

# 机器学习导论

## 作业二

151220097, 孙旭东, 248381185@qq.com

2018 年 4 月 8 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] 给出该对率回归模型的“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.** 1. 因为

$$\ln \frac{p(y = i|\mathbf{x})}{p(y = K|\mathbf{x})} = \mathbf{w}_i^T \mathbf{x} + b_i \quad (1.1)$$

所以

$$\frac{p(y = i|\mathbf{x})}{p(y = K|\mathbf{x})} = e^{\mathbf{w}_i^T \mathbf{x} + b_i} \quad (1.2)$$

进一步有

$$\frac{1 - p(y = K|\mathbf{x})}{p(y = K|\mathbf{x})} = \sum_{i=1}^{K-1} e^{\mathbf{w}_i^T \mathbf{x} + b_i} \quad (1.3)$$

所以可得

$$p(y = K|\mathbf{x}) = \frac{1}{1 + \sum_{i=1}^{K-1} e^{\mathbf{w}_i^T \mathbf{x} + b_i}} \quad (1.4)$$

由此可得当  $k \neq K$

$$p(y = k|\mathbf{x}) = \frac{e^{\mathbf{w}_k^T \mathbf{x} + b_k}}{1 + \sum_{i=1}^{K-1} e^{\mathbf{w}_i^T \mathbf{x} + b_i}} \quad (1.5)$$

综上所述，我们有

$$p(y = k|\mathbf{x}) = \begin{cases} \frac{e^{\mathbf{w}_k^T \mathbf{x} + b_k}}{1 + \sum_{i=1}^{K-1} e^{\mathbf{w}_i^T \mathbf{x} + b_i}} & k \neq K; \\ \frac{1}{1 + \sum_{i=1}^{K-1} e^{\mathbf{w}_i^T \mathbf{x} + b_i}} & k = K. \end{cases} \quad (1.6)$$

我们假设总共有  $m$  个样例

$$\begin{aligned} \ell(\mathbf{w}, \mathbf{b}) &= \sum_{i=1}^m \sum_{j=1}^K \mathbb{I}(y_i = j) \ln p(y_i = j|\mathbf{x}_i) \\ &= \sum_{i=1}^m \sum_{j=1}^{K-1} \mathbb{I}(y_i = j) \ln \frac{e^{\mathbf{w}_j^T \mathbf{x}_i + b_j}}{1 + \sum_{t=1}^{K-1} e^{\mathbf{w}_t^T \mathbf{x}_i + b_t}} + \sum_{i=1}^m \mathbb{I}(y_i = K) \ln \frac{1}{1 + \sum_{t=1}^{K-1} e^{\mathbf{w}_t^T \mathbf{x}_i + b_t}} \\ &= \sum_{i=1}^m \sum_{j=1}^{K-1} \mathbb{I}(y_i = j) (\mathbf{w}_j^T \mathbf{x}_i + b_j - \ln(1 + \sum_{t=1}^{K-1} e^{\mathbf{w}_t^T \mathbf{x}_i + b_t})) - \sum_{i=1}^m \mathbb{I}(y_i = K) \ln(1 + \sum_{t=1}^{K-1} e^{\mathbf{w}_t^T \mathbf{x}_i + b_t}) \\ &= \sum_{i=1}^m \sum_{j=1}^{K-1} \mathbb{I}(y_i = j) (\mathbf{w}_j^T \mathbf{x} + b_j) - \sum_{i=1}^m \ln(1 + \sum_{t=1}^{K-1} e^{\mathbf{w}_t^T \mathbf{x}_i + b_t}) \end{aligned} \quad (1.7)$$

再令  $\beta_j = (\mathbf{w}_j; b_j)$  以及  $\hat{\mathbf{x}} = (\mathbf{x}; 1)$ ，可以有

$$\ell(\beta) = \sum_{i=1}^m \sum_{j=1}^{K-1} \mathbb{I}(y_i = j) (\beta_j^T \hat{\mathbf{x}}_i) - \sum_{i=1}^m \ln(1 + \sum_{t=1}^{K-1} e^{\beta_t^T \hat{\mathbf{x}}_i}) \quad (1.8)$$

2. 由上一问可得：

$$\begin{aligned} \frac{\partial \ell(\beta)}{\partial \beta_j} &= \sum_{i=1}^m \mathbb{I}(y_i = j) \hat{\mathbf{x}}_i - \sum_{i=1}^m \frac{e^{\beta_j^T \hat{\mathbf{x}}_i}}{1 + \sum_{t=1}^{K-1} e^{\beta_t^T \hat{\mathbf{x}}_i}} \hat{\mathbf{x}}_i \\ &= \sum_{i=1}^m (\mathbb{I}(y_i = j) - \frac{e^{\beta_j^T \hat{\mathbf{x}}_i}}{1 + \sum_{t=1}^{K-1} e^{\beta_t^T \hat{\mathbf{x}}_i}}) \hat{\mathbf{x}}_i \\ &= \sum_{i=1}^m (\mathbb{I}(y_i = j) - p(y_i = j|\hat{\mathbf{x}}_i)) \hat{\mathbf{x}}_i \end{aligned} \quad (1.9)$$

## 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 $\mu_1, \Sigma_1$ 及 $\mu_2, \Sigma_2$ 均已知。现在，我们定义“最优分类”：若对空间中的任意样本点，分别计算已知该样本采样于正例时该样本出现的概率与已知该样本采样于负例时该样本出现的概率后，取概率较大的所采类别作为最终预测的类别输出，则我们说这样的分类方式满足“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)方法满足“最优分类”性质。（提示：找到满足最优分类性质的分类平面。）

**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。

表 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.**

表 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	