机器学习导论
作业二

171830635, 俞星凯, yuxk@smail.nju.edu.cn

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. *此处用于写解答(中英文均可)*

(1)

$$\begin{aligned}
(w^*, b^*) &= \arg\min_{w,b} (0 + w - b)^2 + b^2 + (1 - w - b)^2 \\
&= \arg\min_{w,b} 2w^2 - 2w + 3b^2 - 2b + 1 \\
&= \arg\min_{w,b} 2(w - \frac{1}{2})^2 + 3(b - \frac{1}{3})^2 + \frac{1}{6} \\
&= (\frac{1}{2}, \frac{1}{3})
\end{aligned}$$

(2)

$$(w_E, b_E) = \arg\min_{w,b} \|0 + w - b\|_2^2 + \|b\|_2^2 + \|1 - w - b\|_2^2$$

(w_E, b_E)与(w^*, b^*)一致。

(3)

$$(w_S, b_S) = \arg\min_{w,b} |0 + w - b| + |b| + |1 - w - b|$$

(w^*, b^*)不是该问题的解，因为$(\frac{1}{2}, \frac{1}{2})$是更优解。

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. *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1.00。*

(2) *选择*$\theta = 0.55$*，准确率、查准率、查全率都是1.0。*

(3) *准确率、查准率、查全率都是1.0。*

(4) *选择*$\theta = 0.55$*，准确率、查准率、查全率仍然都是1.0。*

(5) *sigmoid函数的图像特点是中间陡峭，两边平缓。因为在*y*方向上变化相同长度时，在*x*方向上的变化两边比中间大，所以当阈值在中间0.5附近变化时，对预测结果影响较小；当阈值在0和1附近变化时，对预测结果影响较大。*

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.

设***P1***的解$\mathbf{w}$，则$k\mathbf{w}$也是解，令$k = \sqrt{\mathbf{w}^\top S_w \mathbf{w}}$，则$(k\mathbf{w})^\top S_w (k\mathbf{w}) = 1$，并且$(k\mathbf{w})^\top S_b (k\mathbf{w})$最大，即$-(k\mathbf{w})^\top S_b (k\mathbf{w})$最小，因此$kw$是***P2***的解。

设***P2***的解$\mathbf{w}$，则$k\mathbf{w}$是问题***P3***的解。

$$\begin{aligned} \max_{\mathbf{w}} \quad & \frac{\mathbf{w}^\top S_b \mathbf{w}}{\mathbf{w}^\top S_w \mathbf{w}} \\ s.t. \quad & \mathbf{w}^\top S_w \mathbf{w} = k^2 \end{aligned}$$

因为k是任意的，所以去除约束条件，$\mathbf{w}$是***P1***的解。

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) *OvO*的复杂度$\frac{C(C-1)}{2}\mathcal{O}(\frac{n}{c})\times 2=\mathcal{O}(Cn)$。
*OvR*的复杂度$C\mathcal{O}(n)=\mathcal{O}(Cn)$。

(2) *1)* 可以纠正$\lfloor\frac{n}{2}\rfloor$个分类器的错误。$c_0$和$c_1$的海明距离为*4*，$c_1$和$c_2$的海明距离为*4*，$c_0$和$c_2$的海明距离为*4*，最小海明距离为*4*。当两个分类器出错时，某些编码无法纠错，例如*00101011*，其与c_0、c_1、c_2的海明距离分别为*2*、*2*、*4*。

2) 码长为*8*，类别数为*4*，海明距离意义下的最优*ECOC*编码，第i行由2^{4-i}个*0*和2^{4-i}个*1*交替组成。

	0	*1*	*2*	*3*	*4*	*5*	*6*	*7*
0	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*
1	*0*	*0*	*0*	*0*	*1*	*1*	*1*	*1*
2	*0*	*0*	*1*	*1*	*0*	*0*	*1*	*1*
3	*0*	*1*	*0*	*1*	*0*	*1*	*0*	*1*

3) 行分离，行之间的海明距离较大。列分离，列之间不相关，即海明距离较大。

4) 满足该条件的概率很低，因为不同类别之间的距离不一样，例如手写数字识别中*5*比*1*更“接近”*6*，所以编码出错概率难以相当，也很难保证独立，这产生的影响是实际误差可能远大于*ECOC*编码的理论误差。

(3) 因为对每个进行了相同的处理，其拆解出的二分类任务中类别不平衡的影响会相互抵消。

4.2
对$K-1$个对数几率使用线性模型

$$\ln\frac{p(y=i|\mathbf{x})}{p(y=K|\mathbf{x})}=w_i^Tx+b_i,\quad i=1,2,\cdots,K-1$$

再考虑概率之和为*1*，得到

$$p(y=i|\mathbf{x})=\begin{cases}\frac{e^{w_i^Tx+b_i}}{1+\sum_{j=1}^{K-1}e^{w_j^Tx+b_j}} & i=1,2,\cdots,K-1\\ \frac{1}{1+\sum_{j=1}^{K-1}e^{w_j^Tx+b_j}}, & i=K\end{cases}$$

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.