**目 录**

1. 绪论
2. 正文
3. 结论

**绪 论**

本身就是对数学算法这类很兴趣，选择了数智工作室作为学习进步的目标，努力完成考核任务，其实也不单单为了有机会加入，更多的是也希望在这个过程中学到更多的知识，掌握更多专业技能，学到更多有趣的算法。选择了难度一，是利用线性回归算法实现对测试集的数据预测。

。

**正 文**

**回归**

1. 线性回归

目标是完成线性回归模型的构建，通过训练模型得到预测值。线性回归，是指用线性函数拟合已知的数据，预测未知数据；解决的是回归问题，属于监督学习。

由训练集可见，有十三个特征，则采用的是多元线性回归模型。首先假设函数为

在该训练集中n为十三。函数的参数为：维度为n+1维。因为相当于偏置项，与之相乘的为一，所以两者相乘结果就是偏置项的结果，让模型精确度更高，拟合效果更好。有了假设函数可以求出代价函数为

代价函数采用了平方差的方法，代价函数作为一项指标，可以判断模型的拟合成程度的好坏，损失值越低就代表模型训练的越好，所以线性回归的目的就是要通过一次一次的迭代实现损失值的下降以实现模型的训练。而损失值得下降，就需要假设函数的结果与训练集中的结果拟合效果更好，就要寻找参数的最佳值。这一过程就需要利用梯度下降。

二、梯度下降

梯度下降的主要目的是通过迭代找到目标函数的最小值，或者收敛到最小值；在线性回归模型中就是希望找到一组theta使损失值能降到最低。首先梯度下降要找到函数降低最快的方向，就要对损失函数求偏导，以找到方向：

通过一次次更新theta的值，是函数损失值降低，让模型的训练效果更好。这其中的α代表的是学习率，他是相当于下山的步长，学习率越低，每次下降的程度越低；学习率高，每一次下降的程度也高，步子更大。合适的步长也很重要，如果学习率太大，可能就会错过最优点，无法收敛；但是学习率过小就会让迭代次数增加，收敛速度过慢。

1. 正规方程

用梯度下降方法，需要经过多次的迭代才能找到theta的最优值，但是利用正规方程，就能够实现直接求出最优的theta。求解最优解θ，我们只需要对J(θ)函数对每个θ求偏导，求出每个θ 使得偏导等于0，这些解组合成对应的向量θ就是最优解。

要找到损失函数的最低点，在最低点时，导数为零；通过梯度下降是每一次迭代找到斜率降低的方向一步一步接近；而正规方程则是，令每一个偏导数都为零，求出其对应的theta的值，将其组合一起就得到了需要的theta的最优值。

此时的代价函数为

对代价函数的每一个theta求偏导，得到正规方程的表达式为：

移项整理得特征参数表达式：

对比：如果特征量较少，用正规方程比较多，误差比较小，计算量不大；但是特征变量多的时候，还是用梯度下降准确率要高。

1. 数据处理

1、对缺失值的处理

对于训练集中出现的nan值，我最初选用的是将其删除，但是经过模型的测试之后，由于样本值不够多，将nan值删除之后，对模型的精确度影响较大，所以又将nan值用该特征的平均值代替。先计算该特征的均值，然后找到nan值的所在位置的索引，然后用均值代nan

