机器学习导论
作业二

学号, 作者姓名, 邮箱

2020 年 4 月 8 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) 该例子中 X, y 为

$$X = \begin{bmatrix} -1 & 1 \\ 0 & 1 \\ 1 & 1 \end{bmatrix}, y = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}. \tag{1.3}$$

从而可得最小二乘法的解为

$$\widehat{\mathbf{w}}^* = (X^\top X)^{-1} X^\top y = \begin{bmatrix} 1/2 \\ 1/3 \end{bmatrix}. \tag{1.4}$$

即 $w^* = 1/2, b^* = 1/3$.

(2) “最小化数据集到线性模型的欧氏距离的平方和”的数学表达式为

$$\min_{w,b} R(w,b) = \frac{1}{w^2+1}\left[(-w+b)^2 + b^2 + (w+b-1)^2\right]. \tag{1.5}$$

令 R 对 w 和 b 的偏导均为 0，可得

$$\begin{cases} 0 = \dfrac{\partial R}{\partial w} = -\dfrac{2w}{(w^2+1)^2}\left[(-w+b)^2 + b^2 + (w+b-1)^2\right] + \dfrac{1}{w^2+1}[4w-2], \\ 0 = \dfrac{\partial R}{\partial b} = \dfrac{1}{w^2+1}[6b-2]. \end{cases} \tag{1.6}$$

求解上述方程组可得

$$\begin{cases} w = (-2 \pm \sqrt{13})/3, \\ b = 1/3. \end{cases} \tag{1.7}$$

可以验证，w 的两个解中 $(-2+\sqrt{13})/3$ 使 R 更小，从而最优解为

$$\begin{cases} w_E = (-2 \pm \sqrt{13})/3, \\ b_E = 1/3. \end{cases} \tag{1.8}$$

该解与 (w^*, b^*) 不一致。

(3) “最小化数据集到线性模型的欧氏距离之和”的数学表达式为

$$\min_{w,b} R(w,b) = \frac{1}{\sqrt{w^2+1}}\left[|-w+b| + |b| + |w+b-1|\right]. \tag{1.9}$$

经计算，有

$$\begin{aligned} R_2(w^*, b^*) = R_2(1/2, 1/3) &= \frac{4\sqrt{5}}{15}, \\ R_2(1/2, 1/2) &= \frac{\sqrt{5}}{5}. \end{aligned} \tag{1.10}$$

由于 $R_2(w^*, b^*) > R(1/2, 1/2)$，从而 (w^*, b^*) 不是该问题的解。

Solution. 助教反馈：

- 注意审题，*2*、*3* 两问要求写出“在该例子中的”表达式
- *3* 问中的优化目标是绝对值函数，通过“在某点偏导不为 *0*”来说明“该点不为最小值点”是错误的，在最小值点偏导也不为 *0*（偏导不存在）
- *3* 问是论述题，给出的结论需要理由支撑

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*)* 闭式解的意思在于把标记作为回归值进行预测，逼近的是 y_i 的值，而并非对率值 $\log \frac{y_i}{1-y_i}$，因此得到回归值后可以使用任意激活函数来完成决策，如 $tanh, sigmoid$。由于把标记作为回归值后 z 的取值更倾向在 $[0, 1]$ 之间，并利用了 $sigmoid$ 函数（注意：在这个问题下选用 $sigmoid$ 函数并非增益操作）在 $z = 0$ 附近梯度较大，即微小误差产生的扰动会对 $sigmoid$ 输出结果产生较大影响，因此此方法需要更加精确的阈值。
数值方法逼近的是对率值 $\log \frac{y_i}{1-y_i}$，z 的范围不仅在 $[0, 1]$ 之间，而 $sigmoid$ 函数在 z 远离 0 时梯度较小，即微小误差对 $sigmoid$ 输出结果影响较小，因此不需要很精确的阈值。

