Web 信息处理与应用：课后作业 3

数据准备 + 分类 + 聚类部分

请于 2020 年 12 月 31 日前将作业电子版发送至课程邮箱：ustcweb2019@163.com

邮件标题与作业文件命名为：PBXXXXX_XXX(姓名)_HW3

1 计算题

1.1 考虑下表中的事务性数据集：

Customer ID	Transaction ID	Items Bought
1	0001	$\{a, d, e\}$
1	0024	$\{a, b, c, e\}$
2	0012	$\{a, b, d, e\}$
2	0031	$\{a, c, d, e\}$
3	0015	$\{b, c, e\}$
3	0022	$\{b, d, e\}$
4	0029	$\{c, d\}$
4	0040	$\{a, b, c\}$
5	0033	$\{a, d, e\}$
5	0038	$\{a, b, e\}$

1) 每个事务 ID 对应一条事务，计算{e}，{b, c}，{b, c, e}的支持度。
2) 使用（1）的计算结果，计算关联规则{b, c}→{e}和{e}→{b, c}的置信度。
3) 从（2）的结果看，置信度是对称的吗？请根据计算公式分析其对称性。

1.2 考虑如下二元分类的数据集：

User interest	User occupation	Click
Tech	Professional	1
Fashion	Student	0
Fashion	Professional	0
Sports	Student	0
Tech	Student	1
Tech	Retired	0
Sports	Professional	1

1) 计算分别以属性 User interest 和 User occupation 划分时的信息增益。构建决策树将会选择哪个属性？
2) 计算分别以属性 User interest 和 User occupation 划分时的 Gini 指数。构建决策树将会选择哪个属性？

1.3 已知正例点 $x_1 = (2.5, 2.5)^T$，$x_2 = (5, 2)^T$，和负例点 $x_3 = (1.5, 1.5)^T$，试用 SVM 对其进行分类，求最大间隔分离超平面，并指出所有的支持向量。

2 问答题（言之有理即可）

2.1 主成分分析的基本流程是什么？与特征值有何关系？：

2.2 如果从信息检索的视角，可以将寻找最近邻的过程视作检索最相关的 K 个文档的过程。那么，这一过程是否可以利用倒排索引的思路加以实现？如何实现？

2.3 无论是 K 最近邻分类还是 K 均值聚类，都涉及到 K 的取值问题。请简述两个问题各自选取合适 K 值的思路，并比较两者在思路上有何不同？

2.4 K-mediods 算法描述:

a) 首先随机选取一组聚类样本作为中心点集

b) 每个中心点对应一个簇

c) 计算各样本到各个中心点的距离(如欧几里得距离)，将样本点放入距离中心点最短的那个簇中

d) 计算各簇种，据簇内各样本点距离的绝对误差最小的点，作为新的中心点

e) 如果新的中心点集和原中心点集相同，算法中止；如果新的中心点集与原中心点集不完全相同，返回 b)

试着:

a) 阐述 K-mediods 算法和 K-means 算法相同的缺陷

b) 阐述 K-mediods 算法相比于 K-means 算法的优势

c) 阐述 K-mediods 算法相比于 K-means 算法的不足