**数智NLP组2022第一阶段考核**

**信息工程学院**

**电子信息类**

**2021级9班**

**3121002477**

**黄维爽**

**2021年3月**

**摘要**：在这个文档里，我将要简要概述第一阶段的考核任务完成情况并进行汇报和总结，任务包括机器学习的线性回归模型、softmax回归模型、k-means算法、以及数据预处理，特征工程，结果可视化，模型评估等等内容,以及利用爬虫爬取豆瓣top250的电影的相关信息的实现。

关键词：机器学习，爬虫

**Abstract**: In this paper, I will give a brief overview of the completion of the assessment tasks in the first stage and report and summarize them. The tasks include linear regression model, softmax regression model, k-means algorithm, as well as data preprocessing, feature engineering, Results visualization, model evaluation, etc., as well as the realization of relevant information of Douban top250 movies using crawler.

**Keywords**: Machine Learning, Crawler.

1. **项目实现过程**

**1.1 前期准备工作**

* 软件的安装与使用：jupyter notebook、pycharm
* 强化数学基础（高数，线代，概率论，统计方法学习），懂得线性回归、softmax、k-means的数学底层逻辑
* 学习python编程语言，掌握三大科学计算库的使用，学会用代码实现算法
* 明确机器学习工作流程：获取数据（现有）数据预处理，特征工程，算法代码实现，训练模型与优化，模型检验（可视化与评估）

**1.2 中期**

* 数据预处理：处理缺失值、处理异常值
* 代码的编写：根据学习过的机器学习算法内容，用python实现，注意代码的命名规则与语法（让代码具有可读性与易懂性）
* 查找资料，咨询师兄，增加知识来源渠道（当我没有思路的时候，这些是好方法）

**1.3 后期收尾工作**

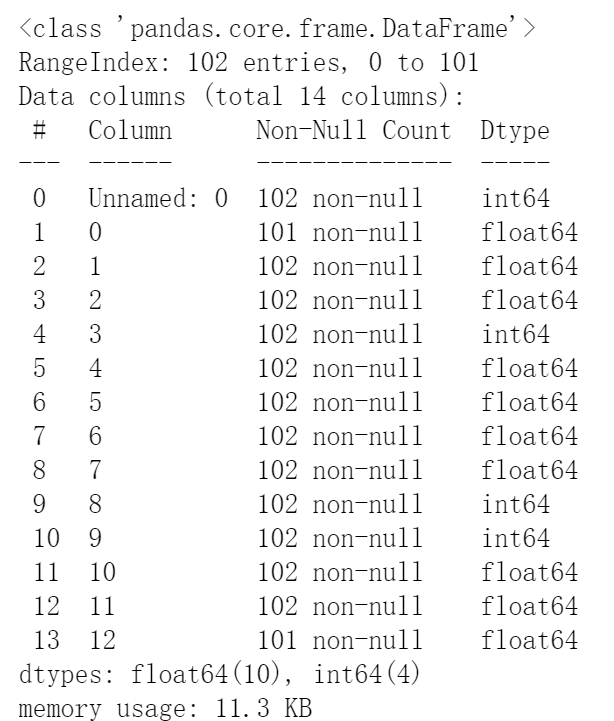
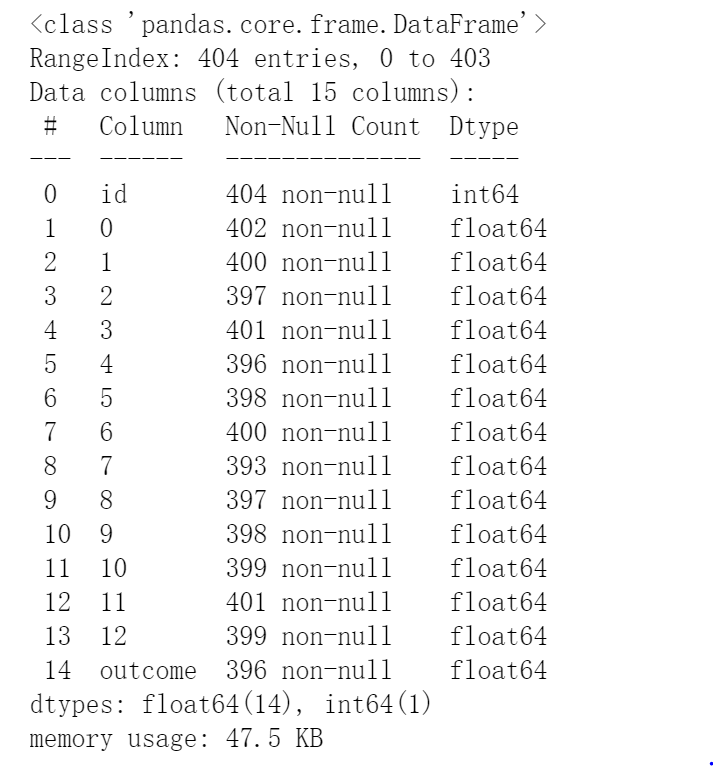
* 模型的优化与调参（如迭代次数，学习率等等）
* 优化数据预处理的代码
* 交流与学习
* 总结出现的问题

