# **三周机器学习总结**

## 郑曼娜

引言

从3.13.到现在半个多月的时间里，我就“线性回归”与“softmax”这两个模型的搭建进行了我对机器学习的探索。

## 1 我的大致学习路线

**1.1什么是机器学习以及一些基本术语**

·机器学习是一门能够让编程计算机从数据中学习的计算机科学。机器学习研究如何让计算机不需要明确的程序也能具备学习能力。研究关于“学习算法”的学问。

·要进行机器学习与模型的搭建，首先要有数据，从数据中学得模型的过程叫做“学习”或“训练”。

·在训练之前，要先对获取的数据集进行划分，分为“训练集”与“测试集”。

**1.2基础工具、基础库的使用**

在第一周的学习时间里，我安装了Anaconda，开始使用Jupyter Notebook。了解了Numpy，pandas库的一些函数，我通过边看视频边做笔记的方式学习这些函数，后来得知，其实并不需要学完很多函数就可以开始学习做模型了，可以在做模型的过程中需要用到什么函数的时候再去查询资料。

**1.3数学知识的补充**

在学习线性回归模型的数学原理的时候，我意识到有一些线性代数的知识是需要掌握而我还没有掌握的，于是在第一周的时间里我也补充学习了一些线代的知识：行列式，向量，矩阵。还有一些高等数学的知识，比如求偏导，这个在我学完了之后课内高数的课程才开始教。

**1.4模型的搭建**

大概到了第一周的周末，我才开始了对线性回归模型的实现，之后又实现了softmax模型。

**1.5模型的评估**

在最后一周完成两个模型的搭建后，我学习了一点点关于模型评估的知识，选了两种评估的方法分别对我的两个模型进行评估，可是其中一个方法我没来得及实现代码。

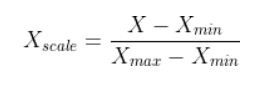
## 2 数据预处理

我用numpy与pandas完成了这几种数据预处理的方法：

**·归一化和标准化**

1. **归一化**：归一化的目的就是使得预处理的数据被限定在一定的范围内（比如[0,1]或者[-1,1]），从而消除奇异样本数据导致的不良影响。用于我的softmax模型的数据预处理。

公式：



1. **标准化**：最大最小标准化，将数据变换为均值为0，标准差为1的分布。用于我的线性回归模型的预处理。

公式：

**·独热编码**：用于处理softmax模型的标签

**·缺失值的处理**：在训练线性回归模型时，训练集中含有许多缺失值（测试集也有），我用均值来代替这些缺失值。

**·异常值的检测和处理**

·**分割数据集：**训练集中包含输入值与输出值，用numpy与pandas中简单的函数便可将它们划分开，以及实现偏置项的加入。

## 3 线性回归

这是我第一个实现的模型，也是我花了最多时间的模型，我使用了梯度下降法来实现。

**3.1简介**

·线性模型就是对输入特征加权求和，再加上一个偏置项（也称为截距项 ）的常数，以此进行预测。使用梯度下降法可以调整模型参数直至训练集上的成本函数调至最低。

·多元线性回归：用多个特征量或变量来预测输出值。

**3.2数学原理**

**·这项任务的目的是：**给定一个数据集（称作训练集）D={(x1,y1),(x2,y2),(x3,y3)…,(xn,yn)},要学得一个线性模型，以尽可能准确地预测另一个数据集（称作测试集）D’={x1,x2,x3,…,xn}的输出。

**·对公式的理解：**





注：θT表示θ的转置向量（列向量）

X为样本的特征向量

y＾为预测值

hθ为假设函数

J (θ)为成本函数

α为学习率

·其中J(θ)为衡量预测结果与正确值的差距的函数，J(θ)越小，结果越准确，为了求出使J(θ)尽可能小的θ，要对最后两个式子进行迭代，更新θ的值，使J(θ)不断趋近于最小，使得用θ预测的输出值逼近正确值。

**3.3代码实现**

对数学原理理解完毕之后，就开始我的代码实现了，首先对数据进行预处理，对于训练集中含有缺失值，我采用以均值代替缺失值的方法。然后我就直接开始写梯度下降的算法了，然后就失败了，我算出来的参数全是“nan”，我逐步检查我的代码，确认了输入值没有“nan”，确认了梯度下降公式的代码没有出错，也多次调整了学习率和迭代次数，但是还是无法得出正常得结果，甚至出现了溢出，这时我才后知后觉地意识到忽略了数据预处理中一些重要的步骤——比如归一化、标准化，在用了标准化来处理我的训练集后，再次用梯度下降开始训练，就得出了不再是“nan”的参数值，然后用这些参数值去预测出了一份数据，提交了，得分是0.37，我再一次感到困惑了，反复调参、预测、上传，还是不能得到正常的预测结果，我觉得是我在数据预处理方面仍有遗漏，于是反复搜索数据预处理的文章和书籍资料，可是却得不到解决的方法，最后发现其实问题是出在没有对测试集的数据也进行归一化或标准化，因为我得出的参数是基于标准化来训练的，用这些参数来预测时，自然要对测试集也进行标准化，这其实是一个简单的问题，我却一直没发现它，从这里我也得到了启发，不懂的东西还是应该多问，完全靠自己盲目地找问题有时候既浪费了很多时间又无法得出正确的结果。

