率的测试

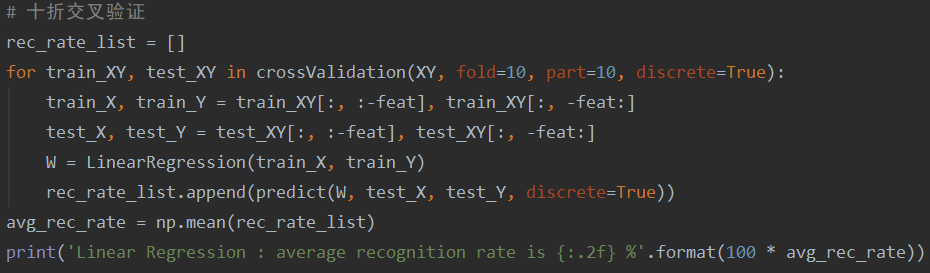


**2. 多元线性回归**

使用one-hot编码生成标签，以及给样本追加常量1，同时，为了可逆运算能够进行，将数据维度降低到400维以下，经过实验，在降低到61维时，能达到较好的效果



通过十折交叉验证，使用训练集得到权重矩阵 ，使用其计算测试集的识别率，测试识别率的方式是，计算one-hot标签的最大值索引是否与真实标签相同

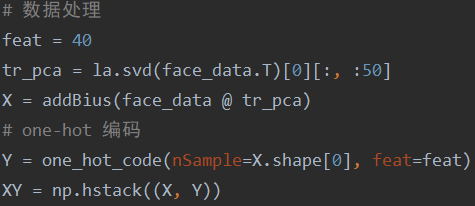


多元线性回归代码如下，直接通过推导的公式得到权重矩阵

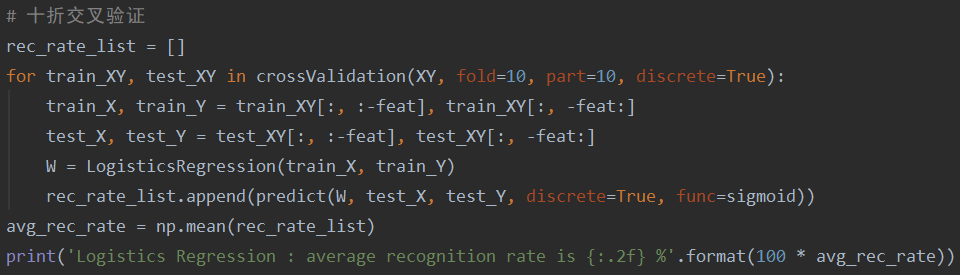


**3. Logistics回归**

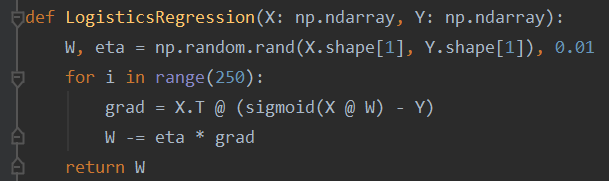
此处回归算法有一个奇怪的点，至今我也没能解决，即在将数据中心化后无论进不进行PCA降维，该回归算法始终都不收敛，于是我直接进行svd分解，将数据变换后低维，再进行人脸识别，首先为样本追加常量1，将数据维度降低到50维，使用one-hot编码生成标签



