机器学习实验报告

逻辑回归（Logistic Regression）

姓名：秦海滨

学号：1180300523

2020 年 10 月 19 号

目录

1 实验目的 **2**

2 实验要求及环境 **2**

2.1 实验要求 2

2.2 实验环境 2

3 基本思想 **2**

4 数据的生成 **3**

5 损失函数 **3**

6 梯度下降法 **4**

6.1 不带惩罚项 4

6.2 带惩罚项 5

7 实验结果分析（以二分类为例） **5**

7.1 实验数据满足朴素贝叶斯假设 5

7.2 实验数据不满足朴素贝叶斯假设 6

7.3 使用 UCI 上的数据进行测试 7

8 结论 **8**

1 实验目的

理解逻辑回归模型，掌握逻辑回归模型的参数估计算法。

2 实验要求及环境

2.1 实验要求

实现两种损失函数的参数估计（无惩罚项和有惩罚项的情况），可以采用梯度下降、共轭梯度或者牛顿法等。

2.2 实验环境

操作系统：Windows10
开发环境：Spider 4.1.4，Python 3.8

3 基本思想

逻辑回归（Logistic Regression）是一种解决分类问题的算法，虽然它的名字中带有“回归”二字，但实际上是利用回归的方法解决分类的问题。能通过不断的优化寻找最优的参数以正确地进行数据分类。常见的可以利用逻辑回归解决垃圾邮件分类问题（二分问题）、用户点击率（二分问题）问题。

实际上在之前的学习中我们就已经接触过一种回归问题，即“线性回归”。线性回归就是给定一系列数据，通过求取一个合适的线性函数，使其尽可能多地包含所有的数据，线性函数的表达式如下：

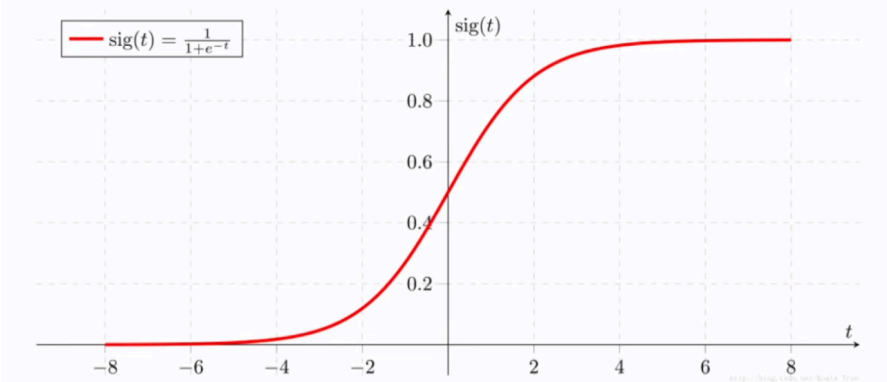
$$f(x) = w^T x + b$$

其中，w 和 b 都是通过学习得到的，我们常用的方法即是最小二乘法，通过最小化数据与模型的均方误差，求得优化后的参数。

实际上，“逻辑回归 = 线性回归 +Sigmoid 函数”。线性回归是一类回归问题，需要我们输出一条连续的曲线以表示数据中变量与变量间的关系；而逻辑回归是一类分类问题，通过数据中各特征维度的值对数据进行分类，输出的是对原始数据的划分结果。由于在线性回归模型中 $f(x) = w^T x + b$ 的值域是 $[-\infty, +\infty]$，我们可以根据线性回归模型得到任意的一个值；但在逻辑回归中，我们需要通过概率对数据进行分类，即模型的值域应该在 $[0, 1]$ 内。此时我们引入 Sigmoid 函数，其表达式为：

$$sig(t) = \frac{1}{1 + e^{-t}}$$

其函数图像如下：



(a) Sigmoid 函数

其连续性以及可导性在之后的优化中显得格外重要。

此时我们就可以将线性回归的回归问题转化为一个逻辑回归的分类问题：

$$sig(w^T x + b) = \frac{1}{1 + e^{-(w^T x + b)}}$$

我们将 Sigmoid 函数的输出视作概率值作为分类的判断依据，若 sig 输出大于等于 0.5，则将其视为正例；若 sig 输出小于 0.5，则将其视作反例。

现在我们要做的，就是在给定样本集的基础上，通过学习、优化求出最优的参数 w^* 和 b^*，使该模型能够对其他的样本输入做出分类。

4 数据的生成

可以通过输入不同类别的每一维数据的均值 μ 和方差 σ 以单独生成满足高斯分布的数据。最后将多个类别的数据混合之后返回。

实验要求我们生成满足朴素贝叶斯假设的数据和不满足朴素贝叶斯假设的数据以测试逻辑回归的结果，我们可以通过调整输入的协方差矩阵以输出需要的数据类型。

5 损失函数

由于逻辑回归解决的是分类问题，所以其损失函数也是按照分类情况进行计算的。以二分类为例，我们定义逻辑回归的损失函数如下：

$$cost_i = \begin{cases} -ln(p_i), & if\ y_i = 1 \\ -ln(1 - p_i), & if\ y_i = 0 \end{cases}, \ p_i = \frac{1}{1 + e^{-(w^T x_i + b)}}$$

