机器学习导论
作业二

学号, 作者姓名, 邮箱

2020 年 3 月 30 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. *此处用于写解答 (中英文均可)*

(1)

$$设E(w,b)=\sum_{i=1}^{m}[y_i-(\mathbf{wx_i}+b)]^2$$

$$\frac{\partial E(w,b)}{\partial w}=2\left((w\sum_{i=1}^{m}x_i^2-\sum_{i=1}^{m}(y_i-b)x_i)\right)$$

$$\frac{\partial E(w,b)}{\partial b}=2\left(mb-\sum_{i=1}^{m}(y_i-wx_i)\right)$$

令以上两式为 *0* 可以解得

$$w^*=\frac{\sum_{i=1}^{m}y_i(x_i-\bar{x})}{\sum_{i=1}^{m}x_i^2-\frac{1}{m}\left(\sum_{i=1}^{m}x_i\right)^2}=\frac{1}{2}$$

$$b^*=\frac{1}{m}\sum_{i=1}^{m}(y_i-wx_i)=\frac{1}{3}$$

$$其中\bar{x}=\frac{1}{m}\sum_{i=1}^{m}x_i$$

(2)

$$(w_E,b_E)=argmin\sum_{i=1}^{m}\frac{(wx_i+b-y_i)^2}{1+w^2}$$

$$=argmin\left(\frac{3b^2-2b+2w^2-2w+1}{1+w^2}\right)$$

$$b_E=\frac{1}{3}\ w_E=\frac{\sqrt{13}-2}{3}\ 并不一致$$

(3)

$$(w^*,b^*)=argmin\sum_{i=1}^{m}\frac{|wx_i+b-y_i|}{\sqrt{1+w^2}}$$

$$=argmin\left(\frac{|b-w|+|b|+|b+w-1|}{\sqrt{w^2+1}}\right)$$

$$if\ b\geq w\ and\ b\geq 1-w, b^*=\frac{1}{2},\ m^*=\frac{1}{2},\ ans=\frac{\sqrt{5}}{5}$$

$$if\ b\geq w\ and\ b\leq 1-w, b^*=0,\ m^*=0,\ ans=1$$

$$if\ b\leq w\ and\ b\geq 1-w, b^*=0,\ m^*=1,\ ans=\frac{\sqrt{2}}{2}$$

$$if\ b\leq w\ and\ b\leq 1-w, b^*=\frac{1}{2},\ m^*=\frac{1}{2}, ans=ans=\frac{\sqrt{5}}{5}$$

$$\therefore b*=\frac{1}{2}, m^*=\frac{1}{2},\ 并不相同,\ bu$$

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?

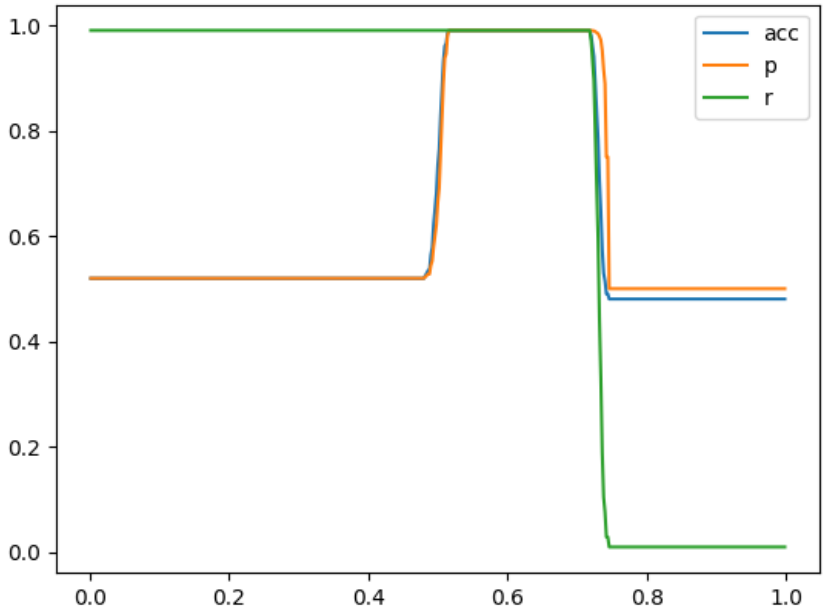
(4) **[10 pts]** 利用所学知识选择合适的分类阈值，并输出数值方法训练所得分类器在 test sets 下的预测结果。

