机器学习导论
习题六

学号, 作者姓名, 邮箱

2020 年 5 月 30 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(BoostMain.py, RandomForestMain.py， 不需要提交数据集)， 将以上三个文件压缩成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**6月11日23:59:59。本次作业不允许缓交，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 [25pts] Bayesian Network

贝叶斯网(Bayesian Network)是一种经典的概率图模型，请学习书本7.5节内容回答下面的问题：

(1) **[5pts]** 请画出下面的联合概率分布的分解式对应的贝叶斯网结构：

$$\Pr(A,B,C,D,E,F,G) = \Pr(A)\Pr(B)\Pr(C)\Pr(D|A)\Pr(E|A)\Pr(F|B,D)\Pr(G|D,E)$$

(2) **[5pts]** 请写出图1中贝叶斯网结构的联合概率分布的分解表达式。

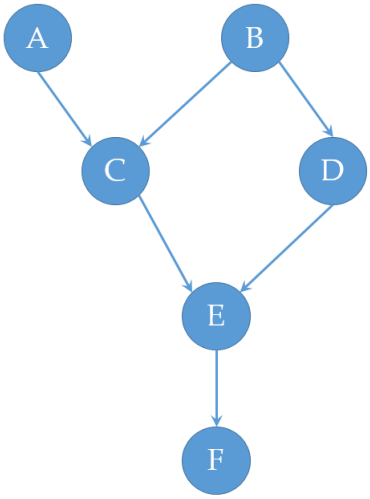

图 1: 题目1-(2)有向图

(3) **[15pts]** 基于第(2)问中的图1, 请判断表格1中的论断是否正确。首先需要作出对应的道德图，并将下面的表格填完整。

表 1: 判断表格中的论断是否正确

序号	关系	True/False	序号	关系	True/False
1	$A \perp\!\!\!\perp B$		7	$F \perp B \mid C$	
2	$A \perp B \mid C$		8	$F \perp B \mid C, D$	
3	$C \perp\!\!\!\perp D$		9	$F \perp B \mid E$	
4	$C \perp D \mid E$		10	$A \perp\!\!\!\perp F$	
5	$C \perp D \mid B, F$		11	$A \perp F \mid C$	
6	$F \perp\!\!\!\perp B$		12	$A \perp F \mid D$	

Solution. 此处用于写解答(中英文均可)

2 [35+10pts] Theoretical Analysis of k-means Algorithm

给定样本集$\mathcal{D} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$, k-means聚类算法希望获得簇划分$\mathcal{C} = \{C_1, C_2, \cdots, C_k\}$使得最小化欧式距离

$$J(\gamma, \mu_1, \ldots, \mu_k) = \sum_{i=1}^{n}\sum_{j=1}^{k} \gamma_{ij} \|\mathbf{x}_i - \mu_j\|^2 \tag{1}$$

其中$\mu_1, \ldots, \mu_k$为k个簇的中心(means), $\gamma \in \mathbb{R}^{n\times k}$为指示矩阵(indicator matrix)定义如下：若$\mathbf{x}_i$属于第j个簇, 则$\gamma_{ij} = 1$, 否则为0，则最经典的k-means聚类算法流程如算法1中所示

Algorithm 1 $k-$means Algorithm

1: Initialize $\mu_1, \ldots, \mu_k$;

2: **repeat**

3: **Step 1:** Decide the class memberships of $\{\mathbf{x}_i\}_{i=1}^n$ by assigning each of them to its nearest cluster center.

$$\gamma_{ij} = \begin{cases} 1, & \|\mathbf{x}_i - \mu_j\|^2 \le \|\mathbf{x}_i - \mu_{j'}\|^2, \forall j' \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

4: **Step 2:** For each $j \in \{1, \cdots, k\}$, recompute μ_j using the updated γ to be the center of mass of all points in C_j :

$$\mu_j = \frac{\sum_{i=1}^n \gamma_{ij}\mathbf{x}_i}{\sum_{i=1}^n \gamma_{ij}} \tag{3}$$

5: **until** the objective function J no longer changes;

(1) **[5pts]** 试证明, 在算法1中, Step 1和Step 2都会使目标函数J的值降低.