**2 考核任务描述**

不用框架实现线性回归模型（使用最小二乘法，或者梯度下降法，或者牛顿法），不用框架实现sotfmax回归模型（多元分类），提供K-means算法与决策树（CART），使用爬虫爬取豆瓣电影top250相关信息，学习以下几种数据处理方法：归一化和标准化，独热编码，连续值离散化处理，缺失值的处理，异常值的检测和处理，分割数据集，使用matplotlib实现可视化。

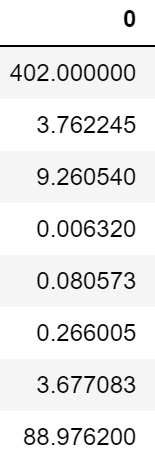
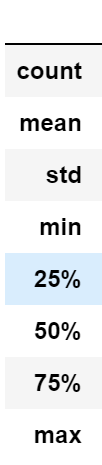
1. **线性回归模型的实现**
   1. 数据预处理

先导入数据集，查看数据集的基本信息，可以发现这两个数据集存在缺失值。



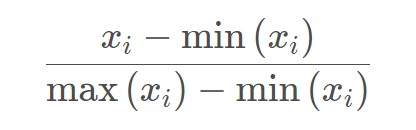
训练集存在outcome值缺失的行要删掉，不能填充，缺失值都要用训练集的均值填充（实际情况测试集未知，不能用测试集均值填充测试集的缺失值）。

训练集存在异常值（最小值，最大值与均值相差较大）。

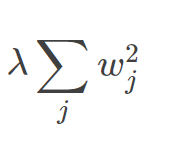


处理异常值使用的是3σ原则，使样本满足正态分布（σ为标准差，μ为均值），即样本数值几乎全部集中于（μ-3σ，μ+3σ）区间上，把不满足条件的样本（这些样本称为粗大误差）剔除。

将训练集与测试集进行线性归一化，统一量纲，数据归一化后，最优解的寻优过程明显会变得平缓，更容易正确的收敛到最优解。归一化将数据集中某一列特征的值缩放到0和1之间。缩放方法如下所示：



L2正则化：梯度和损失函数加入L2范数，防止过拟合。L2范数：

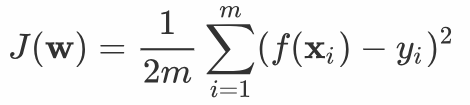


* 1. 算法的实现

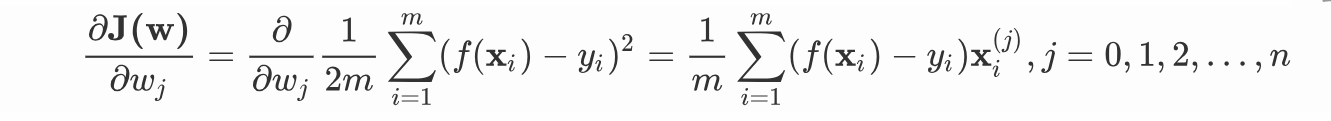
梯度下降法：

梯度下降法是一种常用的一阶优化方法，是求解无约束优化问题的经典方法之一。对于连续可微函数上某一点，有各个方向导数，沿梯度方向的方向导数达到最大值，也就是说，梯度的方向是函数在这点增长最快的方向。函数在某点的梯度是这样一个向量，它的方向与取得最大方向导数的方向一致，而它的模为方向导数的最大值。所以我们可以沿反梯度方向不断一步一步迭代，得到局部极小点。当目标函数为凸函数时，局部极小点就是全局最小点，此时梯度下降法可确保收敛到全局最优解。