通过十折交叉验证，使用训练集得到权重矩阵 ，使用其计算测试集的识别率，测试识别率的方式是，计算one-hot标签的最大值索引是否与真实标签相同

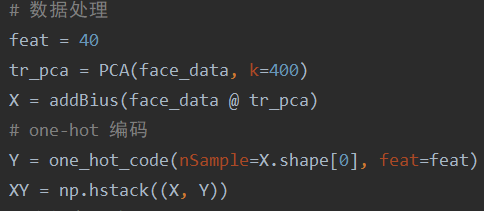


Logistics回归代码如下，直接通过推导的公式进行给定次数的迭代得到权重矩阵

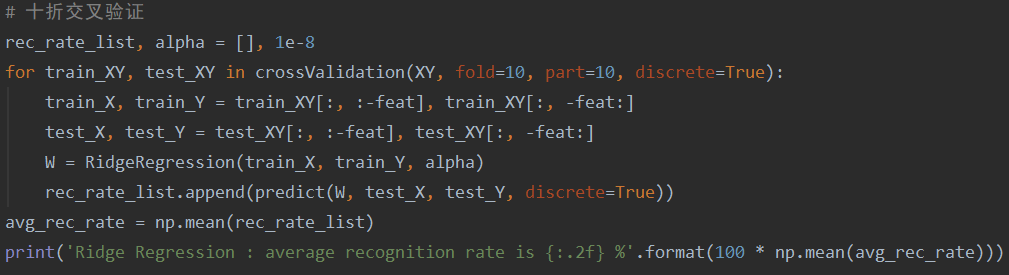


**4. 岭回归**

由于已经可逆，只进行PCA处理至400维（只是为了计算效率以及初步的特征提取，若降低至61维，识别率将与多元线性回归极为相近），随后使用one-hot编码生成标签，以及给样本追加常量1，经过实验，将岭回归正则化参数alpha调整至1e-8能够达到较好效果



通过十折交叉验证，使用训练集得到权重矩阵 ，使用其计算测试集的识别率（需要注意的是，岭回归得到的识别率与多元线性回归不具有可比性，多元线性回归通过PCA进行了较充分的特征提取，而岭回归是多元线性回归的不可逆情况下的补充，测试中并不真正进行充分特征提取，只是为了展现其对多元线性回归的扩展） ，测试识别率的过程依然是利用训练集求得的矩阵在测试集上进行测试

