机器学习导论第二次作业 - 编程题指南

题目

对数几率回归（Logistic Regression, LR) 的两种实现。

相关说明

你的代码需读取train_feature.csv及train_target.csv两个文件作为traning sets，并以题中所述两种方法实现LR已完成分类任务。val_feature.csv及val_target.csv是所提供的validation set，以供评测（也可以选择自己喜欢的validation形式）。但要求给出题中模型在validation set下的表现结果（ACC/P/R)。在验证集上通过测试的模型需要对test_feature.csv中的样本进行预测，并提交其预测结果。注意，由于本作业是具体算法的实现，请勿使用sklearn库函数（提交样例见**output_example.csv**）

相关数据集：**ML2_programming.zip**

train_feature.csv中每一行表示一个样例的特征，train_target.csv中每一行是该样例的标记（0或1）。运行你的代码前，应从当前目录读取这两个文件作为训练集分别运行两个LR算法，进行模型的训练（注意：模型评估的方法是任意的，不做强制要求，你也可以按照你喜欢的validation方法对已训练LR模型进行评估，但对于题目1中所要求的特殊模型，请在validation set上给出其表现结果）。在获得满意模型后请读取test_feature.csv中的样本并使用该模型对其进行预测，预测结果请以"学号_0或1.csv"为格式输出，其中学号后的数字表示用第几种实现完成的训练。例如：我用第一种实现训练得出的预测结果命名为"DZ1937001_0.csv"。

最终需提交你的代码文件和预测数据文件。两种实现的完整代码请分别以"学号_0或1.py"命名，Jupyter notebook文件(*.ipynb) 亦可作为主程序提交，命名方式相同。代码文件中应包含完整的数据读取、模型训练、模型评估和预测输出部分，请通过注释（#）的方式完成分块。

当然，虽然两种不同的实现原理不同，对同一组test数据的预测结果应是一致的。若有同学的预测结果出现分歧请从debug/实现原理/模型参数几方面考虑。

分题具体说明

1. 任务是完成闭式解实现，并输出validation sets下的性能度量。validation sets指的是val_feature.csv及val_target.csv，分别为validation sets数据集的特征及标签。
2. 任务是通过调整阈值，改进前题所实现的分类器（即闭式解方法），并对test sets结果进行预测并输出。test sets指的是test_feature.csv中的样本，数据仅包含特征。完成此题时应取得**学号_0.py 或 .ipynb**及**学号_0.csv** 。
3. 任务是完成数值方法的实现并输出validation sets下的性能度量。
4. 任务是通过数值方法学得β后输出test sets的预测结果。完成此题时应取得**学号_1.py 或 .ipynb**及**学号_1.csv** 。
5. 提示：从z向量模长及sigmoid函数自身性质考虑。

评分标准

- 10/ 40 仅其中某一种算法实现
- 20/ 40 两种算法实现
- 30/ 40 仅其中某一算法对test的预测结果足够准确
- 40/ 40 两个算法对test的预测结果足够准确
- 45/40   完成附加分

语言及环境

仅接受使用Python编写代码。

Python功能丰富，能够完成诸多任务，在机器学习领域很常用。如果你没有学习过这两种语言，请参考[Learn python in 30 min]，助教无法回答编程语言相关的问题。

可能会帮助你的几个函数

```
import pandas as pd                          #pandas库函数
X_train=pd.read_csv('train_feature.csv')#读取训练集特征赋予变量X_train
X_train['add_column'] = 1                    #在X_train的特征中加如一行常数1
import numpy as np                           #numpy库函数
np.dot(x,y)                                  #<x,y>，x,y的内积
np.linalg.pinv(X)                            #X的伪逆矩阵
np.linalg.norm(z)                            #向量z的L2范数
```

最终提交list

1. 学号_0.py 或 .ipynb #通过闭式解实现
2. 学号_1.py 或 .ipynb #通过数值方法实现
3. 学号_0.csv #通过闭式解学得模型的预测结果
4. 学号_1.csv #通过数值方法学得模型的预测结果

请将上述四个文件与**作业.pdf**打包为**学号_姓名.zip**后上传。