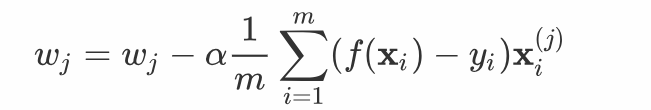
将损失函数（均方误差）



对列向量求导，得到的偏导：

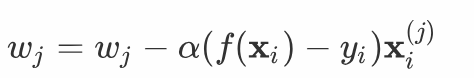


然后对各个分量都以下面形式更新（初始化权重：随机初始化，使得算法更容易找到全局最优解）：

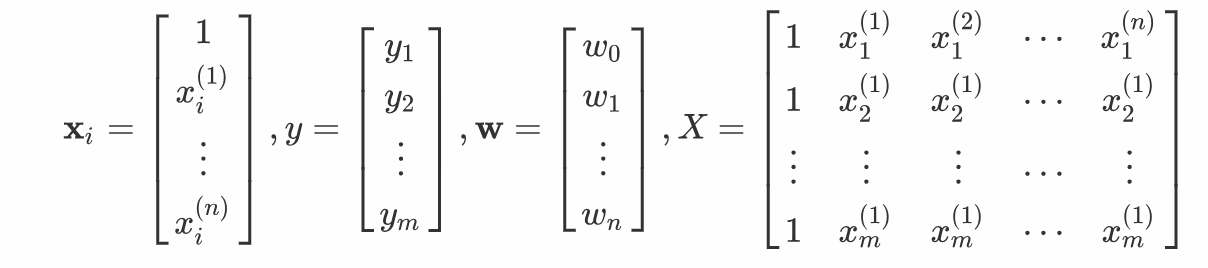


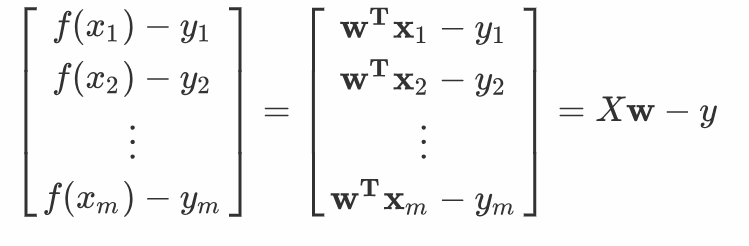
从公式可以看出对于每一个分量进行一次迭代时计算了所有训练样本数据，这种称为**批量梯度下降**。因此在数据量很大的时候，每次迭代都要遍历训练集一遍，开销会很大。

为改善上述情况，可以在每次迭代仅选择一个训练样本去计算代价函数的梯度，然后更新参数。即使是大规模数据集，随机梯度下降法也会很快收敛。这种方法称为**随机梯度下降**。

此时有：

梯度求解过程中用到了求和，代码实现时用循坏太过繁琐，可借助矩阵运算，简洁迅速。

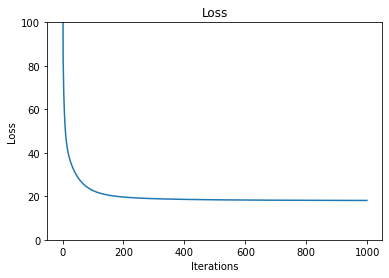




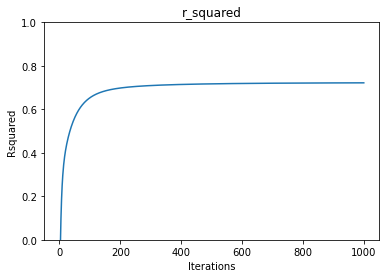
利用了梯度下降法，通过迭代，不断更新梯度与权重，得到使损失函数收敛的特征权重矩阵，再利用矩阵乘法与测试集的特征矩阵相乘得到预测值

* 1. 可视化

通过matplotlib这个库进行画图操作，将损失值作为y轴，迭代次数作为x轴，可以得到损失随着迭代次数的增加而减少，最后曲线变得比较平缓，可得代价函数已经收敛



通过R Squared（相关系数）评价线性回归模型的好坏，R Squared越大说明模型训练得越好，如同所示，随着迭代次数得增加，R Squared逐渐增大，最后趋于平缓，达到了0.72左右，训练效果比较好



