

机器学习导论

作业二

学号, 作者姓名, 邮箱

2018 年 4 月 4 日

1 [25pts] Multi-Class Logistic Regression

教材的章节3.3介绍了对数几率回归解决二分类问题的具体做法。假定现在的任务不再是二分类问题，而是多分类问题，其中标记 $y \in \{1, 2, \dots, K\}$ 。请将对数几率回归算法拓展到该多分类问题。

- (1) [15pts] 给出该对数回归模型的“对数似然”(log-likelihood);
- (2) [10pts] 计算出该“对数似然”的梯度。

提示1: 假设该多分类问题满足如下 $K - 1$ 个对数几率,

$$\begin{aligned}\ln \frac{p(y=1|\mathbf{x})}{p(y=K|\mathbf{x})} &= \mathbf{w}_1^T \mathbf{x} + b_1 \\ \ln \frac{p(y=2|\mathbf{x})}{p(y=K|\mathbf{x})} &= \mathbf{w}_2^T \mathbf{x} + b_2 \\ &\dots \\ \ln \frac{p(y=K-1|\mathbf{x})}{p(y=K|\mathbf{x})} &= \mathbf{w}_{K-1}^T \mathbf{x} + b_{K-1}\end{aligned}$$

提示2: 定义指示函数 $\mathbb{I}(\cdot)$,

$$\mathbb{I}(y=j) = \begin{cases} 1 & \text{若 } y \text{ 等于 } j \\ 0 & \text{若 } y \text{ 不等于 } j \end{cases}$$

Solution. 此处用于写解答(中英文均可)

2 [20pts] Linear Discriminant Analysis

假设有两类数据，正例独立同分布地从高斯分布 $\mathcal{N}(\mu_1, \Sigma_1)$ 采样得到，负例独立同分布地从另一高斯分布 $\mathcal{N}(\mu_2, \Sigma_2)$ 采样得到，其中参数 μ_1, Σ_1 及 μ_2, Σ_2 均已知。现在，我们定义“最优分类”：若分类器在得到测试样例在不同类别的分类概率后，取概率最大的类别作为最终预测的类别输出，则满足“最优分类”性质。

试证明：当两类数据的分布参数 $\Sigma_1 = \Sigma_2 = \Sigma$ 时，线性判别分析(LDA)方法可以达到“最优分类”。（提示：找到定义的最优分类的分类平面。）

Solution. 此处用于写解答(中英文均可)

3 [55+10*pts] Logistic Regression Programming

在本题中，我们将初步接触机器学习编程，首先我们需要初步了解机器学习编程的主要步骤，然后结合对数几率回归，在UCI数据集上进行实战。机器学习编程的主要步骤可参见博客。

本次实验选取UCI数据集Page Blocks（下载链接）。数据集基本信息如表 1所示，此数据集特征维度为10维，共有5类样本，并且类别间样本数量不平衡。

Table 1: Page Blocks数据集中每个类别的样本数量。

标记	1	2	3	4	5	total
训练集	4431	292	25	84	103	4935
测试集	482	37	3	4	12	538

对数几率回归（Logistic Regression, LR）是一种常用的分类算法。面对多分类问题，结合处理多分类问题技术，利用常规的LR算法便能解决这类问题。

- (1) [5pts] 此次编程作业要求使用Python 3或者MATLAB编写，请将main函数所在文件命名为LR_main.py或者LR_main.m，效果为运行此文件便能完成整个训练过程，并输出测试结果，方便作业批改时直接调用；
- (2) [30pts] 本题要求编程实现如下实验功能：
 - [10pts] 根据《机器学习》3.3节，实现LR算法，优化算法可选择梯度下降，亦可选择牛顿法；
 - [10pts] 根据《机器学习》3.5节，利用“一对其余”（One vs. Rest, OvR）策略对分类LR算法进行改进，处理此多分类任务；
 - [10pts] 根据《机器学习》3.6节，在训练之前，请使用“过采样”（oversampling）策略进行样本类别平衡；
- (3) [20pts] 实验报告中报告算法的实现过程（能够清晰地体现(1) 中实验要求，请勿张贴源码），如优化算法选择、相关超参数设置等，并填写表 2，在<http://www.tablesgenerator.com/>上能够方便地制作LaTex表格；
- (4) [附加题10pts] 尝试其他类别不平衡问题处理策略（尝试方法可以来自《机器学习》也可来自其他参考材料），尽可能提高对少数样本的分类准确率，并在实验报告中给出实验设置、比较结果及参考文献；

[注意**]** 本次实验除了numpy等数值处理工具包外禁止调用任何开源机器学习工具包，一经发现此实验题分数为0，请将实验所需所有源码文件与作业pdf文件放在同一个目录下，请勿将数据集放在提交目录中。

实验报告.

Table 2: 算法在测试数据集上泛化性能测试结果，先报告在每个类别上的查全率和查准率，最后报告在整个测试数据集上的准确率。

标记	1	2	3	4	5	准确率
查全率	0.xx	0.xx	0.xx	0.xx	0.xx	0.xx
查准率	0.xx	0.xx	0.xx	0.xx	0.xx	