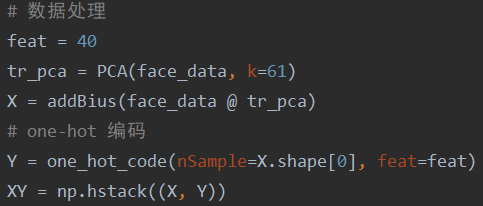


岭回归代码如下，也是直接通过推导的公式计算权重矩阵

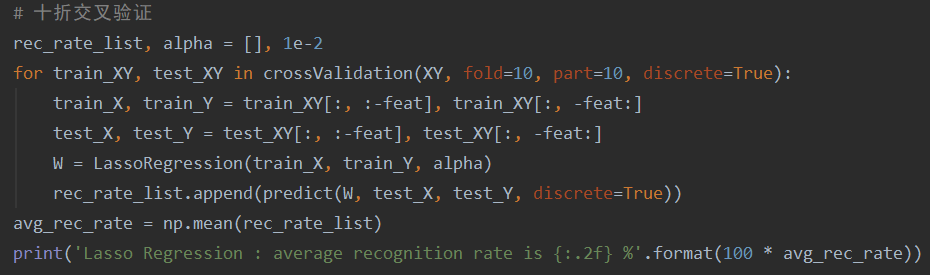


**5. Lasso回归**

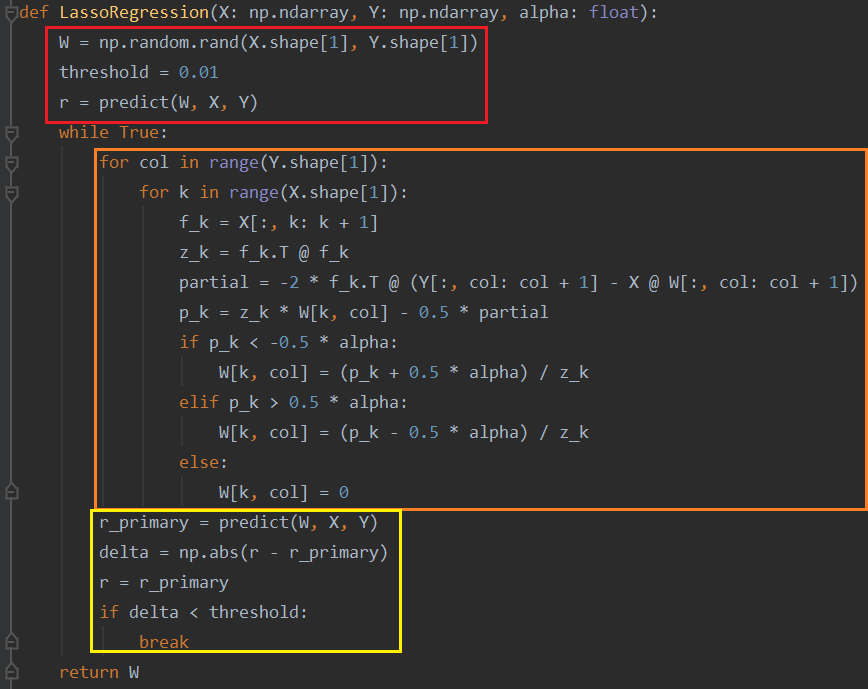
将数据先降维至61维（有一个很奇怪的点是处在一定维度下收敛速度非常慢，而在另一些维度收敛会很快），随后设定阈值为0.01，设置Lasso回归正则化参数为0.001（该正则化参数在较大范围内都能达到一定的识别率，不如岭回归正则化参数敏感）



进行十折交叉验证，利用训练集得到权重矩阵 ，使用其计算测试集的识别率



Lasso回归代码如下，相比前面的回归，Lasso回归的代码更为复杂，图中，红框进行权重矩阵、阈值、和初始误差的初始化，橙框对应推导中的迭代过程的计算，黄框判断是否满足退出迭代的条件



