机器学习导论
作业二

181220031 李惟康 liwk@smail.nju.edu.cn

2020 年 4 月 2 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) 根据线性回归最小二乘法的参数估计的闭式解形式有：

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

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

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

(2) 根据欧氏距离 *(Euclidean Distance)* 的定义我们可将数学表达式重写为：

$$(\mathbf{w_E}, b_E) = \arg\min_{\mathbf{w},b} \frac{\sum_{i=1}^{m}[(\hat{x}_i - x_i)^2 + (\hat{y}_i - y_i)^2]}{\mathbf{w}^2 + 1}$$
$$= \arg\min_{\mathbf{w},b} \frac{\sum_{i=1}^{m}[y_i - (\mathbf{w}\mathbf{x_i} + b)]^2}{\mathbf{w}^2 + 1}$$

将给定数据集代入上式，并分别令损失函数对于$\mathbf{w}$,b的偏导数为0；
因此解得：$w_E = \frac{\sqrt{13}-2}{3}$,$b_E = \frac{1}{3}$。
(3) 根据欧氏距离 *(Euclidean Distance)* 的定义我们可将数学表达式重写为：

$$(\mathbf{w}^*, b^*) = \arg\min_{\mathbf{w},b} \frac{\sum_{i=1}^{m}\sqrt{[(\hat{x}_i - x_i)^2 + (\hat{y}_i - y_i)^2]}}{\sqrt{\mathbf{w}^2 + 1}}$$
$$= \arg\min_{\mathbf{w},b} \frac{\sum_{i=1}^{m}|y_i - (\mathbf{w}\mathbf{x_i} + b)|}{\sqrt{\mathbf{w}^2 + 1}}$$

(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) 使用闭式解方法训练分类器，在分类阈值$\theta = 0.5$的条件下，此分类器在*Validation sets*下的准确率为0.74、查准率为0.67、查全率为1。

(2) 预测结果见附件。

(3) 使用牛顿法训练分类器，在分类阈值$\theta = 0.5$，步长$\gamma = 0.001$的条件下，此分类器在*Validation sets*下的准确率为1、查准率为1、查全率为1。

(4) 预测结果见附件。

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.

考虑问题P1：引入等式约束 $\mathbf{w}^\top S_w \mathbf{w} = 1$，则该问题转化为：

$$\begin{aligned} \max_{\mathbf{w}} \quad & \mathbf{w}^\top S_b \mathbf{w} \\ s.t. \quad & \mathbf{w}^\top S_w \mathbf{w} = 1. \end{aligned}$$

显然该问题的解等价于优化问题P2 的解；

考虑问题P2：显然该问题等价于：

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

根据等式约束易有：$\mathbf{w}\mathbf{w}^T = (S_w)^{-1}$，因此该问题转化为：

$$\min_{\mathbf{w}} \quad -(S_w\mathbf{w})^{-1}(\mathbf{w}^\top)^{-1}\mathbf{w}^\top S_b\mathbf{w} = -\frac{\mathbf{w}^\top S_b\mathbf{w}}{\mathbf{w}^\top S_w\mathbf{w}}$$

故得证

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

[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) 考虑 OvO 策略：这一策略的训练过程当中共需要训练 $\frac{C(C-1)}{2}$ 个分类器，且大小为m的数据训练的时间复杂度为$\mathcal{O}(m)$。注意到每个类别样例均被 $C-1$ 个分类器用作训练样例，则所有分类器的训练集的大小之和为 $(C-1)n$，因此总时间复杂制度为$\mathcal{O}((C-1)m)$。

考虑 OvR 策略：这一策略的训练过程当中共需要训练 C 个分类器，并且每个分类器的训练集大小均为 m，即时间复杂度为$\mathcal{O}(m)$。因此总时间复杂度为$\mathcal{O}(Cm)$。

(2) 1) 由于纠错码之间的最小海明距离为n ，即纠错码之间至少有n位不同。注意到在海明距离的尺度上，每一个分类器的错误都使得我们与正确编码的距离越来越大。若只有$\lfloor \frac{n-1}{2} \rfloor$个错误，则此时距离最近的编码仍是正确编码，即至少可以纠正$\lfloor \frac{n-1}{2} \rfloor$个错误。图1所示的纠错码的最小海明距离为4，因此至少可以纠正1个分类器的错误。