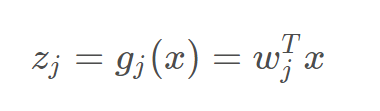
1. **Softmax回归模型的实现：**
   1. 数据预处理

未发现缺失值，不用处理异常值，对标签值进行独热编码，进行归一化

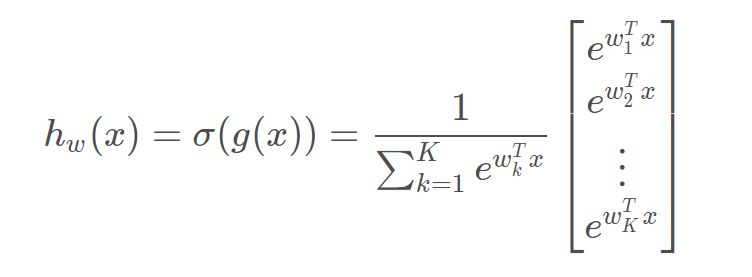
* 1. 算法实现

引用softmax函数（sigmoid函数在多分类的推广），将在负无穷和正无穷的预测结果抓换成指数函数，保证了概率的非负性，然后将转化后的结果除以所有转化后的结果之和，可以理解为转化后结果占总数的百分比，这就是softmax的值。

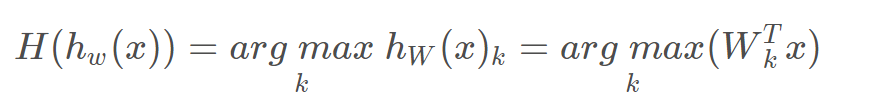
假设有k个分类，softmax对实例x的类别进行预测时需要分别计算x属于每一个类别的概率，因此每个类别拥有各自独立的线性函数



定义softmax函数

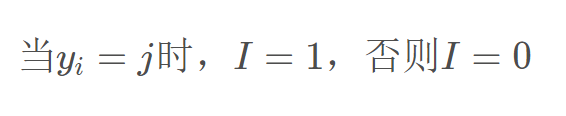
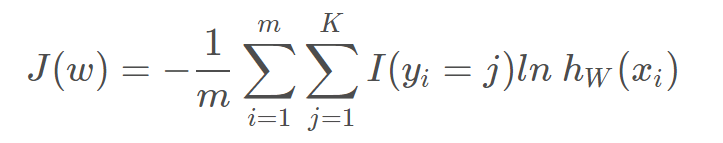


Sotfmax函数的输出是预测x为各类别的概率，通过训练确定了模型参数W,便可构建多元分类函数：

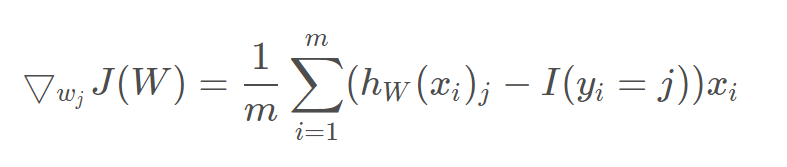


对预测的标签值进行独热编码，由最大似然估计得到的损失函数为交叉熵（不是最小二乘，因为函数会变成非凸的，不容易求解，会得到局部最优，如果用最大似然估计，目标函数就是对数似然函数，是关于w，b的高阶连续可导凸函数，可以方便通过一些凸优化算法求解，如梯度下降，牛顿法）

其中I为示性函数，用独热编码实现

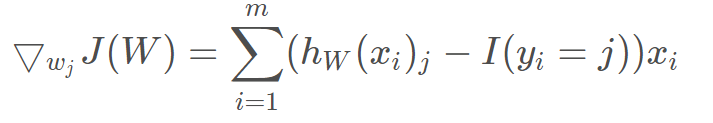


求偏导后梯度为



hW​(xi​)j​−I(yi​=j)：可以解释为模型预测xi​为第j类别的概率与其是为第j类别之间的误差。

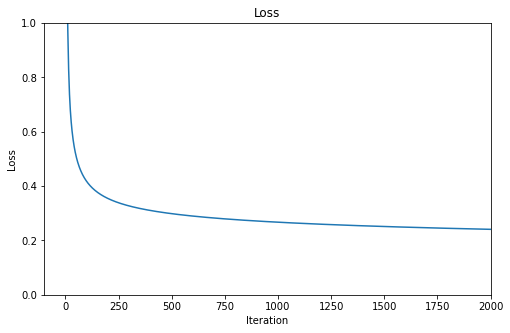
随机梯度下降对应的梯度为



通过迭代，不断更新梯度与权重，得到使损失函数收敛的特征权重矩阵，再利用矩阵乘法与测试集的特征矩阵相乘得到每个样本的属于各个类别的概率，将最大概率对应的标签值作为预测值。

