**中国科学技术大学 计算机学院**

**2013级研究生《机器学习与知识发现》期末考试试题**

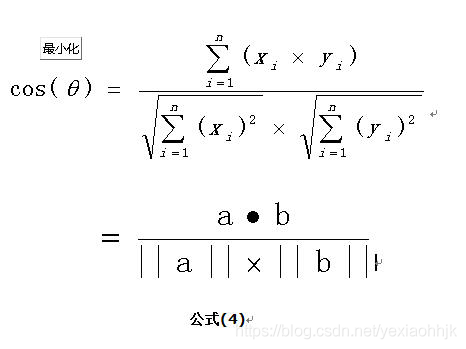
**姓名：\_\_\_\_\_\_\_\_\_\_\_\_ 学号：\_\_\_\_\_\_\_\_\_\_\_\_\_ 成绩：\_\_\_\_\_\_\_\_\_\_\_**

1. 数据对象间的相似性度量计算 （8分）
2. a=(1, 1, 0, 1, 1, 0), b=(0,1,1,0,1,0), c=(1,1,3,3), d=(3,3,1,1), 计算
3. a, b的Jaccard相似系数（Jaccard Cofficient）
4. c, d 向量空间余弦相似度（Cosine Similarity）
5. c, d 的皮尔森相关系数（Pearson Correlation Cofficient）

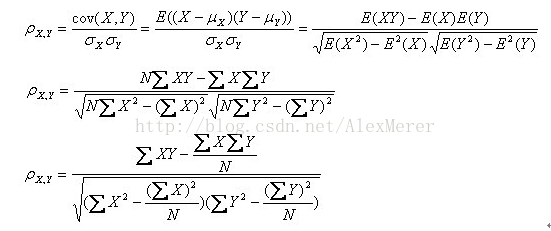
解：



=3/5



=



1. 以上三种度量方法，哪些方法比较适合度量文本数据的相似性？ 简要说明。

答： Jaccard 和 Cosine。。。

1. 考虑下图的决策树。假设产生决策树的数据集包含16个二元属性三个分类C1、C2和C3。根据最小描述长度原则（MDL）计算每棵决策树的总描述长度。

* 树的整体描述长度由下式给出：

Cost(tree, data) = Cost(tree) + Cost(data | tree)

* 树的每个内部节点用划分属性的ID进行编码。如果有m个属性，为每个属性编码的代价是log2(m)个二进位。
* 每个叶节点使用与之相关联的类的ID编码。如果有k个类，为每个类编码的代价是log2(k)个二进位。
* Cost(tree) 是对决策树的所有结点编码的开销。为了简化计算，可以假设决策树的总开销是对每个内部结点和叶结点编码开销的总和。
* Cost( data|tree) 是对决策树在训练集上分类错误编码的开销。每个错误用log2(n)个二进位编码，其中n是训练实例的总数。

根据MDL原则，哪棵决策树更好？（10分）

决策树a的总代价：2× 4+3× 2+7× log2 n = 14+7 log2 n.

决策树b的总代价： 4× 4+5× 2+4×log2 n = 26+4 log2 n.

根据MDL原则，

若n<16。 树a好；

若n>16. 树b好；

若N=16，一样好

.

1. 请简述构建组合（集成）分类器的几种方法，并说明集成分类器能够改善分类器性能的原因。（12分）

书上内容

P172

Boosting、Bagging

什么集成方法会好呢？可以从下面三个方面进行考虑

1.数据量问题（statistical）

学习的过程可以看做是在一个假设空间H中寻找一个最优的假设，如果训练集的数据量特别小的时候，由于训练数据不充分，可以学习到很多不同的假设，而这些假设在训练集上的正确率确是相同的，此时就很难抉择哪个假设在测试集上会得到好的结果了。通过集成这些假设就可以减少选错分类器的风险。

2. 计算问题（computational）

很多学习算法都会遇到局部最优的这种状况，例如，神经网络是通过梯度下降来最小化错误率的，决策树算法是通过贪婪分裂的规则去扩展决策树的，假如训练集是足够充足的，学习算法也很难得到全局最优解的。通过把从不同起始点得到的分类器集成的方法可以更好的靠近全局最优解。

3.假设描述问题（representional）

大部分的应用中，机器学习算法很难用假设空间H中的假设去表达真实函数f，通过对假设空间H中的假设进行加权进而扩展假设空间H的规模或许能够表示真实函数f。

集成分类器能解决单个分类器的训练数据量小，假设空间小，局部最优的三个问题，所以，集成分类器的预测能力会优于单个分类器的预测能力。

1. 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。（14分）
2. 根据上述训练数据，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. 请画出两层决策树模型。
4. 使用朴素贝叶斯方法预测测试样本（outlook=rainy, temperature=cool, humidity=normal, windy=FALSE）的类标号。