Solution. 助教反馈：这里重点说一下闭式解方法。其实际意义在于把标记作为回归值做预测，因此 *0* 类样本的回归值趋向 *0*，而 *1* 类样本的回归值趋向 *1*。题目中使用了 *sigmoid* 函数作为激活函数是为了强调其实际意义仅仅是把回归值映射到 *[0,1]* 区间内，因此 *0-1* 截断函数应是一个更好的激活函数选择。本题的意图在于希望同学们思考阈值的实际意义再做判断，如 *0* 类样本做闭式解线性回归后其样本中心为 *0*，而被 *sigmoid* 函数映射后则变成了 *0.5*。此时如果选用 *0.5* 作为阈值，期望上会有一半的 *0* 类样本被分为 *1* 类，这也就是为什么要输出 P 和 R，即通过 P 和 R 了解错误发生的原因。闭式解之所以可选用的阈值 *gap* 窄就是因为 *0* 和 *1* 经过 *sigmoid* 激活函数后变为了 *0.5* 和 *0.73*，明显是一个不合理的激活函数。希望同学们以后在碰到类似问题的时候思考阈值的实际意义，而不要随便取 *0.5* 或根据正负样本的数量随意带公式。

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 $\mathbf{w}_1$ 是问题 ***P1*** 的一个解,下证 $\mathbf{w}_2 = \mathbf{w}_1/\sqrt{\mathbf{w}_1^\top S_w \mathbf{w}_1}$ 是为题 ***P2*** 的一个解。否则，存在 $\mathbf{w}_3$，使得

$$-\mathbf{w}_3^\top S_b \mathbf{w}_3 < -\mathbf{w}_2^\top S_b \mathbf{w}_2, \mathbf{w}_3^\top S_w \mathbf{w}_3 = 1. \tag{3.3}$$

从而

$$\frac{\mathbf{w}_3^\top S_b \mathbf{w}_3}{\mathbf{w}_3^\top S_w \mathbf{w}_3} = \mathbf{w}_3^\top S_b \mathbf{w}_3 > \mathbf{w}_2^\top S_b \mathbf{w}_2 = \frac{\mathbf{w}_1^\top S_b \mathbf{w}_1}{\mathbf{w}_1^\top S_w \mathbf{w}_1}, \tag{3.4}$$

与 $\mathbf{w}_1$ 是问题 ***P1*** 的最优解矛盾。

若 $\mathbf{w}_2$ 是问题 ***P2*** 的一个解，下证 $\mathbf{w}_1 = \mathbf{w}_2$ 是问题 ***P1*** 的一个解。否则，存在 $\mathbf{w}_3$，使得

$$\frac{\mathbf{w}_3^\top S_b \mathbf{w}_3}{\mathbf{w}_3^\top S_w \mathbf{w}} > \frac{\mathbf{w}_1^\top S_b \mathbf{w}_1}{\mathbf{w}_1^\top S_w \mathbf{w}_1}. \tag{3.5}$$

从而

$$\begin{aligned} -\left(\frac{\mathbf{w}_3}{\sqrt{\mathbf{w}_3^\top S_w \mathbf{w}_3}}\right)^\top S_b \left(\frac{\mathbf{w}_3}{\sqrt{\mathbf{w}_3^\top S_w \mathbf{w}_3}}\right) < \frac{\mathbf{w}_1^\top S_b \mathbf{w}_1}{\mathbf{w}_1^\top S_w \mathbf{w}_1} = \mathbf{w}_1^\top S_b \mathbf{w}_1 = \mathbf{w}_2^\top S_b \mathbf{w}_2, \\ \left(\frac{\mathbf{w}_3}{\sqrt{\mathbf{w}_3^\top S_w \mathbf{w}_3}}\right)^\top S_w \left(\frac{\mathbf{w}_3}{\sqrt{\mathbf{w}_3^\top S_w \mathbf{w}_3}}\right) = 1, \end{aligned} \tag{3.6}$$

