机器学习导论
作业二

171860027, 卫歆, 2523058044@qq.com

2020 年 3 月 29 日

1 [15 pts] Linear Regression

给定数据集 $D = \{(\mathbf{x}_i, y_i)\}_{i=1}^m$，最小二乘法试图学得一个线性函数 $y = \mathbf{w}^*\mathbf{x} + b^*$ 使得残差的平方和最小化，即

$$(\mathbf{w}^*, b^*) = \arg\min_{\mathbf{w},b} \sum_{i=1}^{m} [y_i - (\mathbf{w}\mathbf{x_i} + b)]^2. \tag{1.1}$$

“最小化残差的平方和”与“最小化数据集到线性模型的欧氏距离之和”或是“最小化数据集到线性模型的欧氏距离的平方和”一致吗？考虑下述例子

$$D = \{(-1, 0), (0, 0), (1, 1)\}, \tag{1.2}$$

并回答下列问题。

(1) **[5 pts]** 给出“最小化残差的平方和”在该例子中的解 (w^*, b^*)。

(2) **[5 pts]** 给出“最小化数据集到线性模型的欧氏距离的平方和”在该例子中的数学表达式，并给出其解 (w_E, b_E)，该解与 (w^*, b^*) 一致吗？

(3) **[5 pts]** 给出“最小化数据集到线性模型的欧氏距离之和”在该例子中的数学表达式，(w^*, b^*) 是该问题的解吗？

Solution.

1) 将 *1.2* 代入 *1.1*，化简得 $(\mathbf{w}^*, b^*) = \arg\min_{\mathbf{w},b}[2(w-\frac{1}{2})^2 + 3(b-\frac{1}{3})^2 + \frac{1}{6}]$, $(\mathbf{w}^*, b^*) = (\frac{1}{2}, \frac{1}{3})$.

2) 不一致。$(w_E, b_E) = \arg\min_{\mathbf{w},b} \sum_{i=1}^{m} [(\frac{\mathbf{wx_i}-\mathbf{y_i}+b}{\sqrt{w^2+1}})^2]$, 代入 *1.2* 得 $(w_E, b_E) = \arg\min_{\mathbf{w},b}[\frac{(-w+b)^2+b^2+(w+b-1)^2}{w^2+1}]$

$\frac{\partial d}{\partial w} = \frac{(4w-2)(w^2+1)-2w(2w^2+3b^2-2w-2b+1)}{(w^2+1)^2} = 0$, $\frac{\partial d}{\partial b} = \frac{6b-2}{w^2+1} = 0$. $(w_E, b_E) = (\frac{-2\pm\sqrt{13}}{3}, \frac{1}{3})$.

3) 不是。$(w_E, b_E) = \arg\min_{\mathbf{w},b} \sum_{i=1}^{m} \left|\frac{\mathbf{wx_i}-\mathbf{y_i}+b}{\sqrt{w^2+1}}\right|$, 代入 (w^*, b^*) 计算出值距离之和为 $\frac{4\sqrt{5}}{15}$, 代入 $(\frac{1}{2}, \frac{1}{2})$ 值为 $\frac{\sqrt{5}}{5}$, 小于 $\frac{4\sqrt{5}}{15}$, (w^*, b^*) 不是该问题的解。

2 [40+5 pts] 编程题, Logistic Regression

请结合编程题指南进行理解

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.

1）阈值为 0.5 时，准确率：0.74，查准率：0.67，查全率：1.00.

2）$\frac{y}{1-y} > \frac{m^+}{m^-}$，遍历训练集，带入正例数和反例数，计算得阈值为 0.50833

3）阈值为 0.5 时，准确率：1.00，查准率：1.00，查全率：1.00.

