高级机器学习
作业一

俞星凯 171830635

2020 年 11 月 8 日

学术诚信

本课程非常重视学术诚信规范，助教老师和助教同学将不遗余力地维护作业中的学术诚信规范的建立。希望所有选课学生能够对此予以重视。[1]

(1) 允许同学之间的相互讨论，但是**署你名字的工作必须由你完成**，不允许直接照搬任何已有的材料，必须独立完成作业的书写过程;

(2) 在完成作业过程中，对他人工作（出版物、互联网资料）中文本的直接照搬（包括原文的直接复制粘贴及语句的简单修改等）都将视为剽窃，剽窃者成绩将被取消。**对于完成作业中有关键作用的公开资料，应予以明显引用**；

(3) 如果发现作业之间高度相似将被判定为互相抄袭行为，**抄袭和被抄袭双方的成绩都将被取消**。因此请主动防止自己的作业被他人抄袭。

作业提交注意事项

(1) 请在LaTeX模板中**第一页填写个人的姓名、学号、邮箱信息**；

(2) 本次作业需提交该pdf文件、问题3可直接运行的源码(学号_.py)，将以上两个文件压缩成zip文件后上传。zip文件格式为**学号.zip**，例如170000001.zip；pdf文件格式为**学号_姓名.pdf**，例如170000001_张三.pdf。

(3) 未按照要求提交作业，或提交作业格式不正确，将会**被扣除部分作业分数**；

(4) 本次作业提交截止时间为**11月8日23:59:59**。除非有特殊情况（如因病缓交），否则截止时间后不接收作业，本次作业记零分。

[1] 参考尹一通老师高级算法课程中对学术诚信的说明。

1 [30pts] VC Dimensions

本题探讨 VC 维的性质。

(1) **[10pts]** 请在样本空间 $\mathcal{X} = [0,1]$ 上构造一个有限假设空间 $\mathcal{H}$ 使得 $\mathrm{VC}(\mathcal{H}) = \lfloor \log_2(|\mathcal{H}|) \rfloor$.

(2) **[10pts]** 定义轴平行四边形概念类 $\mathcal{H} = \{h_{(a_1,a_2,b_1,b_2)}(x,y) : a_1 \le a_2 \wedge b_1 \le b_2\}$, 其中

$$h_{(a_1,a_2,b_1,b_2)}(x,y) = \begin{cases} 1 & \text{if } a_1 \le x \le a_2 \wedge b_1 \le y \le b_2 \\ 0 & \text{otherwise} \end{cases}$$

请证明 $\mathcal{H}$ 的 VC 维为 4.

(3) **[10 pts]** 请证明最近邻分类器的假设空间的 VC 维可以为无穷大.

Solution. 此处用于写解答*(*中英文均可*)*

(1) 设$2^n \le |\mathcal{H}| < 2^{n+1}$，将$[0,1]$分为$n$段，第$i$段是区间$[\frac{i-1}{n}, \frac{i}{n})$。从$n$个区间中选出$k$个，分别是第$a_1, a_2, \cdots, a_k$个区间，它们的并区间$t = \cup_{j=1}^{k}[\frac{a_j}{n}, \frac{a_j+1}{n})$。令假设函数$h_t(x) = \mathbb{I}(x \in t)$，可以构造$\sum_{k=0}^{n} C_n^k = 2^n$个这样的假设函数。对于示例集$\{\frac{1}{2n}, \frac{3}{2n}, \cdots, \frac{2n-1}{2n}\}$，其任意一个对分都存在对应假设函数，所以 $\mathcal{H}$ 的 VC 维至少为n。又因为假设函数个数小于2^{n+1}，不可能将大小$n+1$的示例集打散，所以$VC(\mathcal{H}) = n = \lfloor \log_2(|\mathcal{H}|) \rfloor$。