2、min-max归一化（离差标准化）

将数据集里的每一个特征的一组数据都遍历一遍，将max和min记录下来，并将其作为基数；让每一个样本数据与min的差与max与min的差相比，

最后的结果是让数据映射到0-1之间。对于同一个数据集的特征之间相差太大，可以通过归一化处理数据，提升模型的收敛速度，提升精度。

1. 多项式

当训练集中含有特征较多，一次项在拟合过程中表现的并不是很好，所以增添二次项的部分，增加权重的数目，构成多项式回归。

原预测函数：

New函数：

多项式回归通过将决定拟合程度的参数theta的数量增加了一倍，而且这时候模型不再只能满足简单的线性关系，而能适应更多的关系。

1. 正则化

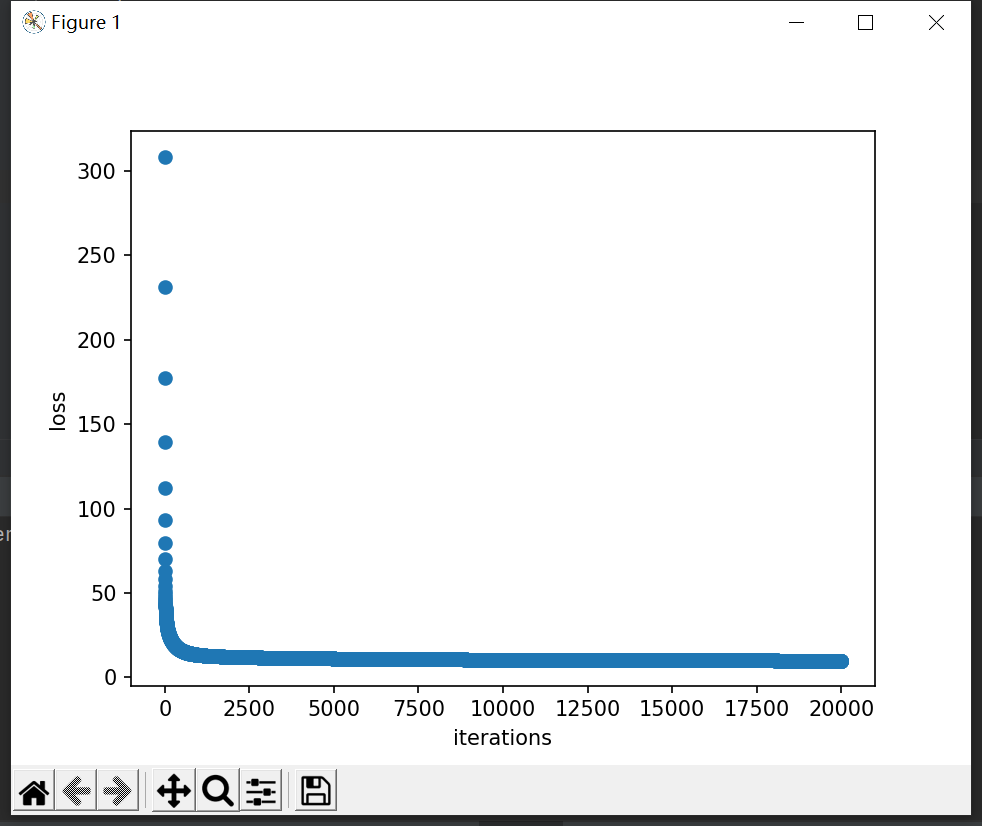
当采用了多项式回归之后，可能会出现过拟合的情况，这时候就要用正则化来解决这个问题。正则化相当于改变权重的重要性原，降低拟合程度，使模型的泛化程度更好

原损失函数：

增加正则化项后

对权重的求取变成：

化简得：

其中的代表的是正则化参数，用来控制两个不同目标之间的取舍；新添加的一项为正则化项，其作用是缩小每一个参数。回归系数反应对应自变量的重要程度，正则化的作用就是减小不太重要的因素参与计算的权重，是对权重的惩罚，权重越大，惩罚也越严重。正则化参数就需要保证既不会失去权重，结果又能够被惩罚到。

1. 尝试、问题与解决

损失值很大

1、为了降低损失值，进行了归一化处理

1. 注意只需要将x值归一，因为通过模型要找到的是因变量与自变量之间的关系，只要将测试集和训练集的x值同时归一两者之间的关系本质上是没有发生改变的。
2. 因为要有偏置项的存在，所以在原数据集内添加新的一列“特征”x(0)等于一，所以要先将原数据集内的数值进行归一，再加上新的一列x(0),才能保证偏置项的正确。
3. 训练集的x也需要归一处理

2、尝试四分位距法

四分位距法，即将数据从小到大排列，只取中间的50%。也就是25%到75%段的数据，可以有效剔除异常值(极大/极小值)。同一特征内数值差距太大，异常值太多，对参数影响较大，需要去除异常值。尝试将25%和75%作为划分界限，但是发现损失值更高了，并且随着选定范围的增大，损失值在逐渐下降，结论就是，四分位距的使用不但没有提高精度，反而让拟合效果变得更差。这个问题还没解决！

问题

现在模型的梯度已经消失，但是损失函数还是很大；同时利用梯度下降和正规方程求取权重，结果损失值都是无法很好的下降，添加了多项式之后拟合效果确实更好了一点，但是kaggle的分数就变得更低了；怀疑是出现了过拟合现象使得他在测试集中无法很好的泛化，但是添加了正则化项之后效果还是不是

很理想。

1. 小结

虽然实现的比较低难度的模型，这个过程也是很曲折，常常会碰到各种各样的问题困难，但是解决他们的过程还是很爽的！

在这个过程中认识到，对回归问题有了更深的了解认识，同时对分类，神经网络等也有了一个比较浅显的认识，也产生了浓厚的兴趣，希望能在接下来的时间里学到更多的东西！