4）选择阈值为 0.50833 进行预测。

5）对于现有的测试数据，相较于闭式解，数值方法下分类阈值的变化对预测结果影响较小。线性回归下闭式解是精确解，取到一个最优阈值后，阈值的改变一定会降低预测的性能。而数值解是近似解，没有函数的精度高，所以预测结果对阈值的变化没有闭式解敏感，存在一定的容错性。

3 [10 pts] Linear Discriminant Analysis

在凸优化中，试考虑两个优化问题，如果第一个优化问题的解可以直接构造出第二个优化问题的解，第二个优化问题的解也可以直接构造出第一个优化问题的解，则我们称两个优化问题是等价的。基于此定义，试证明优化问题 **P1** 与优化问题 **P2** 是等价的。

$$\max_{\mathbf{w}} \frac{\mathbf{w}^\top S_b \mathbf{w}}{\mathbf{w}^\top S_w \mathbf{w}}. \tag{3.1}$$

$$\begin{aligned} \min_{\mathbf{w}} \quad & -\mathbf{w}^\top S_b \mathbf{w} \\ \text{s.t.} \quad & \mathbf{w}^\top S_w \mathbf{w} = 1. \end{aligned} \tag{3.2}$$

Solution.

$\mathbf{w}^\top S_w \mathbf{w} = 1$ 时，令 $\mathbf{w}^\top S_b \mathbf{w} = p \Rightarrow \frac{\mathbf{w}}{\sqrt{\mathbf{p}}}^\top S_b \frac{\mathbf{w}}{\sqrt{\mathbf{p}}} = 1$，设 $\mathbf{x} = \frac{\mathbf{w}}{\sqrt{\mathbf{p}}}$，则 $\mathbf{x}^\top S_b \mathbf{x} = 1, \mathbf{w} = \sqrt{\mathbf{p}}\mathbf{x}$ 是 *3.1* 的解。令 $\mathbf{w}^\top S_b \mathbf{w} = q$ 时取值最小，$\Rightarrow \frac{\mathbf{w}}{\sqrt{\mathbf{q}}}^\top S_b \frac{\mathbf{w}}{\sqrt{\mathbf{q}}} = 1$，则 $\mathbf{w} = \sqrt{\mathbf{q}}\mathbf{x}$ 是 *3.2* 问题的解。这两个解线性相关，即第一个优化问题的解可以直接构造出第二个优化问题的解，第二个优化问题的解也可以直接构造出第一个优化问题的解，所以 *3.1* 和 *3.2* 是等价的。

4 [35 pts] Multiclass Learning

在处理多分类学习问题的时候，我们通常有两种处理思路：一是间接求解，利用一些基本策略 (OvO, OvR, MvM) 将多分类问题转换为二分类问题，进而利用二分类学习器进行求解。二是直接求解，将二分类学习器推广到多分类学习器。

4.1 问题转换

(1) **[5 pts]** 考虑如下多分类学习问题：假设样本数量为 n，类别数量为 C，二分类器对于大小为 m 的数据训练的时间复杂度为 $\mathcal{O}(m)$(比如利用最小二乘求解的线性模型) 时，试分别计算在 OvO、OvR 策略下训练的总时间复杂度。

(2) **[10 pts]** 当我们使用 MvM 处理多分类问题时，正、反类的构造必须有特殊的设计，一种最常用的技术为“纠错输出码”(ECOC)，根据阅读材料 (Error-Correcting Output Codes、Solving Multiclass Learning Problems via Error-Correcting Output Codes[**?**]；前者为简明版，后者为完整版) 回答下列问题：

1) 假设纠错码之间的最小海明距离为 n，请问该纠错码至少可以纠正几个分类器的错误？对于图**??**所示的编码，请计算该纠错码的最小海明距离并分析当两个分类器出错时该编码的纠错情况。
2) 令码长为 8，类别数为 4，试给出海明距离意义下的最优 ECOC 编码，并简述构造思路。
3) 试简述好的纠错码应该满足什么条件？(请参考完整版阅读资料)
4) ECOC 编码能起到理想纠错作用的重要条件是：在每一位编码上出错的概率相当且独立，试分析多分类任务经 ECOC 编码后产生的二类分类器满足该条件的可能性及由此产生的影响。

Class	Code Word							
	f_0	f_1	f_2	f_3	f_4	f_5	f_6	f_7
c_0	0	0	0	0	1	1	1	1
c_1	0	0	1	1	0	0	1	1
c_2	0	1	0	1	0	1	0	1