* 1. 可视化

通过matplotlib这个库进行画图操作，将代价函数的值作为y轴，迭代次数作为x轴，可以得到代价函数值随着迭代次数的增加而减少，最后曲线变得比较平缓，可得代价函数已经收敛。

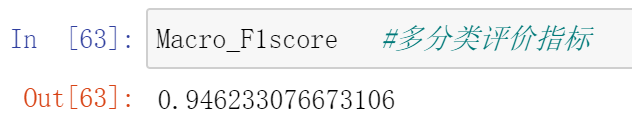


* 1. 模型评估

利用准确率进行模型的评价：准确率越接近1，模型训练越好。把训练集进行训练，得到的标签的预测值与真实的训练集的标签进行比较，得到模型的准确率为0.94705，模型训练较好。



除此之外，另外添加了所要求的评价指标:Macrofscore, Macrofscore值约为0.94623，与准确率相接近，模型训练较好



1. **K-means**

2.3.1 概述

物以类聚，人以群分：k-means算法采用距离作为相似性的评价指标，即认为两个对象越近，其相似度就越大**。**

2.3.2 算法实现

先从样本集中随机选取 n个样本作为簇中心（聚类中心），并计算所有样本与这 n个“簇中心”的距离，对于每一个样本，将其划分到与其距离最近的“簇中心”所在的簇中，对于新的簇计算各个簇的新的“簇中心”，并更新簇中心，不断地划分样本与更新簇中心直到划分结果相同就退出迭代。

1. **爬虫**

Xpath： 定位元素的方法有相对路径（使用标签名与节点属性定位、部分属性值定位、组合元素索引定位、文本内容定位）和绝对路径（可以通过浏览器复制完整Xpath的方法得到）

re模块：通过正则表达式对字符串进行匹配

使用re模块比较少，当xpath不好提取的时候会选择由re正则表达式（.\*?），比较直接。

破解反爬机制：构建合理的请求头：user-agent(浏览器信息)、cookies(用户信息：账号密码等)，模拟浏览器登录的访问、访问频率限制：time.sleep(2)，每次访问都要间隔几秒，模拟人为方位，防止被检测出是爬虫行为、使用代理ip。

**3 评价指标说明**

使用**R-squared**评估线性回归模型的好坏，使用**Accuracy**和**MacroF1Score**对softmax回归模型进行评估

* **R-squared**

R-squared = RSS/TSS=1-ESS/TSS

其中：

TSS (total sum of squares)：总平方和  
RSS (regression sum of squares)：回归平方和（预测值与真实值的差的平方和）  
ESS (error sum of squares) ：残差平方和（真实值与均值的差的平方和）

* **Accuracy**

Accuracy (准确率)：所有预测与真实为同一类的样本数占总样本的比例

* **MacroF1Score**

MacroF1Score = 2\*P\*R/(P+R)

