生物医学数据集机器学习任务

苏智龙

一、生物医学数据集数据概况

选取了胸部X光的照片数据集。

完整数据集：https://data.mendeley.com/datasets/rscbjbr9sj/2

数据集不完整版地址(Kaggle)：https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia

Kaggle版的有5,863张X-Ray照片(JPEG) 。

该数据集标签分为2类：肺炎/正常(Pneumonia/Normal)。

由于设备性能限制，本作业选了其中一部分作为训练集(624张)，测试集为Kaggle上的test全部数据(624张)。

二、机器学习任务

由于数据集里照片的像素大小不一样，而且是RGB格式，有3个通道。所以在进行标准化之前，需要通过openCV把每一张照片的像素大小变的一样。作业中把全部照片变为黑白单通道，224*224像素。

特征数为224*224=50176

训练样本数为624

特征*训练样本数=50176*624=31309824

1、两种数据预处理/标准化

1) 归一化

把每张照片的每个像素值除以255，得到所有值在0-1范围内。

new_value = pixel_value / 255

2) 标准化

把每张照片的每个像素值，先减去该张照片所有像素值的均值，再除以该张照片所有像素值的标准差，化为均值为0，方差为1的数据。

new_value = (pixel_value - mean) / sigma

sigma为该张图片所有像素值的标准差。

3、三种机器学习训练方法

1) 逻辑回归

逻辑回归的代价函数如下：

$$J(\theta) = -\frac{1}{m}[\sum_{i=1}^{m} y^{(i)} \log h_\theta(x^{(i)}) + (1 - y^{(i)}) \log (1 - h_\theta(x^{(i)}))]$$

求出代价函数以后，可以用梯度下降算法来求得能使代价函数最小的参数。

算法为：

Repeat

$$\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta)$$

(simultaneously update all)

}

求导后得到：

Repeat

$$\theta_j := \theta_j - \alpha \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}) x_j^{(i)}$$

(simultaneously update all)