(2) **[5pts]** 试证明, 算法1会在有限步内停止。

(3) **[10pts]** 试证明, 目标函数J的最小值是关于k的非增函数, 其中k是聚类簇的数目。

(4) **[15pts]** 记$\hat{\mathbf{x}}$为n个样本的中心点, 定义如下变量,

$$\begin{aligned} T(X) &= \textstyle\sum_{i=1}^n \|\mathbf{x}_i - \hat{\mathbf{x}}\|^2 / n \\ W_j(X) &= \textstyle\sum_{i=1}^n \gamma_{ij} \|\mathbf{x}_i - \mu_j\|^2 / \sum_{i=1}^n \gamma_{ij} \\ B(X) &= \textstyle\sum_{j=1}^k \frac{\sum_{i=1}^n \gamma_{ij}}{n} \|\mu_j - \hat{\mathbf{x}}\|^2 \end{aligned}$$

试探究以上三个变量之间有什么样的等式关系？基于此请证明, k-means聚类算法可以认为是在最小化$W_j(X)$的加权平均, 同时最大化$B(X)$.

(5) **[Bonus 10pts]**在公式1中, 我们使用ℓ_2-范数来度量距离(即欧式距离), 下面我们考虑使用ℓ_1-范数来度量距离

$$J'(\gamma, \mu_1, \ldots, \mu_k) = \sum_{i=1}^{n}\sum_{j=1}^{k} \gamma_{ij} \|\mathbf{x}_i - \mu_j\|_1 \tag{4}$$

- 请仿效算法1,给出新的算法(命名为$k-$means-ℓ_1算法)以优化公式4中的目标函数J'.

- 当样本集中存在少量异常点(outliers)时, 上述的$k-$means-ℓ_2和$k-$means-ℓ_1算法, 我们应该采用哪种算法？即哪个算法具有更好的鲁棒性？请说明理由。

Solution. 此处用于写解答(中英文均可)

3 [40pts] Coding: Ensemble Methods

本次实验中我们将结合两种经典的集成学习思想：Boosting和Bagging，对集成学习方法进行实践。本次实验选取UCI数据集Adult，此数据集为一个二分类数据集，具体信息可参照链接，为了方便大家使用数据集，已经提前对数据集稍作处理，并划分为训练集和测试集，数据集文件夹为adult_dataset。

由于Adult是一个类别不平衡数据集，本次实验选用AUC作为评价分类器性能的评价指标，可调用sklearn算法包对AUC指标进行计算。

(1) 本次实验要求使用Python3编写，要求代码分布于两个文件中，BoostMain.py, RandomForestMain.py ，调用这两个文件就能完成一次所实现分类器的训练和测试；

(2) **[35pts]** 本次实验要求编程实现如下功能：

- **[10pts]** 结合教材8.2节中图8.3所示的算法伪代码实现AdaBoost算法，基分类器选用决策树，基分类器可调用sklearn中决策树的实现；
- **[10pts]** 结合教材8.3.2节所述，实现随机森林算法，基分类器仍可调用sklearn中决策树的实现，也可以手动实现，在实验报告中请给出随机森林的算法伪代码；
- **[10pts]** 结合AdaBoost和随机森林的实现，调查基学习器数量对分类器训练效果的影响，具体操作如下：分别对AdaBoost和随机森林，给定基分类器数目，在训练数据集上用5折交叉验证得到验证AUC评价。在实验报告中用折线图的形式报告实验结果，折线图横轴为基分类器数目，纵轴为AUC指标，图中有两条线分别对应AdaBoost和随机森林，基分类器数目选取范围请自行决定；
- **[5pts]** 根据参数调查结果，对AdaBoost和随机森林选取最好的基分类器数目，在训练数据集上进行训练，在实验报告中报告在测试集上的AUC指标；

(3) **[5pts]** 在实验报告中，除了报告上述要求报告的内容外还需要展现实验过程，实验报告需要有层次和条理性，能让读者仅通过实验报告便能了解实验的目的，过程和结果。

实验报告.