与 $\mathbf{w}_2$ 是问题 ***P2*** 的解矛盾。

Solution. 助教反馈:

- 证明过程中引入的新符号请定义其含义
- 证明题注意逻辑的严谨性
- 证明题“易证”不得分
- 证明题请严格依据题中条件解题，题干没有假定散度矩阵可逆
- '优化问题的解'指的是'使得目标函数最优的变量赋值'，而不是'最优的目标函数值'
- 两个方向的证明不能同理，*P2* 的解可以直接是 *P1* 的解，但 *P1* 的解不一定是 *P2* 的解
- *P1* 中没有 w 大小的约束，不能 $w^\top S_w w = 1$, 应该在构造 *P2* 解的时候处理

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[1]；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。

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

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。

(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

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. 问题转换

(1) ***[10 pts]*** 假设每个类别的数量为 $n_i, i=1,2,\cdots,n$，则

$$\sum_{i=1}^{C} n_i = n$$

对于 *OvO* 策略，总训练时间复杂度为

$$\begin{aligned}
&\sum_{i<j} \mathcal{O}(n_i+n_j) = \frac{1}{2}\sum_{i\neq j} \mathcal{O}(n_i+n_j) = \frac{1}{2}\left(\sum_{i=1}^{C}\sum_{j\neq i} \mathcal{O}(n_i+n_j)\right) \\
=&\frac{1}{2}\left(\sum_{i=1}^{C}\sum_{j=1}^{C} \mathcal{O}(n_i+n_j) - \sum_{i=1}^{C}\sum_{j=i} \mathcal{O}(n_i+n_j)\right) \\
=&(\sum_{i=1}^{C}\sum_{j=1}^{C} \mathcal{O}(n_i) - \sum_{i=1}^{C} \mathcal{O}(n_i) \\
=&\mathcal{O}(Cn) - \mathcal{O}(n) \\
=&\mathcal{O}(Cn)
\end{aligned}$$

对于 *OvR* 策略，总训练时间复杂度为

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

(2) 1) 至少可以纠正 $[\frac{n-1}{2}]$ 个分类器的错误；图示纠错码的最小海明距离为 *4*；当两个分类器预测出错时，该编码可能无法纠错，比如当最终的编码结果为 $(0,0,1,1,1,1,1,1)$ 时，其到 c_0 和 c_1 的海明距离均为 *2*。

2) 根据论文的构造方法 *(图2)*：

我们可以得到如下长度为 *7* 的编码：

When $3 \leq k \leq 7$, we construct a code of length $2^{k-1}-1$ as follows. Row 1 is all ones. Row 2 consists of 2^{k-2} zeroes followed by $2^{k-2}-1$ ones. Row 3 consists of 2^{k-3} zeroes, followed by 2^{k-3} ones, followed by 2^{k-3} zeroes, followed by $2^{k-3}-1$ ones. In row i, there are alternating runs of 2^{k-i} zeroes and ones. Table 6 shows the exhaustive code for a five-class problem. This code has inter-row Hamming distance 8; no columns are identical or complementary.

图 2: 编码构造方法

	f_0	f_1	f_2	f_3	f_4	f_5	f_6
c_0	*1*	*1*	*1*	*1*	*1*	*1*	*1*
c_1	*0*	*0*	*0*	*0*	*1*	*1*	*1*
c_2	*0*	*0*	*1*	*1*	*0*	*0*	*1*
c_3	*0*	*1*	*0*	*1*	*0*	*1*	*0*

第 8 位可编码为前七位任意一位编码的补码。前七位编码的海明距离最小为 4，加入第八位编码后，最小海明距离仍然为 4，此为海明距离下的最优编码。

	f_0	f_1	f_2	f_3	f_4	f_5	f_6	f_7
c_0	1	1	1	1	1	1	1	*
c_1	0	0	0	0	1	1	1	*
c_2	0	0	1	1	0	0	1	*
c_3	0	1	0	1	0	1	0	*