其中：

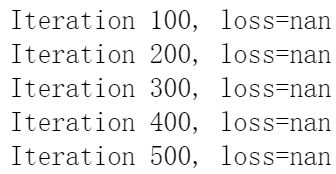
P（精确率）：所有被判定为正类中，真实的正类所占的比例

R（召回率）：所有真实为正类中，被判定为正类占的比例

补充：在softmax多元回归中，如指定一个标签值为正类，其余的都为负类

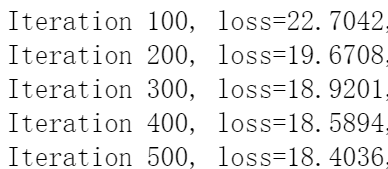
**4 实现过程中遇到的问题与解决方法**

* 线性回归模型：出现了损失值溢出、得不到权重值的现象。原因：数据处理不充分，只处理了缺失值。

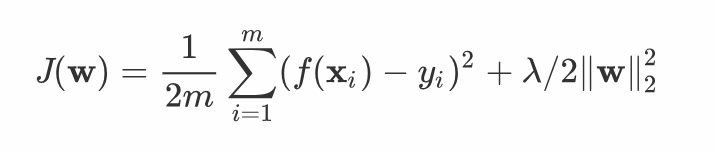


解决办法：归一化（Min-Max scaling）操作，将训练集的特征值转化到[0,1]的范围，统一量纲，使损失函数更好地收敛。

归一化后的效果：



归一化后模型的训练结果不是特别好，在评估网站的分数只有0.4。分数低说明有可能是(直观上)发生过拟合（数学角度上是特征矩阵不可逆，解出的w有多个）(**过拟合Overfitting**：就是太过贴近于训练数据的特征了，在训练集上表现非常优秀，近乎完美的预测/区分了所有的数据，但是在新的测试集上却表现平平，不具泛化性，拿到新样本后没有办法去准确的判断)，加入L2正则项之后



效果并没有体现，分数反减不增，那么得出结论，并不是发生了过拟合，并且存在其他的问题。

解决方法：重新整理思路，在草稿纸复现数据预处理与模型，查找资料。

通过不断地检查，发现是没有将测试集归一化，而正确的操作是将训练集和测试集同步归一化，如果训练集归一化了，那么测试集也要归一化。改正操作后在评估网站得到的分数为0.7，有了较大的提升。

* 还有没有优化方法可以使模型拟合效果更好一些、分数更高一点呢？那么通过巩固知识体系，发现归一化的弊端：受异常值影响较大。原因是因为异常值过多，最大值与最小值与异常值有关，由归一化的公式可知，归一化效果受最大值最小值影响。于是添加了去除异常值的操作，运用 3σ原则，使特征值满足正态分布，而超过3倍的标准差的异常值（几乎不可能出现的异常值），使得得分上升至0.8，有一定的提升。
* 现在的模型训练得比较好，但是也有存在过拟合的几率，针对数据处理和模型优化，测试了多种方法组合（处理异常值与正则化，不处理异常值与正则化，处理异常值与非正则化，不处理异常值与非正则化），发现每次的分数都不一样，其中处理异常值与非正则化的效果最好。

**5 收获与总结**

**5.1 收获**

* 一句业界广泛流传的术语：数据和特征决定了机器学习的上限，而模型和算法只是逼近这个上限而已 —— 这句话使我醍醐灌顶：原来数据处理如此的重要！而且可以调库的情况下，模型和算法占机器学习过程的十分之一，大部分时间都是用在数据预处理和特征工程上。
* 要明确模型实现的框架，知道要做什么，不要做什么，思路很重要，有了思路就可以比较快地写出代码、实现模型，不然就会像无头苍蝇乱撞，会走很多弯路。
* 一些代码编写的技巧：比如说不同的代码都可以实现一个删除缺失值的功能，还有删除某一列的操作，还有某种功能的实现比较繁琐，会有很多代码与方法可以简化操作，比如说均值填充缺失值不用一列一列操作，python有内置函数可以直接填充全部列等等。
* 数学知识：求偏导，对于复杂的多元函数求偏导，矩阵的乘法，一些矩阵运算的规则，一些算法知识，如正规方程法和梯度下降法的底层逻辑与直观理解。最小二乘法的矩阵形式，矩阵求偏导，迭代，类比梯度下降为下山，学习率为布长，梯度为方向导数最大值即下山方向，找到全局最低点意味着代价函数收敛。

**5.2 总结**

大概三个星期的考核一快到尾声了，这篇文档的完稿，是这段学习经历的结束，亦是下一段考核之路的启程，回想起这段时间的压力，喜悦，崩溃，仍历历在目，但是收获颇丰。感谢詹培林等师兄于百忙之中对我的学习的指导，以及待人真诚的品格，给我留下难忘的印象，是我学习的榜样。在此致以真诚的感谢。

**6 参考文献**

[1]周志华.机器学习[M]. 北京：清华大学出版社，2016：23-88

[2]邱锡鹏.神经网络与深度学习[M]. 北京：机械工业出版社，2020：23-69

[3]胡保强.如何理解和使用多元线性回归[OL] <https://zhuanlan.zhihu.com/p/49222906>

[4]Leo Lee.二分类逻辑回归与多分类逻辑回归的公式推导[OL] <https://zhuanlan.zhihu.com/p/97915473>

[5]准确率（Precision）、召回率（Recall）以及综合评价指标（F1-Measure ）https://www.cnblogs.com/bluepoint2009/archive/2012/09/18/precision-recall-f\_measures.html