其中 p 表示 Sigmoid 函数的输出，即代入数据 x 算出的将其分为第一类（$y_i = 1$）的概率值。这个公式表示，若数据属于第一类（$y_i = 1$），则此时判断出

的 p 越小，cost 越大；若数据属于第 0 类（y=0），则此时判断出的 p 越大，cost 越大。为了方便计算，我们将这个分段函数记为：

$$cost_i = -y_i ln(p_i) - (1 - y_i) ln(1 - p_i)$$

所以对于一个给定的样本，我们可以计算样本集内所有数据的损失，即对所有的数据点的损失值求和，但在实际实验过程中我们会发现直接求和会导致溢出问题，故我们在公式前加上一定的系数防止溢出，如下：

$$COST = -\frac{1}{m}\sum_{i=1}^{m}(y_i ln(p_i) + (1 - y_i) ln(1 - p_i))$$

我们也可以通过最大似然估计方式对参数进行估计，最终得出的公式与以上公式类似。

可以证明，以上得出的 COST 函数是关于 w 和 b 的凸函数，我们可以通过梯度下降法求出使 COST 函数取最小值时的参数 w^* 和 b^*，构建出指定样本集上的逻辑回归模型。

为了方便计算，我们记常数项 b 为 w 中的第 0 维参数，即 w_0，并在数据 x 中添加相应的维度，将 p_i 转换为如下形式:

$$p_i = \frac{1}{1 + e^{-w^T x_i}},\ w = \begin{bmatrix} w_0 \\ w_1 \\ \cdots \\ w_n \end{bmatrix},\ x_i = \begin{bmatrix} 1 \\ x_{i1} \\ \cdots \\ x_{in} \end{bmatrix}$$

这样，损失函数 COST 的取值便只与 w 的取值有关，我们可以记该函数为 $COST(w)$，我们只需要求解出使损失函数最小的参数 w^* 即可。

6 梯度下降法

与 Lab1 中类似，在第 i+1 论的迭代中的梯度下降的公式如下，其中 η 为步长（即学习率）：

$$w_{i+1} = w_i - \eta \frac{\partial\, COST(w)}{\partial w}$$

6.1 不带惩罚项

将 Sigmoid 公式代入到 COST 公式中化简，我们可以将 COST 的公式化简为如下向量形式：

$$COST(w) = -\frac{1}{m}\sum_{i=1}^{m}((1 - y_i) w^T x_i - ln(1 + e^{w^T x_i}))$$

对该式求导有：

$$\frac{\partial\ COST(w)}{\partial\ w} = -\frac{1}{m}\sum_{i=1}^{m} x_i((1-y_i) - \frac{e^{w^T x_i}}{1+e^{w^T x_i}})$$

将求出的导数代入到梯度下降的公式中，我们可以通过不断的迭代求出使 $COST(w)$ 最小的 w^* 的值。为了防止无限的迭代，需要我们设置相应的最大迭代次数以及迭代精度。

6.2 带惩罚项

为了防止过拟合现象，我们需要为模型添加一定的惩罚项。类似于 Lab1 中添加惩罚项的方式，我们将损失函数修改为以下形式：

$$COST(w) = -\frac{1}{m}\sum_{i=1}^{m}((1-y_i)w^T x_i - ln(1+e^{w^T x_i})) + \frac{\lambda}{2m}w^T w$$

求导的结果为：

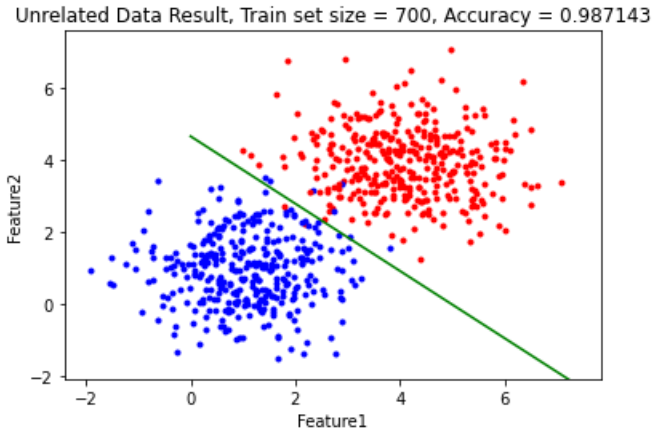
$$\frac{\partial\ COST(w)}{\partial\ w} = -\frac{1}{m}\sum_{i=1}^{m} x_i((1-y_i) - \frac{e^{w^T x_i}}{1+e^{w^T x_i}}) + \frac{\lambda w}{m}$$

