哈尔滨工业大学计算机科学与技术学院

实验报告

课程名称：机器学习

课程类型：选修

实验题目：逻辑回归

学号：1160300312
姓名：靳贺霖

一、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

二、实验要求及实验环境

实验要求：

实现两种损失函数的参数估计（1.无惩罚项；2.加入对参数的惩罚），可以采用梯度下降、共轭梯度或者牛顿法等

实验环境：

操作系统：windows 7
编译环境：python3.7,
编译器：PyCharm

三、设计思想（本程序中的用到的主要算法及数据结构）

1．算法原理

之前我们使用的朴素贝叶斯方法来得到一个分类器的方法是通过学习计算 P(Y)和 P(X|Y)，通过贝叶斯公式来得到 P(Y|X)。现在我们换一种思路，来直接学习 P(X|Y)。直接学习 P(X|Y)的方法之一就是 Logistic Regression。

这个问题中，我们要求得一个分类器 **f : X -> Y**，其中 X 是一个关于要分类的事物的一些特征的向量$< X_1 \dots X_n >$，Y 是一个布尔值 0 或 1。而且假定所有的X_i在给出 Y 的情况下都是条件独立的，而且$\mathrm{P}(X_I|Y = y_k)$满足高斯分布$\mathrm{N}(\mu_{ik}, \sigma_i)$，$\mathrm{P}(Y)$满足 Bernoulli(π)。我们由贝叶斯公式，能得到

$$\mathrm{P(Y = 1|X)} = \frac{P(Y = 1)P(X|Y = 1)}{P(Y = 1)P(X|Y = 1) + P(Y = 0)P(X|Y = 0)}$$

将已知的分布带入到上式，并且经过简化和代换，我们就能得到

$$\mathrm{P(Y = 1|X =}< X_1, \dots X_n >) = \frac{1}{1 + \exp(w_0 + \sum_i w_i X_i)}$$

这样我们就能通过估计$w_0, \dots w_n$来得到P(Y|X)。最直接能想到的就是使用最大似然法来估计这些参数。这里使用最大条件似然估计

$$M_{CLE} = argmax_w \prod_l P(Y^l|X^l, w)$$

对似然函数进行对数处理，并通过变换，得到似然函数

$$\mathrm{L(W)} = \sum_l Y^l \left(w_0 + \sum_i^n w_i X_i^l \right) - \ln(1 + \exp(w_0 + \sum_i^n w_i X_i^l))$$

对于这个似然函数，求出使似然函数最大的 W 即为所要估计的参数。对参数的估计我们这里使用梯度下降法以及牛顿法。之前的报告中已经介绍过梯度下降的原理，这里主要介绍牛顿法的原理。

对于非线性函数$\mathrm{f(x)} = 0$的求解，对其在x_0进行泰勒展开，并且忽略一定的低阶项，则有$f(x_0) + (\mathrm{x} - x_0)f'(x_0) = 0$，若$f'(x_0) \neq 0$，则解为

$$\mathrm{x} = x_0 - \frac{f(x_0)}{f'(x_0)}$$

很明显，这个解明显只是一个近似解，或者换一种说法，只是对解进行一次逼近。这样我们很自然就想到了通过这个式子不断迭代来逼近解析解。其实这就是牛顿法的过程。事实证明，牛顿法可以很快得逼近最优解。

2. 算法的实现

首先要实现的是数据的生成。为了能够明显得在坐标轴中体现，这里生成的每个样本有两个特征值。而且为了能使分类效果明显，以及符合事实的规律，这里对两种不同的类别，相同类别的样本的每一个特征值符合正态分布。然后还要求要生成不满足贝叶斯假设的数据，这里选择的是生成多维正态分布数据。然后对于所有生成的数据，按照 7：3 的比例分别分配给训练样本和测试样本。

首先要实现的是梯度下降来优化求解。可以证明似然函数是凸函数，所以使用梯度下降一定会向最优解逼近而不是陷于局部最优解。首先定义解的初始位置 w（这里使用全为 1 的列向量），然后似然函数对 W 矩阵求导，并且用合适的步长α乘这个导数，得到解要偏移的值，然后就可以进行迭代。迭代函数是

$$\mathrm{w} := \mathrm{w} - \alpha * \left(\frac{1}{m}\right) * x^T * (g(x * w) - T)$$