(2) 对实例集$\{(0.5,0.5),(0.5,1.5),(1.5,0.5),(1.5,1.5)\}$， $\mathcal{H}$ 中存在假设

$$\{h_{(0,1,0,1)}, h_{(0,1,0,2)}, h_{(0,1,1,2)}, h_{(0,1,2,3)}, h_{(0,2,0,1)}, h_{(0,2,0,2)}, h_{(0,2,1,2)}, h_{(0,2,2,3)},$$
$$h_{(1,2,0,1)}, h_{(1,2,0,2)}, h_{(1,2,1,2)}, h_{(1,2,2,3)}, h_{(2,3,0,1)}, h_{(2,3,0,2)}, h_{(2,3,1,2)}, h_{(2,3,2,3)}\}$$

将其打散，所以 VC 维至少为*4*。

对于任意大小为*5*的示例集$\{(x_1,y_1),(x_2,y_2),(x_3,y_3),(x_4,y_4),(x_5,y_5)\}$，设$c_1 = \min_i x_i$，$c_2 = \max_i x_i$，$d_1 = \min_i y_i$，$d_2 = \max_i y_i$，一定存在$(x_j,y_j)$，满足$x_j \ne a_1 \wedge x_j \ne a_2$或者$y_j \ne b_1 \wedge y_j \ne b_2$。考虑这样的对分结果，除了$(x_j,y_j)$之外的四个示例标记为*1*，而$(x_j,y_j)$的标记为*0*。显然为了将其余四个示例标记为*1*，$h_{(a_1,a_2,b_1,b_2)}(x,y)$必须满足$a_1 \le c_1 \le c_2 \le a_2 \wedge b_1 \le d_1 \le d_2 \le b_2$，但此时也有$a_1 \le x_j \le a_2 \wedge b_1 \le y_j \le b_2$，所以$(x_j,y_j)$标记只能为*1*， $\mathcal{H}$ 中不存在任何假设能实现上述对分结果。

于是， $\mathcal{H}$ 的 VC 维为*4*。

(3) 对于大小为m的示例集上的任意对分， $\mathcal{H}$ 中都存在一个最近邻分类器h，它的训练集就是该示例集和该对分组成的样例集，于是h一定能产生该对分，即该示例集能被 $\mathcal{H}$ 打散。因为示例集大小m可以任意大，所以 VC 维可以为无穷大。

2 [40pts] Understanding the Parameters in LVW from A Probabilistic Perspective

课程中，我们介绍了一种包裹式特征选择算法Las Vegas Wrapper（简称LVW），该算法的流程如教材中图11.1所示。首先请各位回顾一下LVW算法，接下来我们将对一种特殊情况

$p(T=1)$	$p(T=2)$	$p(T=3)$

下的LVW做一些分析，过程中复习一些简单的概率知识，并从更理性的角度理解LVW中的参数。

现在，我们获得了一个数据集D，该数据集中一共包含N个特征，设特征集合为$A=\{f_1,f_2,\cdots,f_N\}$。我们知道，对于某个具体任务，特征的质量参差不齐，高质量的特征会带来高质量的性能，低质量的特征会带来低质量的性能。我们设第i个特征的质量为2^{i-1}，即N个特征的质量分别为$1,2,4,\cdots,2^{N-1}$。我们规定：给定一个特征子集A'，学习器在A'上的性能恰好等于A'中包含的所有特征的质量之和。设特征子集A'中特征的质量之和记作$Q(A')$。

(1) **[5pts]** 现执行一次图11.1中第6行的语句，产生一个特征子集A'，试求数学期望$\mathbb{E}[Q(A')]$。

(2) **[10pts]** 现在，LVW算法已经执行了一段时间了，当前得到的最优特征子集为B，我们记$R=Q(B)$。设函数$better(n,r)$表示从前n个特征中随机产生一个特征子集且该特征子集的质量大于r的概率。试求$better(N,R)$。这里我们允许以递归式和递归边界来表达$better(N,R)$。