}

用上述算法，不同的学习率、正则化因子和迭代次数得到的不同代价如图1，准确率如图2。

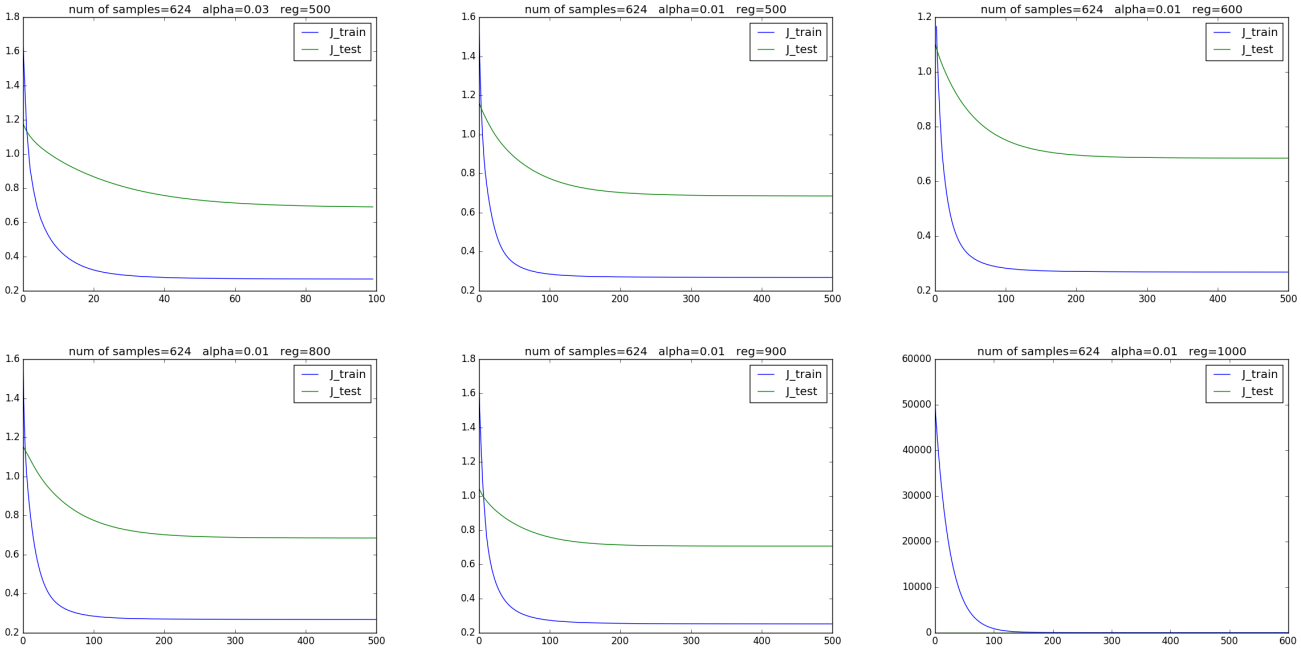

图1 不同学习率、正则化因子和迭代次数得到的代价函数值

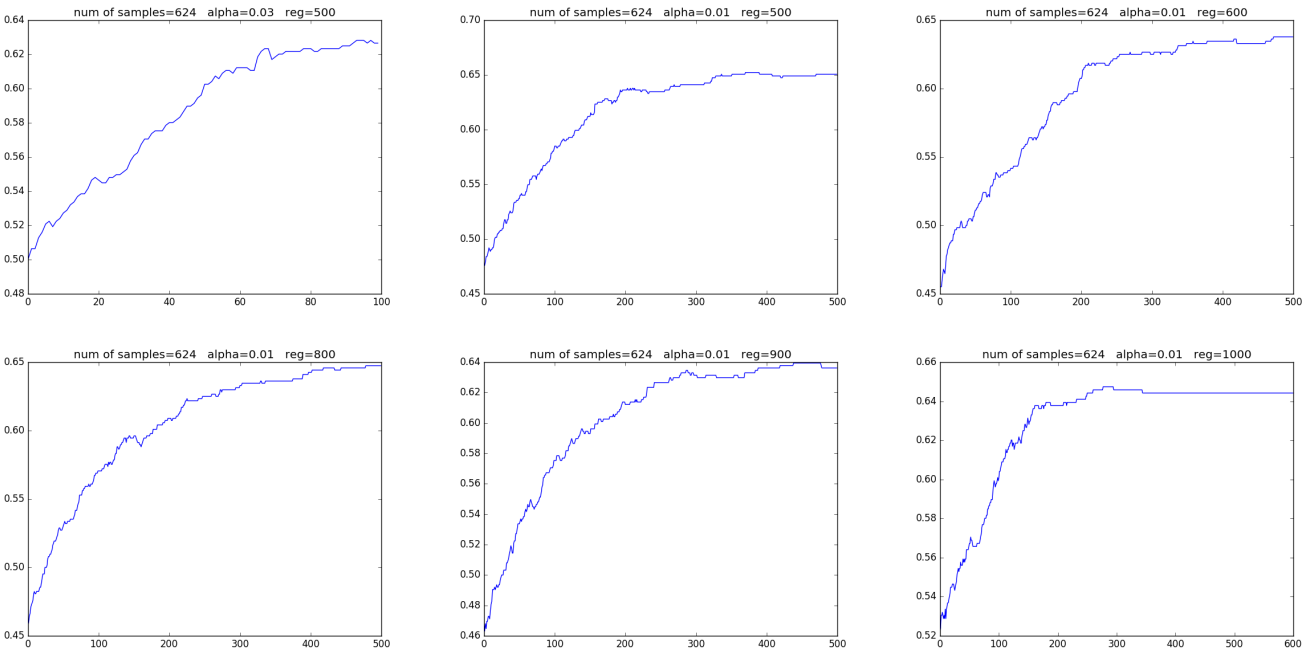

图2 不同学习率、正则化因子和迭代次数得到的准确率

2) SVM

使用python的sklearn库的svm函数实现，核函数使用线形核函数。代码见附件SVM.py。

训练之后的训练集准确率和测试集准确率如图3：

```
accurate of train: 0.990415335463
accurate of test:0.623397435897
```

图3 训练集准确率和测试集准确率

3) 神经网络

网络的一共4层：一层输入层，两层隐藏层，神经元个数分别为10个和5个，和一层输出层，输出结果为0-正常/1-肺炎。

学习率为0.1，迭代次数为1000次。网络的规模可以在代码中的 layer_dims 处修改。

神经网络的代码见附件neuralNetwork.py。

训练的代价函数和训练之后的测试集准确率如图4、图5：

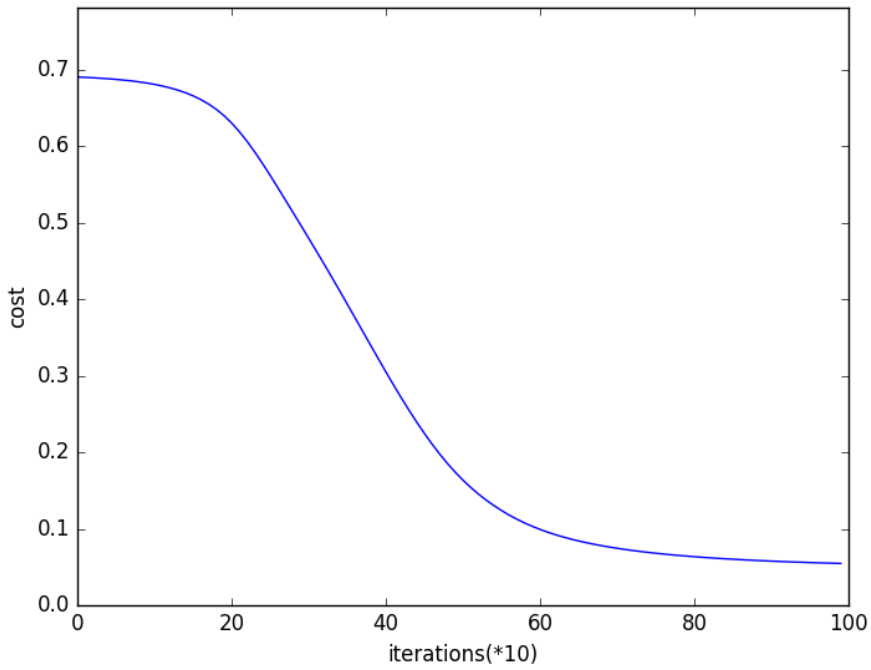

图4 代价函数

```
accurate of test: 0.639423076923
```

图5 测试集准确率

三、结论

三种方法得到的准确率都差不多，在0.62-0.65之间，逻辑回归准确率最好，接近0.65。

在逻辑回归中，正则化因子为800，学习率为0.1效果最好。

神经网络还可以把网络层数、每层的神经元数和学习率作为超参数，达到更好的结果，由于设备运算能力有限，就只选择了其中一种情况作为说明。