2) 根据阅读材料[1]当中给出的构造方法*(Exhaustive Codes)*，在类别数为4时，可行的编码方式有7种，如下表所示：

Class	Code Word						
	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

当码长为 8 时，在 f_6 之后任意的添加一个编码，即为最优编码。注意到因为此时再加任意的编码都是现有编码的反码，此时，类别之间最小的海明距离仍为4。

3) 根据阅读材料[1]：对于多分类学习问题，好的纠错码应当满足以下两个性质：

- 行分离*(Row separation)*：在海明距离的尺度下，每个码字之间的距离应当足够大；
- 列分离*(Column separation)*：任意两个分类器 $f_i, f_j, (i \neq j)$ 的输出相互之间无关联。这一点可以通过使分类器 f_j 编码与其他分类编码的海明距离足够大实现，且与其他分类编码的反码的海明距离也足够大。

4) *1.* 每位编码出错的概率相当：即每一位上的分类器的泛化误差相同，根据教材P66可知，这个条件取决于样本之间的区分难度，即每个编码拆解后类别之间的差异越相同（即区分难度相近），则满足此条件的可能性越大。但实际情况中很难满足。

2. 出错的可能性相互独立：参考资料[1]中给出一个好的纠错输出码应该满足的其中一个条件就是各个位上分类器相互独立*(Column separation)*，当类别越多时，满足这个条件的可能性越大。

教材上同样介绍了产生的影响：一个理论纠错性质很好、但导致的二分类问题较难的编码，与另一个理论纠错性质差一些、但导致的二分类问题较简单的编码，最终产生的模型性能孰强孰弱很难说。

(3) 根据教材P66有：对 OvR 、 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

一个简单而直接的想法是，对于所有K个可能的分类结果，我们运行$K-1$个独立二元逻辑回归模型，在运行过程中把其中一个类别看成是主类别，然后将其它$K-1$个类别和我们所选择的主类别分别进行回归。同样采取将ω和b吸收入向量形式 $\beta=(\omega;b)$，通过这样的方式，如果选择结果K作为主类别的话，我们可以得到：

$$\ln\frac{P(y=1|\mathbf{x})}{P(y=K|\mathbf{x})}=\beta_1\mathbf{x},\ \ln\frac{P(y=2|\mathbf{x})}{P(y=K|\mathbf{x})}=\beta_2\mathbf{x},\ \cdots,\ \ln\frac{P(y=K-1|\mathbf{x})}{P(y=K|\mathbf{x})}=\beta_{K-1}\mathbf{x} \tag{4.1}$$

化简得到：

$$\begin{aligned}P(y=1|\mathbf{x})&=P(y=K|\mathbf{x})\ e^{\beta_1\mathbf{x}}\\P(y=2|\mathbf{x})&=P(y=K|\mathbf{x})\ e^{\beta_2\mathbf{x}}\\&\cdots\cdots\\P(y=K-1|\mathbf{x})&=P(y=K|\mathbf{x})\ e^{\beta_{K-1}\mathbf{x}}\end{aligned} \tag{4.2}$$

由于多类的概率之和为1，即：$P(y=K|\mathbf{x})=1-\sum_{i=1}^{K-1}P(y=i|\mathbf{x})=1-\sum_{i=1}^{K-1}P(y=K|\mathbf{x})\ e^{\beta_i\mathbf{x}}$

故有：$P(y=K|\mathbf{x})=\dfrac{1}{1+\sum_{i=1}^{K-1}e^{\beta_i\mathbf{x}}}$.

代入(4.2)式可得：

$$\begin{aligned}P(y=1|\mathbf{x})&=\frac{e^{\beta_1\mathbf{x}}}{1+\sum_{i=1}^{K-1}e^{\beta_i\mathbf{x}}}\\P(y=2|\mathbf{x})&=\frac{e^{\beta_2\mathbf{x}}}{1+\sum_{i=1}^{K-1}e^{\beta_i\mathbf{x}}}\\&\cdots\cdots\\P(y=K-1|\mathbf{x})&=\frac{e^{\beta_{K-1}\mathbf{x}}}{1+\sum_{i=1}^{K-1}e^{\beta_i\mathbf{x}}}\end{aligned} \tag{4.3}$$

至此，我们就能计算出所有给定未预测样本情况下得到某个结果的概率。之后可通过极大似然

法与二分类问题类似的来估计 β。
由此，我们便完成了对数几率回归在多分类问题上的推广。

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.