(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. *此处用于写解答 (中英文均可)*

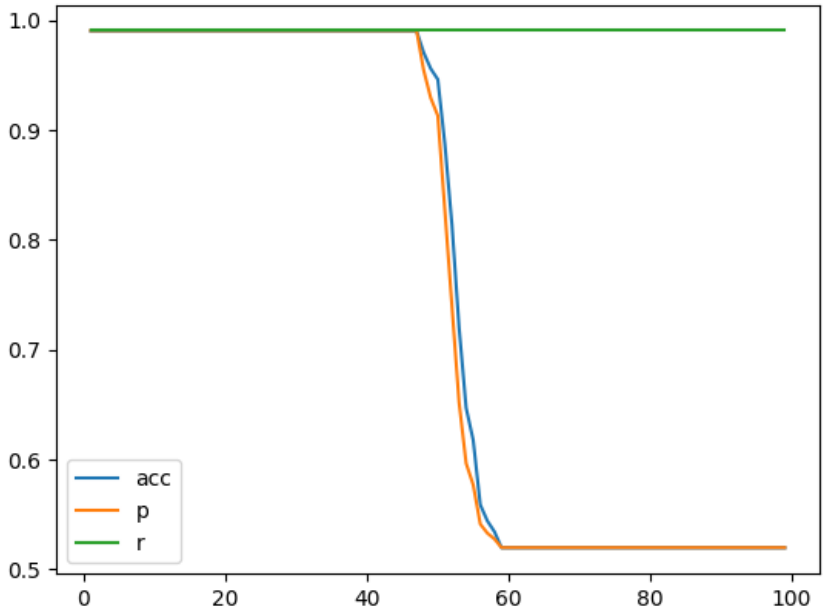
(1) 准确率：0.74，查准率：0.67，查全率：1.0

(2) 在 [0,1] 区间上进行步长为 0.002 的线性搜索，得到最优 θ 值为 0.516，依据为 acc+p+r。完整的度量与 θ 取值关系如图



(3) 本问使用牛顿法。准确率：*1.0*，查准率：*1.0*，查全率：*1.0*

(4) 因为三个率都较高，且在相当大区间内都无变化，于是改为在 *[0,1]* 区间对数尺度上进行线性搜索，横坐标为 *1e-x*, 得到最优 θ 为 *0.05*，判断依据是 *acc+p+r*



(5) 通过在程序中直接增加输出我们可以发现，闭式解方法的 β 的模远远小于牛顿法优化得到的 β。由于 *sigmoid* 函数中含有 e^{-z} 项在分母上，而 $z = \beta X$ 关于不同 X 的变化幅度与 β

模长正相关因此拥有较小模长的闭式解方法输出的 f 分布将相对紧凑，而数值优化得到的 f 分布将更分散，因此改变 θ 的大小，闭式解方法将相对更加敏感，产生的影响更大。

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. 此处用于写解答 (中英文均可)

若 w_1 是问题(*3.1*)的解,则 $\frac{w_1^T}{\sqrt{w_1^T S_w w}} S_w \frac{w_1}{\sqrt{w_1^T S_w w}} = 1$, 且由 $w_1 = \max_{\mathbf{w}} \frac{\mathbf{w}^\top S_b \mathbf{w}}{\mathbf{w}^\top S_w \mathbf{w}}$ 可知 $\frac{w_1^T}{\sqrt{w_1^T S_w w}} S_b \frac{w_1}{\sqrt{w_1^T S_w w}} = max_w(w^T S_b w) = min_w(-w^T S_b w)$. 所以 $\frac{w_1}{\sqrt{w_1^T S_w w}}$ 是问题 (*3.2*) 的解