其中，x 为 n*3 的矩阵，而且第一列为全 1，每一行后两列的元素代表两个特征值。T 为每一个样本对应的分类布尔类型的数，g 函数为 sigmoid 函数，但是要做一些改进能够对列向量的每一个元素求值。这样，迭代合适的次数即可得到近似解。

然后要实现的是牛顿法来优化求解。因为该问题要求解的是$L'(\mathrm{w}) = 0$，所以带入牛顿法中，迭代的函数为

$$\mathrm{w} := \mathrm{w} - \frac{L'(w)}{L''(w)}$$

一阶倒已经在梯度下降法中求解过了，所以主要是要求似然函数的二阶导。经过求导可以进一步得到迭代函数为

$$\mathrm{w} := \mathrm{w} - \alpha * \left(\frac{1}{m}\right) * ((xdiag\{(1 - g(x * w)g(x * w)\}x^T)^{-1}x^T * (g(x * w) - T))$$

然后要考虑的就是正则项的添加。添加正则项在这里就比较容易，如果添加 w 的二范式来作为正则项，那么在梯度下降和牛顿法中只需要正常求一阶导和二阶导就可以实现。

四、实验结果与分析

首先考虑比较几种不同的方法的正确率结果，这几种方法分别是梯度下降、梯度下降加正则项、牛顿法和牛顿法加正则项。对于数据的生成，每种类别生成 50 个数据点，这里取两种分类的第一维均值为 1 和 2，第二维的均值为 1 和 2，两维数据的方差为 0.4 和 0.6，正则项的超参数 lambda=0.001，得到的正确率如下表所示

方法	第一次实验	第二次实验	第三次实验	平均值
梯度下降	0.93	0.92	0.90	0.915
梯度+正则	0.93	0.94	0.91	0.925
牛顿法	0.94	0.93	0.91	0.925
牛顿+正则	0.94	0.94	0.90	0.93