与不带惩罚项的迭代方式相同，我们可以求解出 $w*$ 的值，加入惩罚项后的模型在理论上应该有着防止过拟合的能力，我们会在后续的实验结果分析中对其进行验证。

7 实验结果分析（以二分类为例）

7.1 实验数据满足朴素贝叶斯假设

在生成数据之后，按照 3:7 的比例将数据集分割为训练集和测试集，共生成了 1000 个测试数据，数据均有两个特征。设置超参数为 $\lambda = 0.002$，测试结果如下：



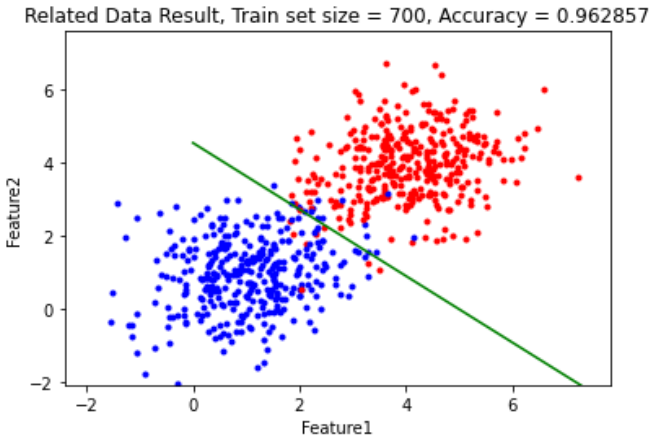
(b) 满足朴素贝叶斯假设的训练集测试结果

(c) 满足朴素贝叶斯假设的测试集测试结果

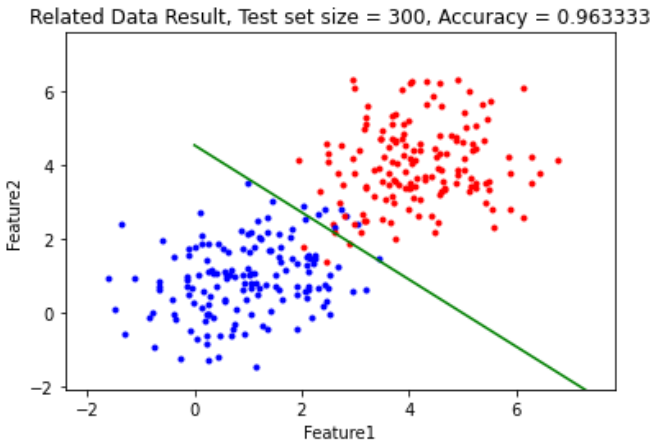
可以看出模型的拟合效果较为准确，在训练集和测试集上表现均良好。

7.2 实验数据不满足朴素贝叶斯假设

与上一节类似的，我们只需要在生成数据时修改各个维度之间的协方差矩阵，使其维度之间不相互独立。同样生成 1000 个二维数据，设置超参数为 $\lambda = 0.002$，测试结果如下：



(d) 不满足朴素贝叶斯假设的训练集测试结果



(e) 不满足朴素贝叶斯假设的测试集测试结果

可以看出模型的拟合效果也较为良好，相较于满足朴素贝叶斯假设的数据，准确度有一定的下滑，但影响不大。

7.3 使用 UCI 上的数据进行测试

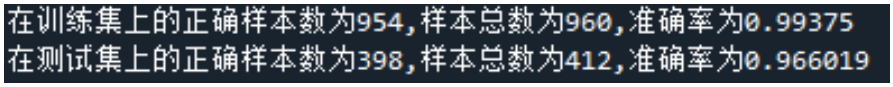
在此选择了 UCI 网站上的名为“banknote authentication Data Set”的数据集，其中包含了从真假钞图像中提取出的四个维度的特征数据，以及对真假钞的分类结果。数据集的摘要如下：

Abstract: Data were extracted from images that were taken for the evaluation of an authentication procedure for banknotes.

Data Set Characteristics:	Multivariate	**Number of Instances:**	1372	**Area:**	Computer
Attribute Characteristics:	Real	**Number of Attributes:**	5	**Date Donated**	2013-04-16
Associated Tasks:	Classification	**Missing Values?**	N/A	**Number of Web Hits:**	267511

(f) 数据集摘要

在读取数据之后，按照 7:3 的比例将其分割为训练集和测试集，最终模型的测试结果如下：



(g) UCI 数据集分类结果

由以上分类结果观察，该模型针对该数据集的分类效果较好。

8 结论

逻辑回归在一定程度上以较好的效果解决了基本的分类问题。其对于满足朴素贝叶斯假设的数据有着十分优秀的分类效果，但从实验结果上观察，该模型对于不满足朴素贝叶斯假设的数据也有着不俗的分类效果。

逻辑回归在现在仍然是一种高效的分类模型，我们利用该模型对于 UCI 上的数据集的分类效果也十分优秀。