Pattern Recognition Assignment 1

Logistic Regression

2031561 郭超政

1. **模型概述**

逻辑回归（Logistic Regression）模型的基本形式为：

在使用逻辑回归模型解决二分类问题时可将分类概率表示为：

为了求解模型的参数，可以通过极大似然法进行估计，则根据(2)(3)得到模型的等价最小化似然函数：

通过牛顿法来求解似然函数的最优解：

根据牛顿法以及(4)，可以得到其迭代更新的公式：

1. **模型实现**

**基本函数定义：**

根据式(4)定义模型的似然函数：

1. # Likelihood function of logistics regression
2. **def** likelihood(data, label, w):
3. likelihood = -label \* (np.dot(data, w)) +
4. np.log(1 + np.exp(np.dot(data, w)))
5. **return** np.sum(likelihood)

根据式(7)(8)，定义相应的梯度以及海森矩阵计算函数：

1. # Gradient of likelihood function
2. **def** gradient(data, label, w):
3. gradient = np.zeros(w.shape[0])
4. **for** x,y **in** zip(data,label):
5. p1 = np.exp(np.dot(w, x)) / (1 + np.exp(np.dot(w, x)))
6. gradient += x \* (y - p1)
7. **return** -gradient
9. # Hessian matrix of likelihood function
10. **def** hessian(data, label, w):
11. hessian = np.zeros((w.shape[0], w.shape[0]))
12. **for** x,y **in** zip(data,label):
13. p1 = np.exp(np.dot(w, x)) / (1 + np.exp(np.dot(w, x)))
14. hessian += np.reshape(x, (x.shape[0], 1)) \* x \* p1 \* (1-p1)
15. **return** hessian

**牛顿法求解过程实现：**

训练阶段模型的输入为训练数据以及对应的数据标签，训练数据data的格式为的矩阵，包含m个数据样本，每个样本包含n个属性值，属性值需要均为数值类型；

根据式(1)中的定义，将线性模型简化为，需要将数据矩阵拓展为，则模型输入的数据矩阵尺寸为；

数据标签label为对应的n维数组，标签对应二分类的结果，使用0，1表示；

对应的模型系数coef为n+1维向量，初始化为零向量；

通过循环迭代更新系数coef估计最优解，迭代的过程通过定义最小步长e以及最大迭代次数max\_it来控制；

迭代结束后返回得到的模型系数；

1. # Solve logistics regression with newton's method
2. **def** logistic\_regression(data, label, e, max\_it=100):
3. # Expand data matrix
4. data = np.c\_[data, np.ones(data.shape[0])]
5. # Initialize coeffiencts of model y = 1 / (1 + e^(-wx))
6. coef = np.zeros(data.shape[1])
7. # Initialize step norm
8. d\_norm = np.inf
9. # Number of iteration
10. it\_count = 0
12. **while** d\_norm > e **and** it\_count < max\_it:
13. **print**("Step:", it\_count, "Likelihood:", likelihood(data, label, coef))
14. d = np.dot(np.linalg.pinv(hessian(data, label, coef)),
15. gradient(data, label, coef))
16. # Update coefficient
17. coef = coef - d
18. # Calculate step norm
19. d\_norm = np.linalg.norm(d)
20. it\_count += 1
22. **return** coef

**模型评估部分实现：**

使用计算的到的模型系数预测未知数据：

1. # Predict novel data with trained coefficient
2. **def** predict(data, w):
3. data = np.c\_[data, np.ones(data.shape[0])]
4. p0 = 1 / (1 + np.exp(np.dot(data, w)))
5. p1 = 1 - p0
6. res = np.c\_[p0,p1]
7. res = np.argmax(res, axis=1)
8. **return** res

计算预测结果的准确率：

* # Calculate the percision of predicted result
* **def** score(predict, ground\_truth):
* count = (predict == ground\_truth).astype(int).sum()
* **return** (count/len(predict))

1. **模型测试**

模型的测试选择了UCI上的Breast Cancer数据集以及Abalone数据集：

**Breast Cancer数据集**：

数据集描述的是不同的乳腺癌病例分类，病例分为良性以及恶性两个类别，分别表示为0，1；数据包含了团块厚度等9个用于描述病例状况的数值属性，属性均为数值类型，原始数据中还包含一列样本ID，实验中弃之不用；数据样本数量为699；

