对数几率回归进行多分类任务

基于Iris数据集

计算机1604 王殊 1611640413

数据集

Iris 数据有4个属性，共三种鸢尾花的类型，分别为：Iris-versicolor、Iris-virginica和Iris-setosa。为了方便将数据的分布打印在屏幕上。我选取了花的属性“sepal length”和“petal length”。三种花关于这两个属性的分布：

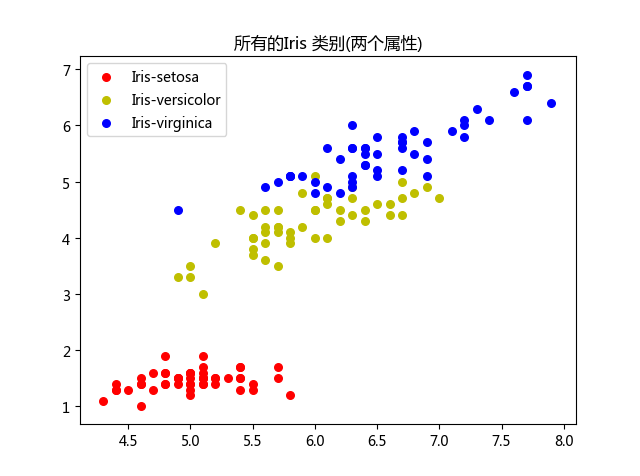


图1

数据划分

三种类型需要用到三个二分类。按照要求，数据集被划分为训练集和测试集，比较不同训练集的样本

数量占全部样本的50%、70%和90%对于测试的准确性的影响。

三种以不同花为正例的二分类器分别估计这个样本是正例的可能性，最终判断花的类型是基于三种二分类器产生的估计的概率的最大值。

原理

|  |  |
| --- | --- |
| 符号 | 意义 |
| n | 样本的特征 |
| m | 样本数量 |
|  | 第i+1个样本 |
|  | 第i+1个样本的标签 |
|  | 对第i+1个样本估计的标签 |
|  | 假设函数 |
|  | 需要拟合的线性方程 |

在一般的线性回归中，我们希望估计的输出与样本的真实输出更可能地接近。然而在对数几率回归中，我们需要预测的是样本是正例的可能性。所以需要使用sigmoid函数将拟合的方程变成概率：

和b可以认为是位置的参数，通常将他们增广后的新的作为权值，同时给样本**X**添加一行用来保持不变。根据离散型的最大似然估计，概率质量函数相乘：

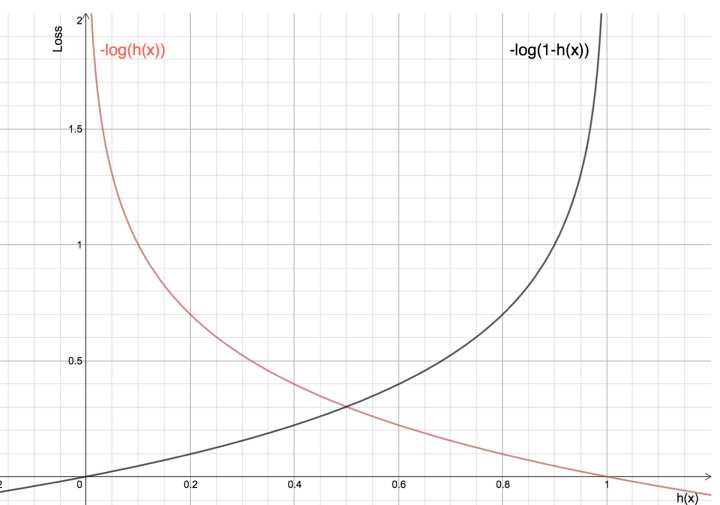
因为对于浮点数的乘法溢出的可能，所以可以取对数规避这个问题：

同时转换成矩阵相乘的形式：

优化的目标J：

这里的损失函数可以认为是，最大似然估计估计的是使最大的参数，所以需要使用-cost作为优化的目标。所以J可以表述为：

观察函数和在(0,1)的图像可以发现是凸函数：



那么由凸函数的性质，两个凸函数相加还是凸函数。就可以使用诸如梯度下降算法等优化算法求解。以梯度下降为例子。

求出梯度：

当然也可以使用矩阵的形式：

然后更新权重：

只要满足某个条件就可以退出循环。程序使用的是Batch的梯度下降。

实验结果

* 采用50%的样本作为训练集：

|  |  |
| --- | --- |
| 图像 | 说明 |
|  | 以Iris-setosa为正例的决策边界 |
|  | 以Iris- versicolor为正例的决策边界 |
|  | 以Iris- virginica为正例的决策边界 |
|  | 以Iris-setosa为正例时，随着迭代次数的增加，优化目标函数值的变化。 |
|  | 以Iris-versicolor为正例时，随着迭代次数的增加，优化目标函数值的变化。 |
|  | 以Iris-virginica为正例时，随着迭代次数的增加，优化目标函数值的变化。 |

三个二分类器的最终训练得到的权值：

|  |  |  |
| --- | --- | --- |
| 正例 |  | 迭代次数 |
| Iris-setosa |  | 100000 |
| Iris-versicolor |  | 85980 |
| Iris-virginica |  | 100000 |

训练集的准确率：92% ,错误6个样本。

测试集的准确率：96% ,错误3个样本。

* 采用70%的样本作为训练集：

|  |  |
| --- | --- |
| 图像 | 说明 |
|  | 以Iris-setosa为正例的决策边界 |
|  | 以Iris- versicolor为正例的决策边界 |
|  | 以Iris- virginica为正例的决策边界 |
|  | 以Iris-setosa为正例时，随着迭代次数的增加，优化目标函数值的变化。 |
|  | 以Iris-versicolor为正例时，随着迭代次数的增加，优化目标函数值的变化。 |
|  | 以Iris-virginica为正例时，随着迭代次数的增加，优化目标函数值的变化。 |

三个二分类器的最终训练得到的权值：

|  |  |  |
| --- | --- | --- |
| 正例 |  | 迭代次数 |
| Iris-setosa |  | 100000 |
| Iris-versicolor |  | 100000 |
| Iris-virginica |  | 100000 |

训练集的准确率：94.29% ,错误6个样本。

测试集的准确率：95.56% ,错误2个样本。

* 采用90%的样本作为训练集：

|  |  |
| --- | --- |
| 图像 | 说明 |
|  | 以Iris-setosa为正例的决策边界 |
|  | 以Iris- versicolor为正例的决策边界 |
|  | 以Iris- virginica为正例的决策边界 |
|  | 以Iris-setosa为正例时，随着迭代次数的增加，优化目标函数值的变化。 |
|  | 以Iris-versicolor为正例时，随着迭代次数的增加，优化目标函数值的变化。 |
|  | 以Iris-virginica为正例时，随着迭代次数的增加，优化目标函数值的变化。 |

三个二分类器的最终训练得到的权值：

|  |  |  |
| --- | --- | --- |
| 正例 |  | 迭代次数 |
| Iris-setosa |  | 100000 |
| Iris-versicolor |  | 79054 |
| Iris-virginica |  | 100000 |

训练集的准确率：94% ,错误8个样本。

测试集的准确率：86.67% ,错误2个样本。

实验结果分析

随着训练集的比重上升，准确率基本上也基本在上升。实验结果随机划分测试集和训练集，准确率并不是一成不变的。