DSCI 6001P 数据科学基础

作业 3 集成、关联规则、贝叶斯、神经网络

提交截止日期：11.30 号晚上 24 点之前
提交方式：电子版发送至 卢嘉栋 (ljd_pb19000334@mail.ustc.edu.cn)
邮件主题和 PDF 命名格式：HW3-姓名-学号，如 HW3-张三-SA24123123

1. 考虑下表的购物篮事务：

事务 ID	购买项
1	{牛奶，啤酒，尿布}
2	{面包，黄油，牛奶}
3	{牛奶，尿布，饼干}
4	{面包，黄油，啤酒}
5	{啤酒，饼干，尿布}
6	{牛奶，尿布，面包，黄油}
7	{面包，黄油，尿布}
8	{啤酒，尿布}
9	{牛奶，尿布，面包，啤酒}
10	{啤酒，饼干}

(1) 从这些数据中，能够提取出的关联规则的最大数量是多少（包括零支持度的规则）？
(2) 能够提取的频繁项集的最大长度是多少（假定最小支持度>0）？
(3) 写出从该数据及中能够提取的 3-项集的最大数量的表达式。
(4) 找出具有最大支持度的项集（长度为 2 或更大）。

2. 数据库有 5 个事务。设 min_sup =60% ， min_conf = 80%。

TID	购买的商品
T100	|M, O, N, K, E, Y|
T200	|D, O, N, K, E, Y|
T300	|M, A, K, E|
T400	|M, U, C, K, Y|
T500	|C, O, O, K, I, E|

(a)分别使用 Apriori 算法和 FP-growth 算法找出频繁项集。比较两种挖掘过程的有效性。
(b)列举所有与下面的元规则匹配的**强**关联规则(给出支持度 s 和置信度 c)，其中，X 是代表顾客的变量，$item_i$是表示项的变量(如“A”，“B”等)：

$$\forall x \in \text{transaction}, \text{buys}(X, \text{item}_1) \wedge \text{buys}(X, \text{item}_2) \Rightarrow \text{buys}(X, \text{item}_3) \ [s, c]$$

3. 集成学习：
(1) 试析随机森林为何比决策树 Bagging 集成的训练速度更快？

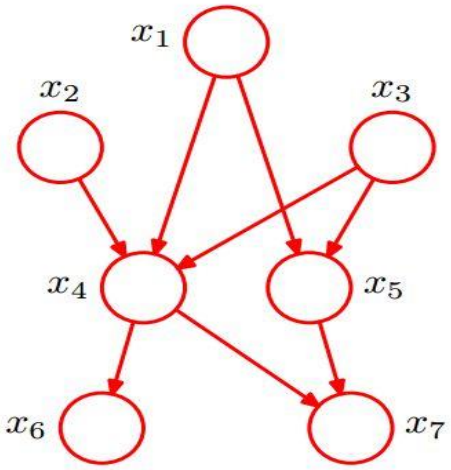
(2) 集成学习中多样性增强的方法有哪些？分别阐述这些方法适用的前提。
(3) Bagging 能否提升朴素贝叶斯分类的性能？为什么？
(4) 分析 GradientBoosting [Friedman, 2001]和 AdaBoost 的异同？

4. 给定一个贝叶斯网络如下图所示：

（1） 在给定x_1，x_3的情况下，x_5，x_6是条件独立的吗？

（2） 在给定x_2，x_3的情况下，x_5，x_6是条件独立的吗？

（3） 写出x_1，x_2,…x_7的联合概率分布



https://blog.csdn.net/weixin_43499292/article/details/118733376

5. 试由下表的训练数据学习一个朴素贝叶斯分类器并确定 x = （2，S，T）的类判别结果 y。表中 X（1）， X（2） X（3）为特征，Y 为类标记。

	1	2	3	4	5	6	7	8	9	10
$X^{(1)}$	1	1	1	2	2	1	2	2	3	3
$X^{(2)}$	S	M	M	S	S	L	M	M	L	S
$X^{(3)}$	T	T	F	F	F	T	F	T	T	F
Y	-1	-1	1	1	-1	-1	-1	1	1	1

6. 给定如下表所示训练数据。假设每一个个体学习器由 x（输入）和 y（输出）产生，其阈值 v（判定正反例的分界线）使该分类器在训练数据集上分类误差率最低。（y=1 为正例，y=-1 为反例），请使用 Adaboost 算法集成多个个体学习器，得到最终的分类器。

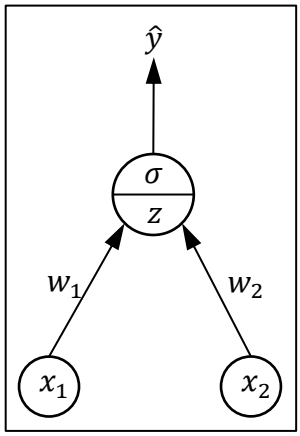
序号	1	2	3	4	5	6	7	8	9	10
x	0	1	2	3	4	5	6	7	8	9
y	1	1	1	-1	-1	-1	1	1	1	-1

https://blog.csdn.net/zzqingyun/article/details/124649176

7. 请证明：Adaboost 算法是前向分步算法的特例。这时，模型是由基本分类器组成的加法模型，损失函数是指数函数。

8. 已知二维空间中的 3 个点 $x_1 = (1,1)^T$, $x_2 = (5,1)^T$, $x_3 = (4,4)^T$, 试求在 p 取不同值时，L_p 距离下 x_1 的最近邻点。

9. [神经网络-链式法则] 一个简单的神经网络如右图所示。其中 x_1 和 x_2 为输入，w_1 和 w_2 为模型权重（参数），σ 表示激活函数，模型输出为 $\hat{y}$。



假设激活函数为 Sigmoid 函数，损失函数为二元交叉熵损失函数。前向传播过程可以用以下公式表示（y为真实标签）：

$z = w_1x_1 + w_2x_2$ ，

$\hat{y} = \sigma(z) = \frac{1}{1+e^{-z}}$ ，

$L(y,\hat{y}) = -ylog(\hat{y}) - (1-y)log(1-\hat{y})$ 。

请基于链式法则，推导出损失函数 L 对模型权重 w_1 和 w_2 的偏导数，即 $\frac{\partial L}{\partial w_1}$ 和 $\frac{\partial L}{\partial w_2}$ 的表达式。（结果尽量用 x_1，x_2，y 和 $\hat{y}$ 表示）

10. [神经网络-CNN 卷积计算] 已知一个灰度图像，其像素值可表示为如下 4×4 矩阵：

$$\begin{bmatrix} 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 \\ 0 & -2 & 0 & -2 \end{bmatrix}$$

将该图像输入到一个卷积层中。卷积层的参数设定为：步幅（stride）为 1，无填充（no padding），且无偏置（no bias）。卷积核（kernel）大小为 3×3。

（1）当卷积核（kernel）为 $\begin{bmatrix} -\frac{1}{2} & 1 & -\frac{1}{2} \\ -\frac{1}{2} & 1 & -\frac{1}{2} \\ -\frac{1}{2} & 1 & -\frac{1}{2} \end{bmatrix}$ 时，计算输出矩阵；

（2）当卷积核（kernel）为 $\begin{bmatrix} -\frac{1}{2} & -\frac{1}{2} & -\frac{1}{2} \\ 1 & 1 & 1 \\ -\frac{1}{2} & -\frac{1}{2} & -\frac{1}{2} \end{bmatrix}$ 时，计算输出矩阵。