可以看出，其实不同的方法得到的结果正确率都挺高，但是由于线性分类器的局限性，所以很难达到完全正确，这里给出一次实验得到的分类结果

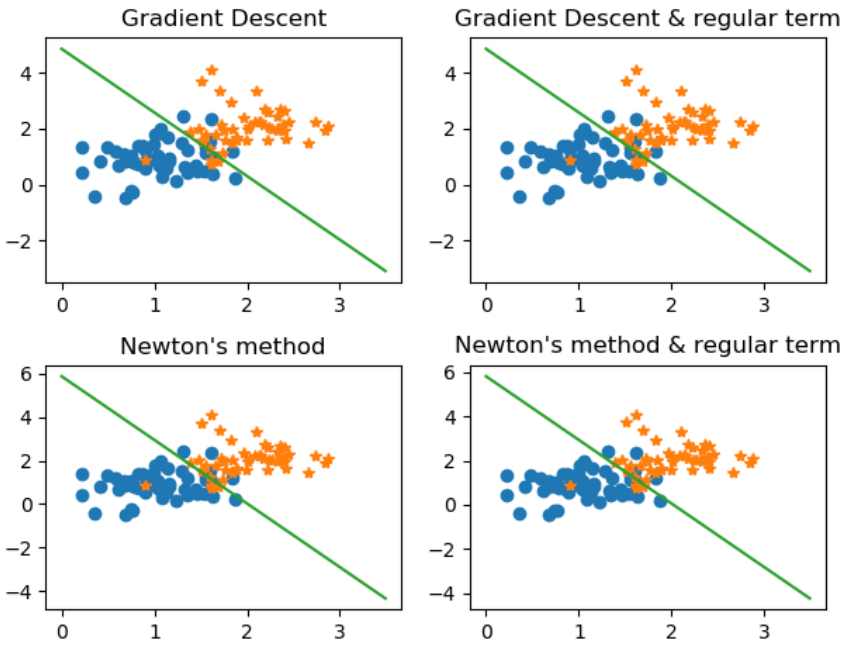

figure 1:四种不同方法得到的分类结果，可以看出差别不大

然后考虑极端情况，即两维均值对应坐标相差较远的情况，其他情况不变，得到正确率为

方法	第一次实验	第二次实验	第三次实验	平均值
梯度下降	1	1	1	1
梯度+正则	1	1	1	1
牛顿法	1	1	1	1
牛顿+正则	1	1	1	1

可以看出，这样情况下分类效果是明显的，几乎能达到完全正确，得到结果图如下所示

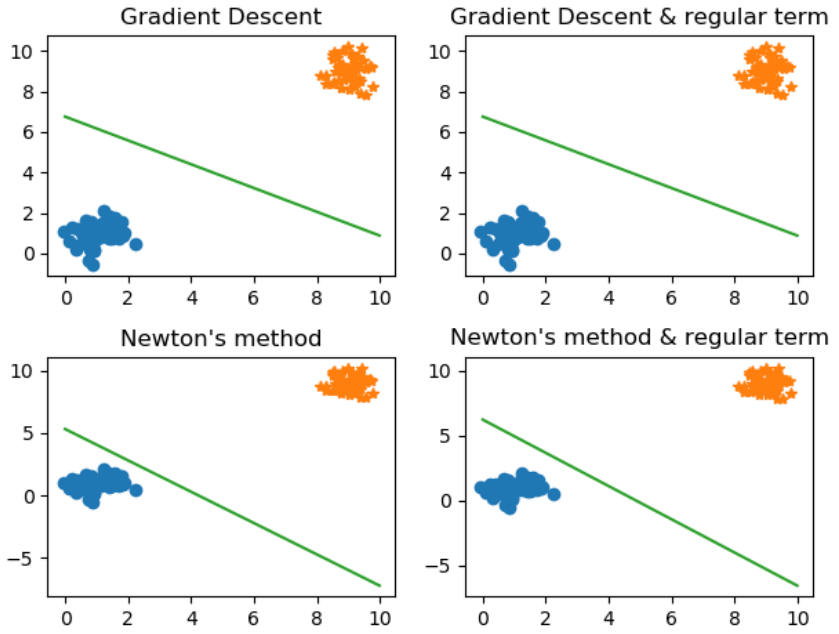

figure2:当均值相差较大时候的结果，这时候结果不同，但是分类效果都很好

可以看出这时候的分类效果很好，但是结果的斜率截距明显不同。然后考虑均值完全相同的情况，其他条件不变，得到的正确率如下

方法	第一次实验	第二次实验	第三次实验	平均值
梯度下降	0.52	0.53	0.51	0.52
梯度+正则	0.52	0.53	0.51	0.52
牛顿法	0.5	0.5	0.5	0.5
牛顿+正则	0.57	0.48	0.51	0.53

可以看出，这时得到的正确率和随机猜测的几率大概相同，只有一半的概率能划分正确，其实这时对应的两类数据其实可以看作是一类的，因为已知的特征值都是相同的，得到图像如下所示

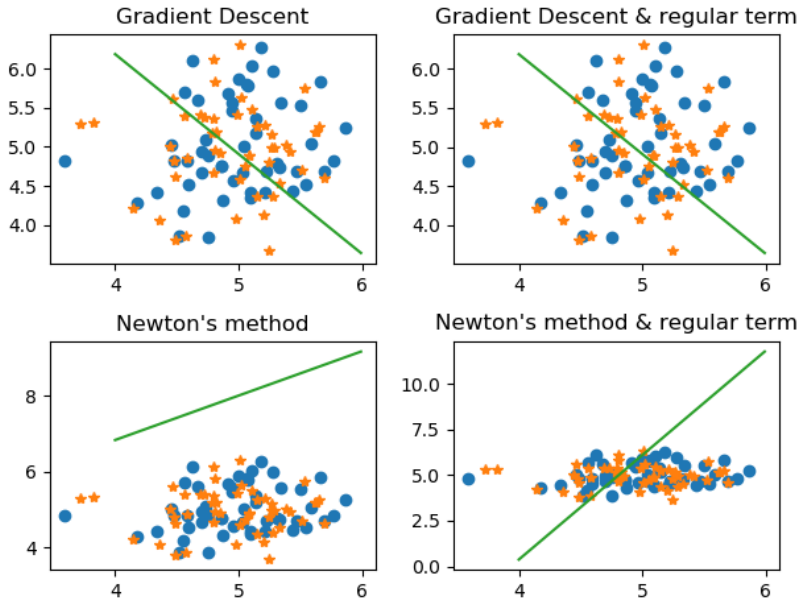

figure3:当两类数据的特征均值都相同时的结果

然后考虑当给出的数据不满足朴素贝叶斯假设的情况，这时候第一维数据满足均值为[2,8]，第二维数据的均值为[3,7]，方差均满足协方差阵[[0.5, 0.4], [0.4, 0.5]]，得到的结果正确率如下表

