**数智2022第一阶段考核**

**学院：计算机学院  
 专业：计算机类**

**姓名：李金强**

**学号：3121004863**

摘要：通过机器学习，我们应学会对训练集进行学习和训练，从而得到一个模型用于预测分析，这对于智能化的社会而言是至关重要的。在这个文档中，我将会简要概述“线性回归”和“softmax”两个模型、K-means聚类算法、数据预处理及通过爬虫从网上爬取需要的资源。

**关键词：**机器学习、模型、数据预处理、爬虫

**Abstract**: Through machine learning, we should learn to learn and train on the training set to obtain a model for predictive analysis, which is crucial for an intelligent society. In this document, I will give a brief overview of the "linear regression" and "softmax" models, the K-means clustering algorithm, data preprocessing, and the resources needed to crawl the web through a crawler.

**Keywords**: machine learning, model, data preprocessing, crawler

1. **任务实现路线**
   1. **前期预备工作**

**·** 下载使用Pycharm、jupyter notebook、Anaconda软件

· 大致了解所需的学习内容方向，提前做好部署规划

· 通过多种途径获取所需的学习资源，充分利用好学习资源进行学习

* 1. **任务内容的学习**

**·** 加强高等数学中偏导数和线性代数中矩阵部分知识的学习，为后面的 线性回归模型和softmax模型的学习夯实基础

· 学会使用python语言，掌握python语言的基础知识

· 通过相关书籍视频对模型进行学习，多看应用生活中的事例和代码

* 1. **处理任务出现的问题**

· 模型出现过拟合问题，对模型进行调参从而达到优化的效果

· 训练模型的数据出现异常，学会对异常缺失的数据进行处理

· 爬虫遇到反爬机制，通过请求头或其他爬取方式手段进行爬取

* 1. **后期进行总结**

**·** 与师兄进行学习交流探讨

**·** 考核后进行复盘，做得不够好的地方进行改进

· 对考核过程中所学知识进行总结做成笔记，方便日后复习

1. **任务的学习**

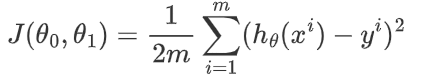
**2.1 线性回归模型**

**2.1.1 模型原理**

假设有一个函数，先设，接下来不停地一点点改变这两个参数值，使得不断变小，直到我们找到的极小值。若目标函数为凸函数，该极小值则为函数最小值，可确保收敛到全局最小解。

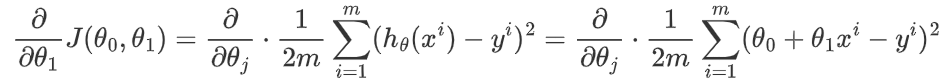
**2.1.2 模型分析**

该模型的代价函数（预测值和实际值的差的平方误差和）为：

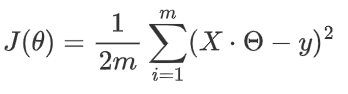


优化目标为：

对公式进行化简得到：



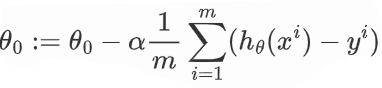
对模型进行向量化后，可得到：

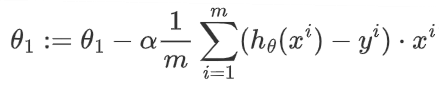


**2.1.3 模型学习率**

学习率用于控制梯度下降的速度。当学习率过小时，结果就是函数值只能一点一点的挪动，需要很多步才能到达全局最低点。如果学习率过大时，那么梯度下降很有可能会越过最低点，甚至可能无法收敛或发散。

但是在我们模型运行过程中，我们无需中途改变模型的学习率，这是因为当我们接近局部最低点时，梯度下降法会自动采取更小的幅度。这是由于在接近函数的最低点时，在该点处的导数值就会越小。





如果梯度下降算法正常工作的话，每一步迭代之后J(\theta)都应该下降。通过代价函数随迭代步数变化的曲线能判断梯度下降算法是否已经收敛。

**2.1.4 特征缩放**

如果两个变量的取值范围相差甚远，那么参数的等值线图将为一个很瘦扁的椭圆。梯度下降的过程因来回震荡而十分缓慢。

当我们对特征变量进行缩放时，梯度下降的收敛速度会变得更快，算法的效率也得以提高。除此之外，我们还能通过均值归一化提高运算效率。

**2.2 softmax模型**

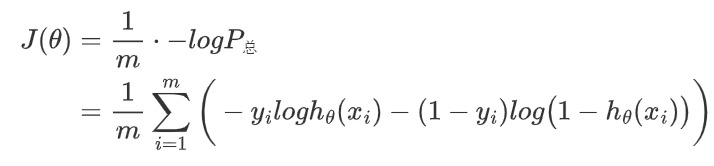
**2.2.1 模型概述**

对于给定的输入测试x，分别计算在每个类别中的概率，将概率最高的类别作为预测的结果。Softmax模型能够确保概率的非负性，同时概率的所有之和都为1。

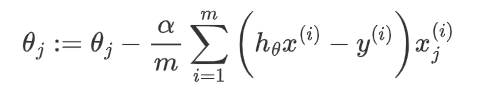
**2.2.2 代价函数**

在softmax模型中，我们不能直接引用线性回归的代价函数作为softmax模型的代价函数。如被引用时，代价函数将会有很多的极小值，称为非凸函数。若对非凸函数使用梯度下降法，无法确保能取到最小值，因为有可能落在在函数极小值无法跨越。