P(no|x)=2/5 \* 1/4 \* 1/4 \*1/4 \*1/2 = 1/320

P(yes|x)=3/5 \* 3/6\* 3/6 \*4/6\*5/6= 1/4

预测为yes

1. 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？离散化（等区间离散、等数据离散），高斯分布模拟

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

1. 考虑下表中显示的购物篮事务的例子。（12分）

|  |  |
| --- | --- |
| 事务ID | 购买项 |
| 1 | ｛B，C｝ |
| 2 | ｛R，D，M｝ |
| 3 | ｛M，T，C｝ |
| 4 | ｛M，B，T｝ |
| 5 | ｛B，C，T｝ |
| 6 | ｛M，T，R，D｝ |
| 7 | ｛R，D，T｝ |
| 8 | ｛B，T｝ |
| 9 | ｛R，D，C｝ |
| 10 | ｛M，T，R，D｝ |

* 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）=602
  2. 能投提取的频繁项集的最大长度是多少？（假定最小支持度>0）4
  3. 写出从该数据集中能够提取的3-项集的最大数量的表达式。
  4. 找出一个具有最大支持度的项集（长度为2或更大）。(D,R)
  5. 找出一对项a和b，使得规则和具有相同的置信度。

(B,C) or (D,R)

**课后习题：6.6 ，修改了数据集中的相应数据： BBread与RBeer交换；TButter与DDiapers交换。**

1. 聚类分析 （14分）
2. 在聚类分析中，传统的K-means算法都有哪些局限性？有哪些相应的改进方法？
3. 请简要描述聚类与关联分析的主要相似点和不同点。
4. 请举出一个采用聚类作为主要的数据挖掘方法的实际应用例子。
5. SVM计算题 （12分）

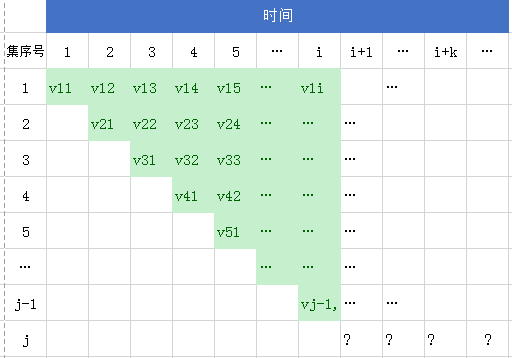
已知正例点x1 = (1, 2)T，x2 = (2, 3)T ，负例点x3 = (2, 1)T，试求最大间隔分离超平面和分类决策函数，并在图上画出分离超平面、间隔边界及支持向量。

拉格朗日乘子：(3,2,5). W1=-3, w2=7, b=-6. ???在算一遍？？

1. 优酷网是国内领先的视频分享网站，兼具影视、综艺、资讯三大内容形态。电视剧因其独特的表现形式，一直备受大众欢迎。优酷网提供了专门的电视剧播放频道，用户可以进入页面搜索浏览自己喜欢看的电视剧并下载收藏。目前，新剧《[神探夏洛克 第三季](http://v.youku.com/v_show/id_XNjU2MjIzOTQw.html)》已在优酷网播放若干集，部门经理希望通过预测后续剧集的点击率，决定是否继续引进该剧。（18分）

已知的数据包含很多部已上映完的电视剧的基本信息和播放记录，具体如下：

* 基本信息包括：电视剧名称、类型（剧情、警匪、悬疑…）、上映时间、导演、演员、集数、集均播放量、评论数、收藏数、豆瓣评分
* 播放记录指电视剧上映后每集每天的播放量，如下图所示，假设每部电视剧每天只更新一集，Vji代表第j集上映后第i天的播放量



我们的目标就是：如上图所示，给定当前电视剧前j-1集的第1至第i天的播放记录（例如j=11，i=7），预测第j集上映后每天的播放量即Vj,i+1。

请利用你所学过的机器学习和数据挖掘的方法帮助部门经理解决该问题。

1. 你认为上述已知的数据中，哪些特征或数据对于预测新一集的播放量有明显作用？
2. 结合(1)中选择的特征或数据，使用什么技术、什么算法来解决预测问题，请具体说明之。
3. 如果想预测一部新电视剧（所有集都没有上映）第一集上映当天的播放量，请简要给出一个解决方案。