在对测试集也完成了标准化后，我终于输出了正常的结果，上传预测后成功上榜了。

**3.4评估模型**

我选择了自助法（自助采样法）和R-Square对我的线性回归模型进行评估，从训练集中取了1/5的样本作为验证集。

## 4 softmax

**4.1简介**

·softmax算法是逻辑回归算法在多分类问题上的推广，主要处理多分类问题。我使用了梯度下降法来实现它。

·相比于第一个线性回归模型，我在学习这个模型的时候，更加的有方向了，知道该从哪些方面去入手了，但是相比于逻辑回归，这个模型的公式又复杂了很多，所以我还是花了比较久的时间去实现这个模型，并且完成效果也很一般，比如没有实现正则化，没有实现模型的评估。

**4.2数学原理**

·这项任务的目的是：对于一个给定的实例x，softmax模型首先计算出每个类别k的分数Sk（x），然后对这些分数应用softmax函数，算出每个类别的概率。再将概率值最高的类别作为预测出的结果（即标签）。

·对公式的理解



注：K是类别的数量

S（x）是实例x每个类别的分数的向量

σ（s（x））k是给定的类别分数下，实例x属于类别k的概率

y^即为预测值

J（Θ）为代价函数

J（Θ）为梯度函数

同样，要对最后两个式子进行迭代，更新θ的值。

**4.3代码实现**

数据预处理采用的是归一化，在补充学习完独热编码后，梯度下降算法的代码实现起来就比较顺利了，但是当我将预测出的数据上传之后，我发现可以用正则化来试试提高准确率，但是由于我没有空了就没有去实现了。

我想用macro-F1的方法来实现对这个模型的评估，但是我没有代码实现：分别计算出每个类别的“宏查准率”（macro-P）、“宏查全率”（macro-R），再根据公式算出相应的“宏F1”（macro-F1）。

## 5 模型评估

在验证集上对模型的泛化误差进行评估，这对模型的选择有较大意义（我猜想的）。用于评估测试的数据集称为“验证集”。

2.4.1评估模型的方法（这里指把训练集另外划分为训练集和验证集的方法）

·留出法

·交叉验证法

·自助法

我用自助法从训练集中抽取了1/5的样本作为验证集对我的模型进行评估，完成模型的评估后，进行调参，再重新于训练集（包含后来分出的验证集）中训练模型，便得到我们的最终模型。

2.4.2模型的性能度量方法

·回归：MSE、MRE、R-Square等等

·分类：Accuracy、Precision、Recall、F1 Score等等

我选择了R-Square对我的线性回归模型进行评估，macro-F1对我的softmax模型进行评估，但是很遗憾后者没能实现：

1. R-Square

公式：R-Square=1-

1. Macro-F1

精确率：P=

召回率：R=

Macro-F1=

其中macro-P为各类别的P加起来的均值，macro-R为各类别的R加起来的均值

## 6 总结与反思

这是我在图书馆待的日子最多的三个星期了，我每天除了上课写作业，就是面对吴恩达，面对《机器学习实战》，面对CSDN，虽然我学到的东西比较少，人也很迟钝，但总的来说我还是投入了挺多时间的，也挺开心自己能够坚持下来学的，虽然在面对难题时有过想偷懒不想学了的心理，但是想到我已经为这个考核投入一些时间、精力、情绪了，就舍不得放弃了呜呜。意外的收获是，这项任务非但没有打乱我的课内学习安排，反而让我每一天的安排都变得井井有条起来，我认真规划起我一天要干什么，我要在哪个时间前赶紧写完作业然后来进行我的机器学习，现在一起床就知道自己该干什么了，拖延症拜拜。

反思就是：我应该学会多主动地问问题，面对完全新的知识自己盲目地找问题太浪费时间了，又解决不了又耽误进程，有不懂的地方还是应该抓紧时间问，要有时间观念。其实三周的时间过去，我并没有达到自己理想的目标——我没有做完那个Kmeans，甚至很匆忙地在赶任务，比如现在已经周日下午了，我还在赶论文，我希望自己在二轮地考核当中能够更有时间观念一点，争取多完成一点任务，不再被ddl追着跑。

## 7 参考文献

[1]《机器学习实战》Aurelien Geron（机械工业出版社）

[2]《机器学习》周志华（清华大学出版社）