H2 机器学习作业一

20185670 计科卓越班 蔡嘉轩

H3 1.1 求相应的版本空间

改动过的表1.1如下

编号	色泽	根蒂	敲声	是否好瓜
1	青绿	蜷缩	浊响	是
2	乌黑	稍蜷	沉闷	否

假设空间的大小为 (2+1)*(2+1)*(2+1)+1 =28;将假设空间列表得

编号	色泽	根蒂	敲声	是否好瓜
1	乌黑	蜷缩	浊响	否
2	乌黑	蜷缩	沉闷	否
3	乌黑	蜷缩	*	否
4	乌黑	稍蜷	浊响	否
5	乌黑	稍蜷	沉闷	否
6	乌黑	稍蜷	*	否
7	乌黑	*	浊响	否
8	乌黑	*	沉闷	否
9	乌黑	*	*	否
10	青绿	蜷缩	浊响	是
11	青绿	蜷缩	沉闷	否
12	青绿	蜷缩	*	是
13	青绿	稍蜷	浊响	否
14	青绿	稍蜷	沉闷	否

编号	色泽	根蒂	敲声	是否好瓜
15	青绿	稍蜷	*	否
16	青绿	*	浊响	是
17	青绿	*	沉闷	否
18	青绿	*	*	是
19	*	蜷缩	浊响	是
20	*	蜷缩	沉闷	否
21	*	蜷缩	*	是
22	*	稍蜷	浊响	否
23	*	稍蜷	沉闷	否
24	*	稍蜷	*	否
25	*	*	浊响	是
26	*	*	沉闷	否
27	*	*	*	否
28	NULL	NULL	NULL	否

结合样例，检验得版本空间为

编号	色泽	根蒂	敲声
1	青绿	蜷缩	浊响
2	青绿	蜷缩	*
3	青绿	*	浊响
4	青绿	*	*
5	*	蜷缩	浊响
6	*	蜷缩	*
7	*	*	浊响

1.3 设计一种归纳偏好

在数据中包含噪声的情况下，首要思路是对噪声进行处理防止出现二义性。可选择如下的处理的方法：

a. 若存在属性取值都相同标记却不同的两个案例，则只保留标记为正例的样例或者只保留负例的样例，在此基础上求出版本空间。

b. 将属性相近的两个数据分为同一类。若相同属性出现了两种不同的分类，则认为它属于与他最临近几个数据的属性。

a留下的数据无误差，但是可能会丢失部分信息，b不会丢失信息，但是对于某些数据可能存在一些误差。

1.5 简述机器学习在互联网搜索环节的作用

1.网页排序：使用词位置加权的搜索引擎--基于词频统计进行排序，机器学习统计词频，关键词在文档中词频越高，出现的位置越重要，则被认为和检索词的相关性越好。

2.搜索引擎直接给出搜索答案：神经网络被用来分析大量的数据来完成特定的任务，比如从相关的网页中检索长句子和段落，然后提供关于问题答案的信息。

3.搜索图片、视频和其他多元数据：搜索引擎直接从视频和图片中提取信息、谷歌推出了视频智能API，不仅可以从视频中提取出具体的信息，还可以总结视频的上下文，记录视频中的场景，从而对视频进行准确的分类。

4.使得搜索结果有更精确的顺序：使用神经网络、决策树等基于web页面排序算法:RankNet、LambdaRank和LambdaMART。

5.会话智能交互搜索：例如，百度语音搜索，Siri搜索或谷歌助手。它涉及自然语言处理、知识映射和神经网络。

6.对垃圾网站的筛选（模式识别）：对于部分垃圾网站进行模式识别与筛选可以通过离群值测试来实现，或者搜索引擎优化算法来进行网站筛选。

7.对用户行为进行全面分析：在电商等互联网搜索环节中，很多电商都会收集用户信息基于机器学习算法对用户进行画像等分析，再对其进行商品推荐。