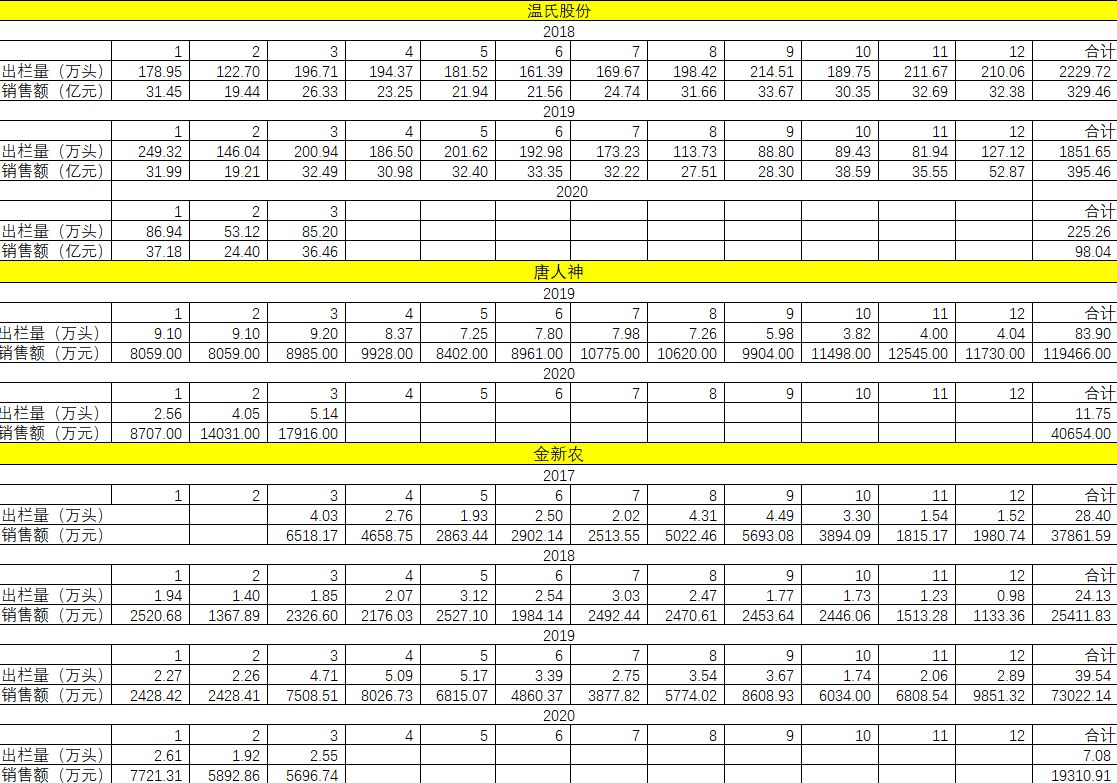
**6. 识别率比较**

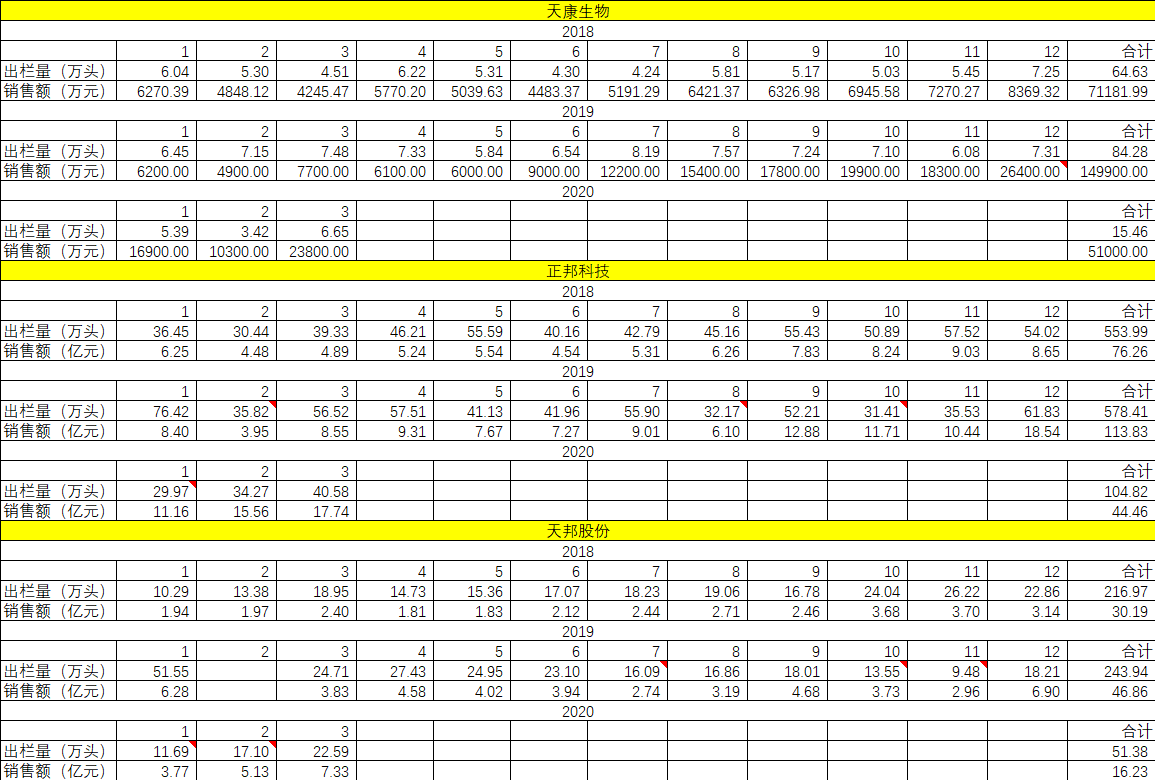
下面给出四种回归对于ORL人脸库的识别率

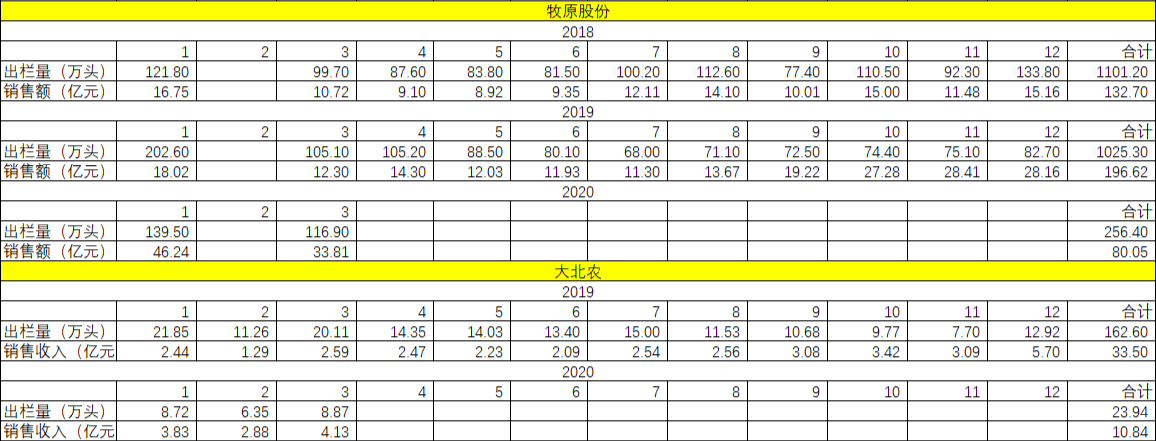
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Linear Regression | Logistics Regression | Ridge Regression | Lasso Regression |
| Recognition rate | 97.25% | 83.75 % | 94.50% | 97.25% |