图 1: 3 类 8 位编码

(3) **[10 pts]** 使用 OvR 和 MvM 将多分类任务分解为二分类任务求解时，试论述为何无需专门这对类别不平衡进行处理。

4.2 模型推广

[10 pts] 对数几率回归是一种简单的求解二分类问题的广义线性模型，试将其推广到多分类问题上，其中标记为 $y \in \{1, 2 \ldots, K\}$。

提示：考虑如下 $K-1$ 个对数几率

$$\ln \frac{p(y=1|\mathbf{x})}{p(y=K|\mathbf{x})}, \ln \frac{p(y=2|\mathbf{x})}{p(y=K|\mathbf{x})}, \cdots, \ln \frac{p(y=K-1|\mathbf{x})}{p(y=K|\mathbf{x})}$$

Solution.

4.1

1）OvO: $\frac{C(C-1)}{2}\mathcal{O}(\frac{2n}{C}) = \mathcal{O}(Cn-n)$

OvR: $C\mathcal{O}(n) = \mathcal{O}(Cn)$

2.1）该纠错码至少可以纠正 (n-1)/2 个分类器的错误。图中最小海明距离为 4，若 f_0 与 f_7 出错，则可以纠错，若 f_0 和 f_7 中有一个出错，也能够纠错，若 f_0 和 f_7 都没错，其他两个分类器出错，则无法纠错。

2.2）海明距离下最优 ECOC 编码应该相互之间最小的海明距离尽量大。先考虑相互距离最大的两行编码，假设距离是 8，那么其他两行与这两行最大的距离只可能有 4；假设最大距离是 7，即有以为相同，那么其他两行与这两行最大的最小距离也是 4；假设最大距离为 6，即有 2 位相同，那么其他两行与这两行最大的最小距离为 2+6/2=5，构造得：

0	*1*	*0*	*1*	*0*	*1*	*1*	*1*
1	*0*	*1*	*1*	*0*	*1*	*0*	*0*
0	*1*	*0*	*0*	*1*	*0*	*0*	*0*
1	*0*	*1*	*0*	*1*	*0*	*1*	*1*

2.3）每一行编码与其他号的编码都应该保持适当的海明距离；每一位的位置函数应该与其他位的函数不相关，所以每一列编码应该与其他列的海明距离尽量大，且与其他列的补码之间的海明距离也较大。

2.4）多分类任务中对不同的分类器提供的训练集不同，分类器对于不是自己训练集的数据不敏感，所以容易出错，但不能保证不同分类器在测试不是自己训练集的数据时出错的概率是相当的。这和数据集中每一类的大小的具体分布情况有关。

3）对 *OvR* 和 *MvM* 来说，由于对每个类进行了相同的处理，其拆解出的二分类任务中类别不平衡的影响会相互抵消，因此通常不需专门处理。

4.2

类似用 *OvR* 的方法。

设

$$\ln \frac{p(y=a|x_i;\beta)}{p(y=K|x_i;\beta)} = (\theta_a^\top - \theta_K^\top)x_i,\ a \in [1, K-1]$$

$$\frac{p(y=a|x_i;\beta)}{p(y=K|x_i;\beta)} = e^{(\theta_a^\top - \theta_K^\top)x_i},$$

$$\sum_{a=1}^{K}(\frac{p(y=a|x_i;\beta)}{p(y=K|x_i;\beta)}) = \frac{1}{p(y=K|x_i;\beta)} = \sum_{a=1}^{K}(e^{(\theta_a^\top - \theta_K^\top)x_i}) = \frac{\sum_{a=1}^{K}(e^{\theta_a^\top x_i})}{e^{\theta_K^\top x_i}},$$

$$p(y=K|x_i;\beta) = \frac{e^{\theta_K^\top x_i}}{\sum_{a=1}^{K}(e^{\theta_a^\top x_i})},$$

可得

$$p(y=j|x_i;\beta) = \frac{e^{\theta_j^\top x_i}}{\sum_{a=1}^{K}(e^{\theta_a^\top x_i})},$$

$$\frac{1-p(y=K|x_i;\beta)}{p(y=K|x_i;\beta)} = \frac{\sum_{a=1}^{K-1}(e^{\theta_a^\top x_i})}{e^{\theta_K^\top x_i}},$$

然后用对数似然和梯度下降法求出每个 θ_i 的最优解。