3) 好的纠错码应该满足以下两个条件：一是**行分离**，即不同类别编码的海明距离应该尽可能的大，以保证纠错码可以尽可能纠正较多的分类器预测错误。二是**列分离**，即不同分类器编码之间的海明距离尽可能的大，否则，意味着不同分类器的分类性能相近，因而可能会同时犯错，这会导致纠错码失效。与此同时，希望分类器的编码也其他分类器的编码的反码距离也足够大，因为有些二分类学习器交换输入的类别后训练出的模型是一样的 (比如决策树 C4.5)。通常而言，k 个分类一共可以形成 2^k 个编码，去除其中一半的补码，只剩下 2^{k-1} 个编码，再去除掉全零或者全一的没有意义的编码，只剩下 $2^{k-1}-1$ 个编码可以用。

4) **在每一位编码上出错的概率相当意味着各个分类器的泛化性能相当**，这意味着多分类转换为不同的二分类问题的难易程度相当，但这在现实问题中很难满足。有的二分类问题，两个类别重叠程度较低，则学出的分类器泛化性能更好；有的二分类问题，两个类别重叠程度较高，则学出的分类器泛化性能较差。**在每一位编码上出错的概率独立即不同的分类器彼此间相互独立**。如前所述，k 个分类只有 $2^{k-1}-1$ 个编码可以用，因而当类别数较小时，可用编码数目过少，难以保证分类器之间的独立性；当类别数较多时，可用编码数目较多，更容易保证分类器之间是相对独立的。

(3) 在 OvR 中，对于每一类都进行了相同的处理；在 MvM 中，对每一类进行了相近的处理，其类别不平衡的问题在对二分类问题进行综合的时候会相互抵消掉或者近似抵消掉，因此不需要专门对类别不平衡问题进行处理。

模型推广

考虑

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

立得

$$
\begin{aligned}
p(y=1|\mathbf{x}) &= \frac{e^{\mathbf{w}_1^{\mathrm{T}}\mathbf{x}+b_1}}{\sum_{i=1}^{K-1} e^{\mathbf{w}_i^{\mathrm{T}}\mathbf{x}+b_i}+1} \\
p(y=2|\mathbf{x}) &= \frac{e^{\mathbf{w}_2^{\mathrm{T}}\mathbf{x}+b_2}}{\sum_{i=1}^{K-1} e^{\mathbf{w}_i^{\mathrm{T}}\mathbf{x}+b_i}+1} \\
&\cdots \\
p(y=K-1|\mathbf{x}) &= \frac{e^{\mathbf{w}_{K-1}^{\mathrm{T}}\mathbf{x}+b_{K-1}}}{\sum_{i=1}^{K-1} e^{\mathbf{w}_i^{\mathrm{T}}\mathbf{x}+b_i}+1} \\
p(y=K|\mathbf{x}) &= \frac{1}{\sum_{i=1}^{K-1} e^{\mathbf{w}_i^{\mathrm{T}}\mathbf{x}+b_i}+1}
\end{aligned}
$$

助教反馈：

< 问题转换 > 第一问

• 计算 OvO 复杂度从平均的角度去计算不是很严谨，更加严谨的做法是计算 $\sum_{i<j}(n_i+n_j)$，n_i 是第 i 类的样本数目。

< 问题转换 > 第二问

• 第二小问：前 7 位按照论文的编码构造方式即可，**第 8 位取前面任意一位的补码**（注意：不要取全 0 或者全 1，分类器无意义！也不要和前面 7 位编码的某一位编码一样，造成分类器重复）。

• 第四小问：当**分类数目较多**时，更容易保证每一位编码上出错的概率独立

参考文献

[1] Thomas G Dietterich and Ghulum Bakiri. Solving multiclass learning problems via error-correcting output codes. *Journal of artificial intelligence research*, 2:263–286, 1994.