**股票预测**

**1. 公司数据**



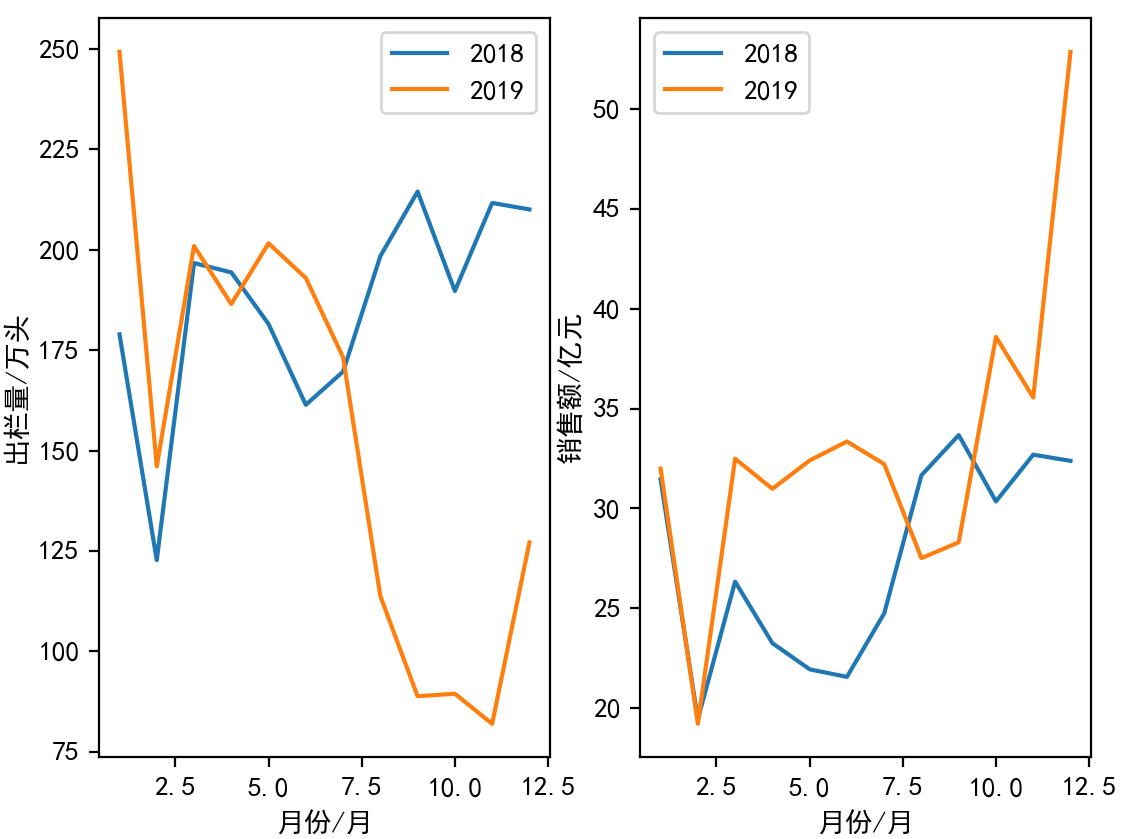




**2. 市场分析**

以温氏股份为例，按照往年数据，春秋两季的生猪出栏量相对全年较高；在正月，人们已经囤积足够粮食，肉类销量并不高，但是在这之前的几个月，肉类销量是相对较高的

受非洲猪瘟影响，2019年末生猪出栏量减少，且在2020年初，新冠肺炎导致的人们过年的氛围降低，会导致人们对肉类的需求进一步降低，但是当国内疫情基本结束后，肉类市场将会出现报复性增长，可以预见的是，各公司将会加大存栏量，且采取薄利多销的策略刺激国内肉类市场的消费



