Machine Learning
Assignment 4
(LR & Model Selection and Evaluation & NN)
Due: May, 5

2023 年 4 月 21 日

1 [15pts] AUC

对于有限样例，请证明

$$\mathrm{AUC} = \frac{1}{m^+ m^-} \sum_{x^+ \in D^+} \sum_{x^- \in D^-} \left(\mathbb{I}(f(x^+) > f(x^-)) + \frac{1}{2} \mathbb{I}(f(x^+) = f(x^-)) \right)$$

Proof. 此处用于写证明 *(中英文均可)*

□

2 [40+5 pts] 编程题, Logistic Regression

请结合编程题指南 (见下页) 进行理解

试考虑对率回归与线性回归的关系。最简单的对率回归的所要学习的任务仅是根据训练数据学得一个 $\beta = (\omega; b)$，而学习 β 的方式将有下述两种不同的实现：

0. [闭式解] 直接将分类标记作为回归目标做线性回归，其闭式解为

$$\beta = (\hat{X}^T \hat{X})^{-1} \hat{X}^T y \tag{2.1}$$

, 其中 $\hat{X} = (X; \vec{1})$

1. [数值方法] 利用牛顿法或梯度下降法解数值问题

$$\min_{\beta} \sum_{i=1}^{m} (-y_i \beta^T \hat{x_i} + \ln(1 + e^{\beta^T \hat{x_i}})). \tag{2.2}$$

得到 β 后两个算法的决策过程是一致的，即：

(1) $z = \beta X_i$

(2) $f = \frac{1}{1+e^{-z}}$

(3) 决策函数

$$y_i = \begin{cases} 1, & \text{if } f > \theta \\ 0, & \text{else} \end{cases} \tag{2.3}$$

其中 θ 为分类阈值。回答下列问题：

(1) **[10 pts]** 试实现用闭式解方法训练分类器。若设分类阈值 $\theta = 0.5$，此分类器在 Validation sets 下的准确率、查准率、查全率是多少?

(2) **[10 pts]** 利用所学知识选择合适的分类阈值, 并输出闭式解方法训练所得分类器在 test sets 下的预测结果。

(3) **[10 pts]** 利用数值方法重新训练一个新的分类器。若设分类阈值 $\theta = 0.5$，此分类器在 Validation sets 下的准确率、查准率、查全率是多少?

(4) **[10 pts]** 利用所学知识选择合适的分类阈值，并输出数值方法训练所得分类器在 test sets 下的预测结果。

(5) **[选做][Extra 5 pts]** 谈谈两种方法下分类阈值的变化对预测结果的影响，简要说明看法。

Solution. 此处用于写解答

3 [15pts] Friedman 检验 [编程题]

在数据集 D_1, D_2, D_3, D_4, D_5 运行了 A, B, C, D, E 五种算法，算法比较序值表如表1所示:

表 1: 算法比较序值表

数据集	算法 A	算法 B	算法 C	算法 D	算法 E
D_1	2	3	1	5	4
D_2	5	4	2	3	1
D_3	4	5	1	2	3
D_4	2	3	1	5	4
D_5	3	4	1	5	2
平均序值	3.2	3.8	1.2	4	2.8

使用 Friedman 检验 ($\alpha = 0.05$) 判断这些算法是否性能都相同。若不相同，进行 Nemenyi 后续检验 ($\alpha = 0.05$)，并说明性能最好的算法与哪些算法有显著差别。本题需编程实现 Friedman 检验和 Nemenyi 后续检验。(预计代码行数小于 50 行)

Solution. *此处用于写解答 (中英文均可)*

4 [30pts] BP 算法推导

请给出教材《机器学习》5.3 节 BP 算法的完整推导过程。注意符号的一致性。

Solution. *(中英文回答均可)*

5. ROC & ROC 曲线 [编程题]

（注：本题供同学自行选做，不要求提交）

现有 500 个测试样例，其对应的真实标记和学习器的输出值如表2所示 (完整数据见 data.csv 文件)。该任务是一个二分类任务，1 表示正例，0 表示负例。学习器的输出越接近 1 表明学习器认为该样例越可能是正例，越接近 0 表明学习器认为该样例越可能是负例。

表 2: 测试样例表

样本	x_1	x_2	x_3	x_4	x_5	...	x_{496}	x_{497}	x_{498}	x_{499}	x_{500}
标记	1	1	0	0	0	...	0	1	0	1	1
输出值	0.206	0.662	0.219	0.126	0.450	...	0.184	0.505	0.445	0.994	0.602

(1) 请编程绘制 P-R 曲线

(2) 请编程绘制 ROC 曲线，并计算 AUC

本题需结合关键代码说明思路，并贴上最终绘制的曲线。建议使用 Python 语言编程实现。(预计代码行数小于 100 行)

提示:

- 需要注意数据中存在输出值相同的样例。
- 在 Python 中，数值计算通常使用 Numpy, 表格数据操作通常使用 Pandas, 画图可以使用 Matplotlib (Seaborn), 同学们可以通过上网查找相关资料学习使用这些工具。未来同学们会接触到更多的 Python 扩展库，如集成了众多机器学习方法的 Sklearn, 深度学习工具包 Tensorflow, Pytorch 等。

Solution. *此处用于写解答 (中英文均可)*

机器学习课内实验2 - 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：**ML4_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。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。

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。

请注意：具体提交要求请见助教的说明。