若 w_2 是问题 (*3.2*) 的解，同理易见其也是问题 (*3.1*) 的解，且任意问题 (*3.1*) 的解 w, aw 也是问题 (*3.1*) 的解

综上，问题 (*3.1*) 与 (*3.2*) 等价

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1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。

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) 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。

(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. 此处用于写解答 *(中英文均可)*

4.1

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

OvR: $T_{OvR} = CO(n) = O(Cn)$

(2) 1) 最小海明距离为 n, 则至少可纠正 $\lfloor \frac{n-1}{2} \rfloor$ 个分类器的错, 因为改变输出编码的任意 $\lfloor \frac{n-1}{2} \rfloor$ 位, 将不能足以使测试样本的海明距离离某个错误分类比原本的正确分类近。

图中 $d_{01} = 4, d_{12} = 4, d_{20} = 4$, 故最小海明距离为 *4*, 当两个分类器出错时, $\lfloor \frac{4-1}{2} \rfloor = 1 < 2$, 因此不能纠错, 但是可以检测到出错。

2) 参考简略版论文, 在类别较小的时候可以采用 *Exhaustive code* 的方式进行设计, 不过这里要求的码长为 *8*, 比标准 *Exhaustive code* 要多一位。构造思路是依次使用 *2* 的降序幂次个 *1* 与 *0* 的间隔序列构造每一个类的编码, 构造的结果为:

c0:11111111 c1:00001111 c2:00110011 c3:01010101

这样构造的 *ECOC* 编码同时兼具下面会具体讲到的 *row seperation* 与 *column separation*, 是最佳的 *ECOC* 码

3) *1.Row Seperation:* 任何一个编码序列使用汉明距离应该良好可分到某个类
2.Column Seperation: 每个位置的编码（每个分类器）与其余位置编码不相关，即与其他位置的编码以及其取反结果的汉明距离应尽可能大

4) 由于保证了 *Column Seperation*, *ECOC* 编码的每一位尽可能不相关, 即每一个分类器尽可能不相关，因此他们在分类出错情况上也应该有较低的相关性，因此能较好的满足每一位上编码出错的概率独立的假设。至于每一位上出错概率，因为使用相同算法所以算法导致的分类误差相当，由于对数据集中的每一个类其被划分的情况是公平的，由划分导致的误差将相互抵消, 因此每一位编码上出错率相当的假设也成立。综上, 使用 *ECOC* 编码后将能较好的提供针对各类的大致相当的正确分类能力和一定纠错能力。

(3) 虽然 *OvO* 与 *MvM* 中确实存在很显然的类别不平衡，不过划分的方法已经确定了这种不平衡对么一个（组）类别都会均等出现，因此类别不平衡产生的影响是相互抵消的，总体上基本是平衡的，因此一般无需专门处理。

4.2 分别训练 *k-1* 个二分类器（或认为 w 与 b 从标量变成了向量），使得 $\ln \frac{p(y=i|\mathbf{x})}{p(y=K|\mathbf{x})} = w_i^T x + b_i, i = 1, 2 \dots k-1$. 则 $p(y=i|x) = e^{w_i^T x + b_i} p(y=K|x)$, 将概率归一化处理产生 $p(y=K|x)$ 后可得 $p(y=i|x) = \frac{e^{w_i^T x + b_i}}{\sum\limits_{j \in K} e^{w_j^T x + b_j}}, i = 1, 2 \dots K$. 此时进行判断的依据即针对输入数据分别求其属于每个类别的概率，并将概率最大的类别作为预测分类输出。至于训练方法，可以采取牛顿法或梯度下降法对 w 与 b 的每一维仍旧采用与二分类类似的优化方法优化得到。