(3) **[10pts]** 从现在开始，我们设$better(n,r)$是一个已知函数。在LVW算法中，当连续T个随机生成的子集不比当前的最优子集更好时，算法结束。我们仍然设当前得到的最优特征子集为B，B的质量为$R=Q(B)$。在本小问中，我们还设T是一个已知参数。设布尔型随机变量e，$p(e=1|T)$表示经过恰好T次循环后LVW算法结束的概率，$p(e=0|T)$表示经过恰好T次循环后LVW算法没有结束的概率。试求分布$p(e|T)$。

(4) **[5pts]** 由第（3）小问可见，参数T越大，LVW算法结束就越（[回答“容易”或“困难”]）；当前最优子集质量越高，LVW算法结束就越（[回答“容易”或“困难”]）。

(5) **[10pts]** 现在，我们引入Bayes观点，认为参数T是服从某个先验分布$p(T)$的随机变量，且是未知的。我们仍然设当前得到的最优子集为B，B的质量为$R=Q(B)$。现在，LVW算法继续执行，在恰好执行了T个循环后成功退出了。如果我们设T的先验分布为整数区间$[1,10]$上的均匀分布，请对T做最大后验估计。这与最大似然估计的结果相同吗？为什么？

(6) **[Bonus 5pts]** 现在，设随机变量T只能在$\{1,2,3\}$中取值，且小明设置了一个先验$p(T)$。小红观察到，当前得到的最优子集为B，B的质量为$R=Q(B)$，且LVW在执行恰好T个循环后成功退出了，于是小红试图对T做最大后验估计。小红发现，T的后验概率在$\{1,2,3\}$上是均匀的，那么小明设置的先验$p(T)$有可能为：（填写一种可能的答案即可）

Solution. *此处用于写解答(中英文均可)*

(1) 设$A'_i\subset A'$有i个特征

$$\mathbb{E}[Q(A')]=\frac{\sum_{i=1}^{N}C_N^i\mathbb{E}[Q(A'_i)]}{2^N-1}=\frac{\sum_{i=1}^{N}C_N^i\frac{i(2^N-1)}{2}}{2^N-1}=\frac{\sum_{i=1}^{N}iC_N^i}{2}=\frac{\sum_{i=1}^{N}NC_{N-1}^{i-1}}{2}=2^{N-2}N$$

(2)

$$better(N,R)=\begin{cases}1 & if\ R\le 0\\ 0 & if\ N\le 0\wedge R>0\\ 0.5\ better(N-1,R-2^{N-1})+0.5\ better(N-1,R) & otherwise\end{cases}$$

(3) $p(e=0|T)=better(N,R)(1-better(N,R))^{T-1}$
$p(e=1|T)=(1-better(N,R))^{T}$

(4) 困难，困难

(5) 因为$P(T|B)=\frac{P(T)P(B|T)}{P(B)}\propto P(T)P(B|T)$，而$T$的先验是均匀分布，于是$P(T|B)\propto P(B|T)$，最大后验与最大似然结果相同。因为$(1-better(N,R))^{T}$单调递减，所以$T$的最大后验估计就是$T$。

3 [30pts] Semi-supervised SVM in practice

参照教材中图13.4所示的TSVM算法，在所提供的半监督数据集上进行训练，报告模型在未标记数据集以及测试集上的性能。

本次实验的数据集为一个二分类的数据集，已提前划分为训练数据和测试数据，其中训练数据划分为有标记数据和无标记数据。数据的特征维度为30，每一维均为数值类型。数据文件的具体描述如下:

- `label_X.csv`,`label_y.csv`分别是有标记数据的特征及其标签。
- `unlabel_X.csv`,`unlabel_y.csv`分别是无标记数据的特征及其标签。
- `test_X.csv`,`test_y.csv`分别是测试数据的特征及其标签。

注意，训练阶段只可以使用`label_X.csv`,`label_y.csv`,`unlabel_X.csv`中的数据，其他的数据只可以在测试阶段使用。

(1) 本次实验要求使用Python3编写，代码统一集中在`tsvm_main.py`中，通过运行该文件就可以完成训练和测试，并输出测试结果。

(2) 本次实验需要完成以下功能:

- **[10pts]** 参照教材中图13.4，使用代码实现TSVM算法。要求:
  1. 不允许直接调用相关软件包中的半监督学习方法。
  2. 可以直接调用相关软件包的SVM算法。
  3. 可以使用诸如`cvxpy`等软件包求解QP问题。
- **[10pts]** 使用训练好的模型在无标记数据和测试数据上进行预测，报告模型在这两批数据上的准确率和ROC曲线以及AUC值。
- **[10pts]** 尝试使用各种方法提升模型在测试集上的性能，例如数据预处理，超参数调节等。报告你所采取的措施，以及其所带来的提升。

Solution. 此处用于写解答*(*中英文均可*)*

sklearn 的 *LinearSVM* 默认损失函数是*Hinge*平方，需要指定*Hinge*损失，其余采用默认设置，准确率、*ROC*曲线和*AUC*如下：

	Label	Unlabel	Test
ACC	0.9035	0.8596	0.9203
AUC	0.9760	0.9508	0.9924

表 1: 原始设置ACC和AUC

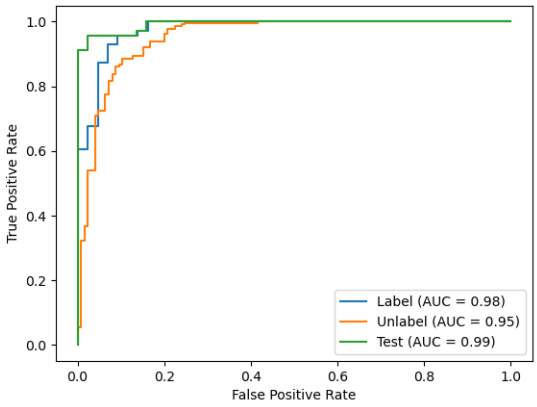

图 1: 原始设置ROC

引入类间平衡设置，几乎所有指标都获得略微提升：

	Label	Unlabel	Test
ACC	0.9386	0.8889	0.9292
AUC	0.9810	0.9519	0.9879

表 2: 类间平衡ACC和AUC

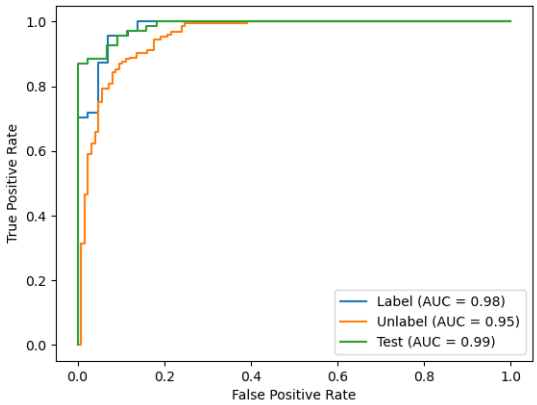

图 2: 类间平衡ROC

加入标准化数据预处理，性能获得大幅度提升：

	Label	Unlabel	Test
ACC	1.0	0.9678	0.9734
AUC	1.0	0.9906	0.9996

表 3: 标准化+类间平衡ACC和AUC

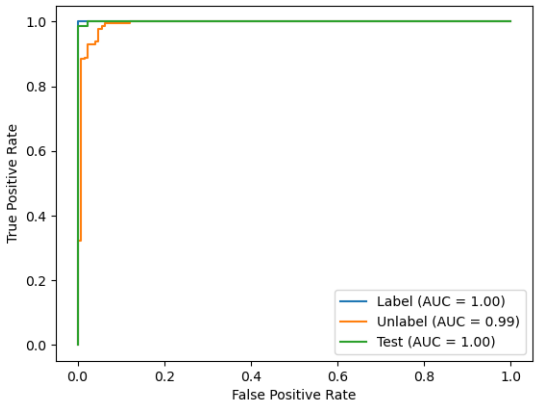

图 3: 标准化+类间平衡ROC