**3. 股票预测**

说明：以下股票信息均出自同花顺APP



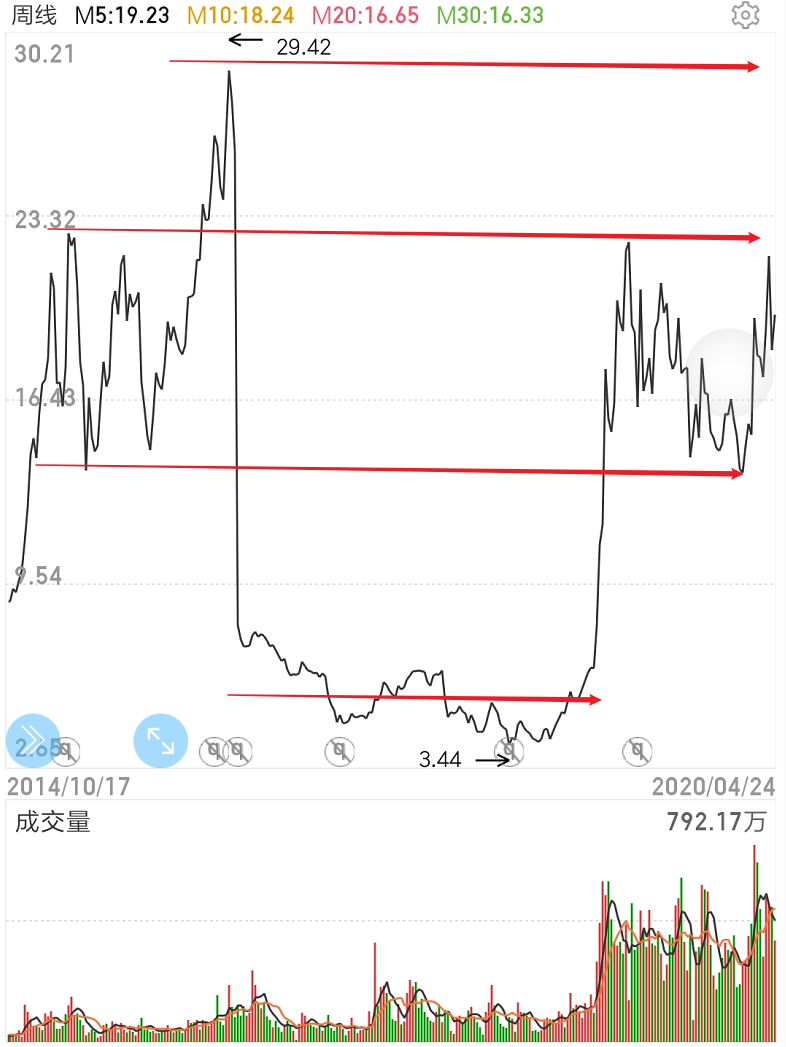
**【牧原股份】**

利用周K分析法，首先注意股票首次大幅拉升距离股票首次进入股市时间约为5年，主力在4-5年内吸筹，随后将大概率在随后1年进行三次以上的大幅度拉升，到目前为止(2020/05)似乎已经完成了主要的拉升操作，从最新股价发现已经有触及压力面的信号，从换手率也能发现成交量近期开始增加，接下来的三年可能都将处于停滞不前和熊市阶段，不建议入手



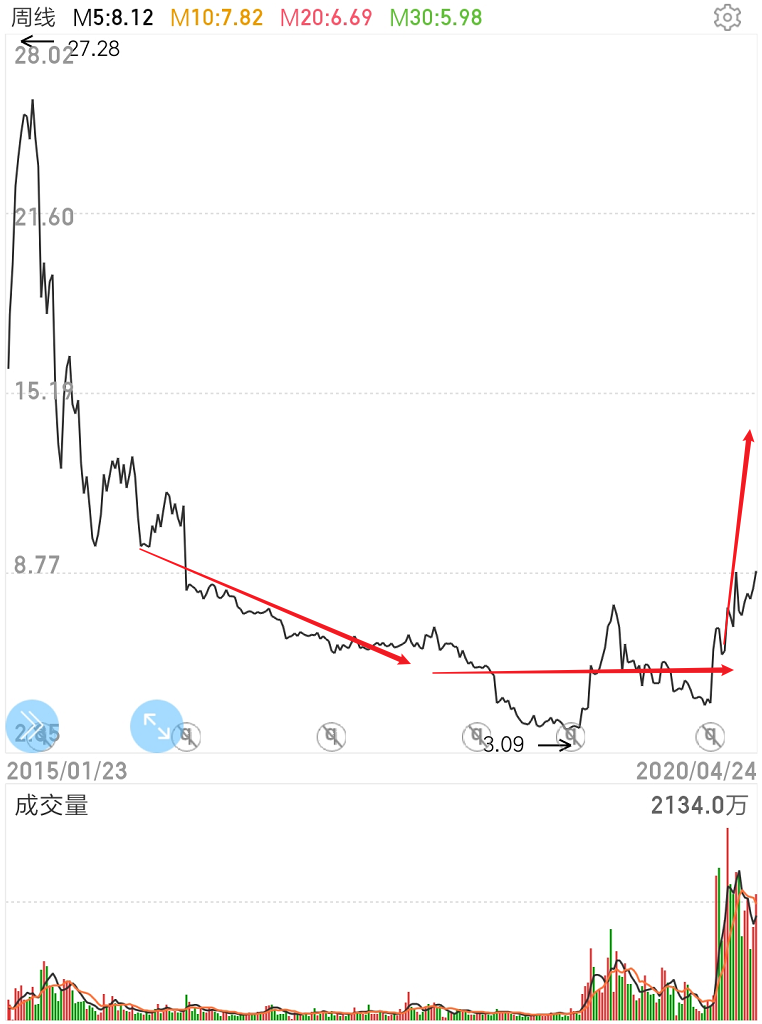
**【温氏股份】**

很显然，该股刚刚经历主力的大手笔出货阶段，现在主要为散户抛售股票，开始了熊市的阶段，三年内都不建议入手



**【正邦科技】**

可以发现，该股在横盘后期的突然拉升非常瞩目，可能是主力想吸引人们关注该股，随后在拉升一定量后，触及历史高位，主力开始大量抛售股票出货，从成交量也能看出主力意图，随后该股出现了向下的趋势，但是在触及历史低位后再次反弹，可以认为这是较强的支撑面，但是该股后期局势不明，若突破了历史高位，可以考虑短期入手，但是目前建议持观望态度，并且即使该股能够升高，也不会有较大涨幅