方法	第一次实验	第二次实验	第三次实验	平均值
梯度下降	0.96	0.94	0.95	0.95
梯度+正则	0.98	0.96	0.94	0.96
牛顿法	0.98	0.96	0.98	0.97
牛顿+正则	1.0	0.98	0.98	0.99

可以看出，当数据不满足朴素贝叶斯假设的情况，一些情况下逻辑回归仍然能得到比较好的结果，这里给出一次实验得到的结果图

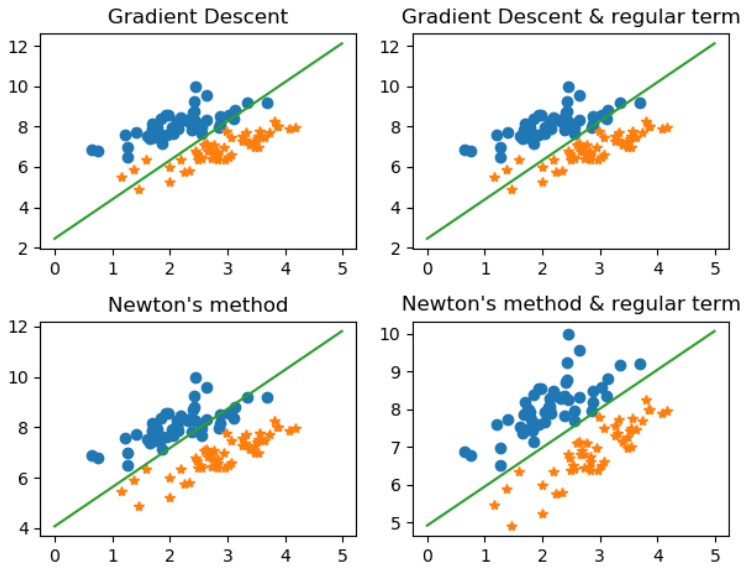

figure 4：不满足朴素贝叶斯的情况

最后，考虑在 UCI 上找一组实际数据来对实验结果进行测试。这里选择的是一共 100 个样本数据，数据有九个特征值，特征值对应属性如下所示

Attribute Information:

Season in which the analysis was performed. 1) winter, 2) spring, 3) Summer, 4) fall. (-1, -0.33, 0.33, 1)

Age at the time of analysis. 18-36 (0, 1)

Childish diseases (ie , chicken pox, measles, mumps, polio) 1) yes, 2) no. (0, 1)

Accident or serious trauma 1) yes, 2) no. (0, 1)

Surgical intervention 1) yes, 2) no. (0, 1)

High fevers in the last year 1) less than three months ago, 2) more than three months ago, 3) no. (-1, 0, 1)

Frequency of alcohol consumption 1) several times a day, 2) every day, 3) several times a week, 4) once a week, 5) hardly ever or never (0, 1)

Smoking habit 1) never, 2) occasional 3) daily. (-1, 0, 1)

Number of hours spent sitting per day ene-16 (0, 1)

Output: Diagnosis normal (N), altered (O)

然后对数据进行相应的处理，使结果判别改为 0 和 1 的判别，这样就能够适配之前写的代码。这里只使用梯度下降法，分别将 100 组数据中的前 70 组作为训练样本，后 30 种作为测试样本，得到如下的结果：

方法	第一次实验	第二次实验	第三次实验	平均值
梯度下降	1	1	1	1

可以看出这种情况以及数据的情况下分类结果显示出特别好的结果，多次实验均为所有测试均通过。

五、结论

1. 使用逻辑回归实现线性分类器能过得到比较好的结果；
2. 使用逻辑回归实现线性分类器正则项的添加与否对于分类器的有效性的影响不大；
3. 当不同类别的特征值差别较大时，分类器的效果最好，几乎能达到百分百正确率判断分类；
4. 当数据不满足朴素贝叶斯条件时，逻辑回归实现的线性分类器有时仍然能得到比较好的结果；

5. UCI 上的实际数据训练显示，逻辑回归实现的线性分类器在现实数据中也能有很好的分类效果。

六、参考文献

[1] https://en.wikipedia.org/wiki/Newton%27s_method

[2] https://en.wikipedia.org/wiki/Gradient_descent

[3] https://www.cnblogs.com/alfred2017/p/6627824.html

[4] Qi L, Sun J. A nonsmooth version of Newton's method[J]. Mathematical Programming, 1993, 58(1-3):353-367.

[5] http://archive.ics.uci.edu/ml/datasets/Fertility

七、附录：源代码（带注释）

已将源代码发送