**中国科学技术大学 计算机学院**

**2016级研究生《机器学习与知识发现》期末考试试题**

姓名： 学号： 成绩：

1. 请画出训练误差和测试误差随模型复杂度变化而变化的曲线图，并解释Ockham’s Razor（奥卡姆剃刀原则）。（5分）
2. (朴素贝叶斯)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。（10分）

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 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 |

三． (决策树)如下表数据，前四列是天气情况（阴晴outlook，气温temperature，湿度humidity，风windy），最后一列是类标签，表示根据天气情况是否出去玩。（15分）

1. “信息熵”是度量样本集合纯度最常用的一种指标，假定当前样本集合D中第k类样本所占的比例为 （k=1, 2, …, K）,请问当什么条件下，D的信息熵Ent(D)取得最大，最大值为多少？
2. 根据表中训练数据，基于信息增益决策树应该选哪个属性作为第一个分类属性？
3. 对于含有连续型属性的样本数据，决策树和朴素贝叶斯分类能有哪些处理方法？
4. 在分类算法的评价指标中，recall和precision分别是什么含义？
5. 若一批数据中有3个属性特征，2个类标记，则最多可能有多少种不同的决策树？（不同决策树指同一个样本在两个决策树下可能得到不同的类标记）

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| outlook | temperature | humidity | windy | play |
| sunny | hot | high | FALSE | no |
| sunny | hot | high | TRUE | no |
| rainy | cool | normal | TRUE | no |
| sunny | mild | high | FALSE | no |
| sunny | cool | normal | FALSE | yes |
| rainy | mild | normal | FALSE | yes |
| overcast | cool | normal | TRUE | yes |
| rainy | cool | normal | FALSE | yes |
| rainy | mild | high | FALSE | yes |
| overcast | hot | high | FALSE | yes |

四、(SVM) 已知正例点x1 = (1, 2)T，x2 = (2, 4)T，负例点x3 = (2, 1)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。（10分）

五、(关联规则)考虑下表的购物篮事务：（12分）

|  |  |
| --- | --- |
| 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）。

1. (集成学习)回答以下问题（10分）
2. 试析随机森林为何比决策树Bagging集成的训练速度更快？
3. 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。

七．(聚类分析)回答以下问题（14分）

1. 领导者算法用一个点（称作领导者）代表一个簇，并将每个点指派到最近的领导者对应的簇，除非距离大于用户指定的阈值。在那样的情况下，改点成为一个新簇的领导者。与K均值相比，领导者算法的优点和缺点是什么？
2. 使用如下表中的相似度矩阵进行单链和全链层次聚类。绘制树状图显示结果，树状图应清楚地显示合并的次序。

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | P1 | P2 | P3 | P4 | P5 |
| P1 | 1.00 | 0.10 | 0.41 | 0.25 | 0.35 |
| P2 | 0.10 | 1.00 | 0.64 | 0.47 | 0.98 |
| P3 | 0.41 | 0.64 | 1.00 | 0.44 | 0.85 |
| P4 | 0.25 | 0.47 | 0.44 | 1.00 | 0.76 |
| P5 | 0.35 | 0.98 | 0.85 | 0.76 | 1.00 |

八．(降维)回答以下问题（8分）

1. 采用PCA方法降维后，有部分信息被舍弃了，简述为什么舍弃这部分信息是必要的？
2. 对于高维空间到低维空间的函数映射不是线性关系的情况，可以采用什么方法进行降维？请阐述其核心思想。

九．综合题（16分）

口碑是阿里巴巴集团与[蚂蚁金服](http://baike.baidu.com/view/15171507.htm" \t "_blank)集团整合双方资源，联手打造的一家互联网本地生活服务平台。与淘宝、天猫不同，口碑面向的是线下实体商家。在推广口碑的过程中，我们遇到了一些商家推荐上的困难，因此需要你的帮助。

现有数据如下表：

|  |  |
| --- | --- |
| 数据集名称 | 具体数据项 |
| 在线用户行为 | 用户id，线上商家id，购买商品id，商品类别，在线动作（0表示点击，1表示购买），行为时间 |
| 用户线下行为 | 用户id，线下商家id，地理位置，行为时间 |
| 商家信息 | 商家id，商家可接纳的最高人数，商家地理位置 |

我们需要在特定的位置给用户推荐线下商家，商家个数可以是多个。由于口碑的用户使用记录较少，这给我们的商家推荐工作带来了一定的困难。而阿里巴巴拥有丰富的淘宝、天猫用户在线行为记录，我们希望能通过用户的在线行为数据来帮助预测其线下行为，改善线下推荐效果。

(1) 对用户的线下商家推荐，可以使用哪些机器学习算法？

(2) 你觉得用户的在线行为数据对线下行为是否能帮助线下行为的预测（绝大部分口碑用户有对应的线上行为记录）？请阐述理由。如果可以，请简要给出方案。

(3) 你认为哪些数据对推荐结果有较大影响？