**【大北农】**

该股处于低位的时间较长，在横盘后期出现了振盘，随后股价开始突破历史高位，对应的成交量也有所体现，可以预见的是，该股将大概率会开始牛市阶段，一年内能够赚钱，并且可能突破历史高位，随后两年内将出现震荡，然后进入熊市

预测：

|  |  |  |  |
| --- | --- | --- | --- |
| 时间 | 股价 | 时间 | 股价 |
| 2020/06/30 | 15.00 | 2020/12/31 | 30.00 |
| 2021/12/31 | 25.00 | 2022/12/31 | 8.00 |



【天康生物】

该股已经处于历史高位，在未确定突破之前，不能贸然入手，况且前期没有横盘，主力的力量不足，长期持有也并不妥，因此不太建议入手



**【天邦股份】**

该股最近几个月见涨，但是后期又大幅度下跌，局势不明朗，可以继续观望，但是有可能出现缓慢的牛市，涨幅不见得很大，但是震荡可能比较厉害，风险较大，不建议入手



**【罗牛山】**

该股的情况发展较慢，前些年总体缓慢升高，但是在触及历史高位后，可能受到压力面的作用，不能进一步提升，随后再次下跌，往后几年局势不明朗，不建议入手



**【傲农生物】**

该股目前股价已经接近历史高位，并且目前处在均线上方，前期也出现过大提升和横盘，随后下降，此时可能仍然处于熊市，不建议入手



**【唐人神】**

该股目前处于熊市中期，即使后期熊市结束也会有长时间横盘，不建议入手



**【上海梅林】**

该股目前似乎仍然处于洗盘的中后期，即使洗盘结束了，主力吸筹期间可能进行各种洗盘等操作，需要等待开始拉升至突破历史高位再考虑入手，因此目前也不建议入手



**【新五丰】**

该股目前处于短期下跌阶段，且之前处于历史高位，现在入手是不明智的选择



**【金新农】**

该股目前似乎也属于下跌横盘的阶段，不建议入手

**4. 预测总结**

短期我推荐购买大北农，4月17日交易日价格是8.46元，它到6月30日/7月6日（最后一次课下午收市价为准）的价格是15元，涨幅为77.3％。预测人姓名：唐健龙

长期（2020、2021、2022年末）我推荐购买大北农，4月17日交易日价格是8.46元，它到（2020、2021、2022年末12月30日）的预估价格分别是（30、25、8元），涨（跌）幅分别为（254.6％、195.5％、-5.4%）预测人姓名：唐健龙

# 四、实验结论或体会

**实验结论**：线性回归能够较好地用于线性分布数据的预测，回归可能发生过拟合从而无法拥有较好的泛化能力，需要适当进行正则化处理；可以融合多种机器学习技术如感知机等进行回归

**实验体会**：线性回归虽然能够预测一些线性分布的数据，但是也需要配合基础的PCA等降维算法降低计算的复杂度以及进行特征提取，才能显著发挥其作用；另外还有一点值得注意的是，传统的线性回归算法，是无法处理非线性分布的问题的，不过如今已经有很多研究指出，我们可以借助各种Kernel trick将非线性问题转化为线性问题，并且其在实验表现也相当不错

|  |
| --- |
| 指导教师批阅意见：  成绩评定：  指导教师签字：  年 月 日 |
| 备注： |

注：1、报告内的项目或内容设置，可根据实际情况加以调整和补充。