**数据处理：**

数据中包含少数的缺失值，共16个样本存在属性缺失，由于数量较少，将缺失样本直接去除，清理后的数据包含683个样本；

数据属性值的数值分布较为平均，且不同属性的数值量纲基本相同，固不作更多的数据预处理；

数据随机打乱后，按照4:1的比例划分为训练集及测试集，各包含546及137个样本；

数据属性值大致分布：

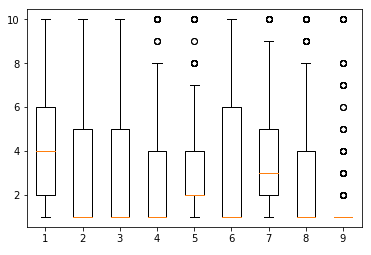
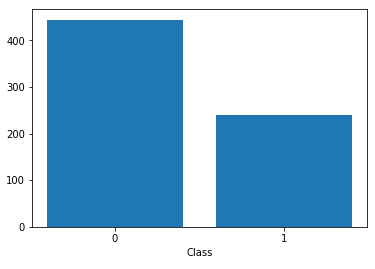
 

图1 - Breast Cancer 数据集属性取值箱线图 图2 - Breast Cancer 数据集类别数量分布

**模型运行结果：**

经过9轮迭代后模型收敛，得到的模型系数为：

[ 0.55860282, 0.06059213, 0.26964914, 0.39501854, 0.04491559,

0.41040066, 0.34391083, 0.212355 , 0.44955754, -9.68085486]

测试集中预测结果准确率为：0.97810

**模型可视化**

选择数据集中的前两个属性绘制相应的分类边界：

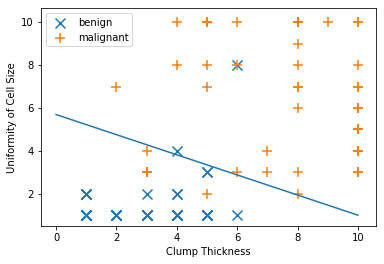


图3 - Breast Cancer逻辑回归分类边界

**Abalone数据集：**

数据集描述了不同年龄段的鲍鱼个体，通过体重、身长等物理数据来推断鲍鱼的年龄；数据集中对鲍鱼年龄的描述为整数数值，为了将其构建为一个二分类任务，将年龄字段划分为两个类别，以及，分别使用0，1表示；样本数量共4177个，属性数量8，其中包含一个性别属性为Nominal类型，其余属性均为数值类型，实验中仅选取其中7个数值类型的属性；

**数据处理：**

数据无缺失值，且无明显异常数值，同样按照4:1的比例划分训练集以及测试集，得到训练集大小3341，测试集大小836；

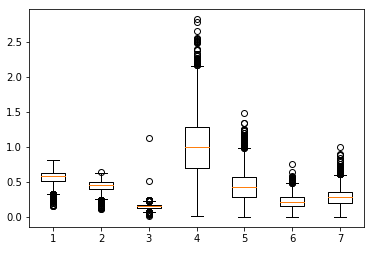
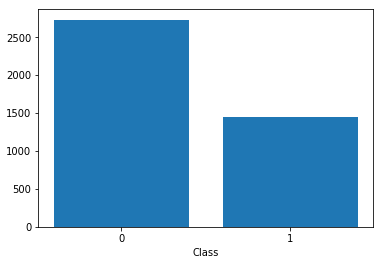
 

图4 - Abalone数据集属性取值箱线图 图5 - Abalone 数据集类别数量分布

**模型运行结果：**

经过5轮迭代后模型收敛，得到的模型系数为：

[ -4.37363705, 2.5186944 , 2.03648743, 18.36825708,

-18.90044524, -3.0565954 , 9.01004543, -1.91022282]

测试集中预测结果准确率为：0.76435

**模型可视化：**

选择数据集中的两个属性绘制相应的分类边界：

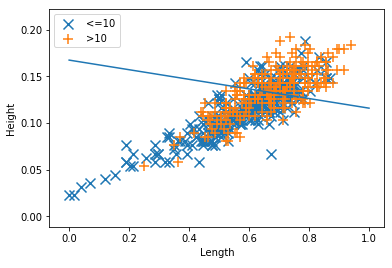


图6 - Abalone逻辑回归分类边界

1. **结果分析**

**逻辑回归优势：**

逻辑回归的过程简单，计算过程也相当迅速，能够在极短时间内得到一个粗略的模型，且模型的可解释性较强，能够较为清晰的展示不同属性之间的关系；

**牛顿法缺陷：**

在使用牛顿法求解逻辑回归模型时，因为涉及对Hessian的求解，在数据规模较大的时候Hessian矩阵的尺寸也会相应变大，增加计算的复杂程度；

**线性表达能力较弱：**

逻辑回归是广义线性模型，其表达能力较弱，在数据属性间关系较为复杂的状况下无法准确描述属性之间的关系，导致模型的效果较差；