代价函数是衡量我们模型输出结果和实际输出结果之间的差距的函数，因此我们希望它越小越好，而我们希望合事件发生概率越大越好，所以得到代价函数：



这个函数的最小值，我们使用梯度下降法：



**2.3 数据预处理**

**2.3.1 类别数据预处理**

**· 序号编码：**用来处理类别间具有大小关系的数据，比如成绩(高中低)

步骤：1.确定x中的唯一值的个数k，将唯一值作为关键字；

1. 生成k个数字作为键值，即Value=[0,1,2,…，k]
2. 每一个唯一的类别型元素对应着一个数字，即键值对 dict={key1:0, key2:1, …,key k:k-1}

**· 独热编码：**用于处理类别间不具有大小关系的特征，比如血型。独热 编码会把血型变成一个稀疏向量，A型血表示为(1,0,0,0)，B型血表示 为(0,1,0,0)。

在独热编码下，特征向量只有某一维取值为1，其余值均为0，因此****可以利用向量的稀疏来节省空间。****如果类别型的唯一类别元素较多，可能会造成****维度灾难****，因此需要****利用特征选择来降低维度****。

**· 二进制编码：**先用序号编码给每个类别赋予一个类别ID，然后将类 别ID对应的二进制编码作为结果。

**2.3.2 结构化数据预处理**

**缺失值处理：**：

**·删除**（当缺失量大于总数据量的40%~50%及以上时删除属性，当整个数据集中缺失值较少或者缺失值数量对于整个数据集来说可以忽略不计的情况下删除样本）

**·统计值填充**（泛指平均值、中位数、众数、最大值、最小值）

**·统一值填充**（**自定义指定的某一个常数**）

**·前后向值填充**（使用缺失值**前一个或后一个数值**作为填充值进行填充）

·**插值法填充**（在x范围区间挑一个或自定义一个数值，然后代进**插值模型**公式中求出数值作为缺失值的数据）

**·建模预测填充**

**·具体分析**

**2.4 K-means聚类算法**

**2.4.1 聚类概述**

在无监督学习中，要将这系列无标签的数据输入到算法中，然后让算法找到一些隐含在数据中的结构，上图中能找到的结构就是两组分开的点集，而这些能找出这些点集的算法被称为聚类算法。

**2.4.2 算法实现**

首先随机生成两个点，这两点叫做聚类中心。K-Means算法中每次内循环的第一步要进行簇分配，也就是说，遍历每个样本点，根据每个样本与聚类的距离远近将每个样本点分配个聚类中心。K-Means算法中每次内循环的第二步要移动聚类中心，要做是找到所有属于某一聚类的点并计算出它们的均值，然后把这类的聚类中心移到那里，红色聚类中心也是一样的操作。然后接着继续做簇分配和移动聚类中心，迭代多次之后，就会完成最终的两个点集的聚类，这就可以说K-Means算法已经聚合了。

**2.5 爬虫**

**2.5.1 re模块**

正则表达式(regular expression)描述了一种字符串匹配的模式（pattern），可以用来检查一个串是否含有某种子串、将匹配的子串替换或者从某个串中取出符合某个条件的子串等。

**.\*?** 意味着进入了非贪婪模式匹配，这是.\*尽可能匹配少的字符

尽量使用泛匹配、使用括号得到匹配目标、尽量使用非贪婪模式

**2.5.2 Xpath**

**XPath 使用路径表达式来选取 XML 文档中的节点或节点集。节点是通过沿着路径来选取的。定位元素可以通过**相对路径**(标签名、属性节点定位，文本内容定位)或**绝对路径**（可在网页上直接复制Xpath）**

****2.5.3 解决部分反爬机制问题****

**· 构建请求网页的请求头（user-agent）**

**· 使用代理ip**

**· 限制访问频率，防止访问频繁访问被机器识别**

1. ****任务问题的处理****
   1. ****数据预处理****

**在提供的训练集中，有部分数据缺失，用均值将数据填充。**

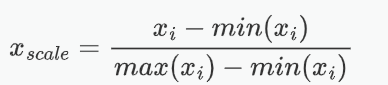
## **同时，需要对异常值进行检测。对异常值的处理采取3σ原则**。

3σ原则指若数据集服从正态分布，异常值被定义为一组测定值与其平均值的差的绝对值超过3倍的标准差。注意：使用前要预先判断数据是否符合正态分布规律。

* 1. **数据归一化**

当特征变量的数值相差甚远时，会采用归一化进行处理。梯度下降的收敛速度会变得更快，算法的效率得以提高。一般来说，归一化的目的是将特征的取值约束到(-1,1)这个区间中。

数据归一化的处理公式为：



1. ****收获与感悟****

**（由于一二轮考核学习知识的时间只有十二天，且这两轮考核的感悟都基本相同，所以此部分内容和第二轮考核文档的收获与反思结合在一起了）**

1. ****参考文献****
2. **埃里克·马瑟斯.python编程从入门到实践[M].北京:人民邮电出版社**
3. **黑马程序员.数据预处理[M].北京:人民邮电出版社**
4. **崔庆才.Python3网络爬虫开发实战[M].北京：人民邮电出版社**