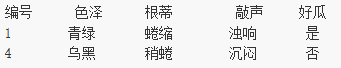
**注意：此习题为笔者自己所解，仅供参考，如有错误，欢迎指正！**

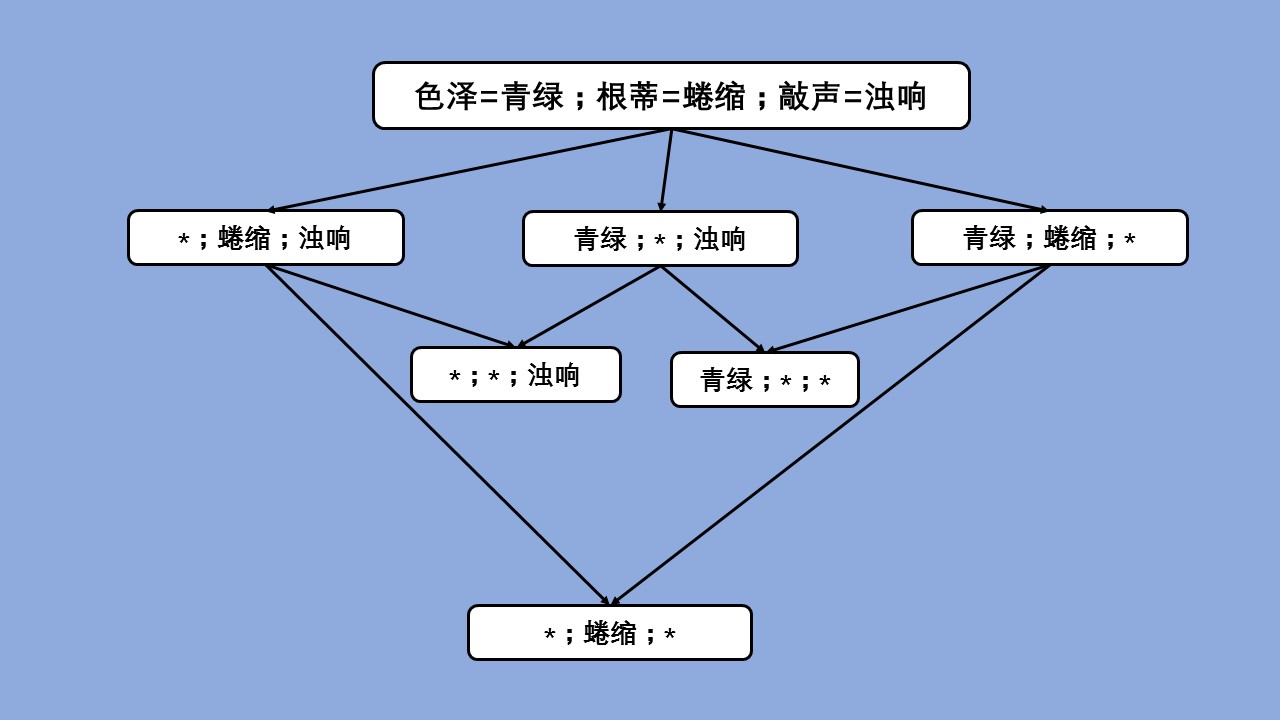
习题

* 1. **表1.1中若只包含编号为1和4的两个阳历，试给出相应的版本空间。**



答：版本空间的定义是与训练集一致的假设空间的假设集合。根据表格可以看到总的假设空间有（2+1）\*（2+1）\*（2+1）+ 1 = 28种假设。

版本空间是包含于假设空间中的，即符合训练集中好瓜描述的假设空间的假设集合，1号和4号中三个属性都是不相同的，所以我们不能确定到底是瓜的哪个属性决定了瓜的好坏，所以对应于好瓜的版本空间如下：



可以看到对应于好瓜的版本空间中含有7中假设。

* 1. **与使用单个合取式来进行假设表示相比，使用“析合范式”将使得假设空间具有更强的表示能力例如：**

**好瓜<-->(（色泽=\*）∩（根蒂=蜷缩）∩（敲声=\*）)∪（（色泽=乌黑）∩（根蒂=\*）∩（敲声=沉闷））**

**会把（色泽=青绿）∩（根蒂=蜷缩）∩（敲声=清脆）以及（色泽=乌黑）∩（根蒂=硬挺）∩（敲声=沉闷）都分类为好瓜。若使用最多包含k个合取式的析合范式来表达表1.1的西瓜分类问题的假设空间，试估算共有多少种可能的假设。**

答：表1.1中包含4个样例，三种属性，假设空间总共有（2+1）\*（3+1）\*（3+1）+1=49种假设。在不考虑冗余的情况下，最多包含k个合取式来表达假设空间，这里k的最大取值就是49，每次从中选出k个来组成析合式的话，根据排列组合可以有∑C49k =2^49种可能，但是这里面包含了很多冗余的情况。关于这道题的解法，详见这个链接：<http://blog.csdn.net/icefire_tyh/article/details/52065626>

* 1. **若数据包含噪声，则假设空间种有可能不存在与所有训练样本都一致的假设。在此情形下，试着设计一种归纳偏好用于假设选择。**

答：1.通常认为两个数据的属性越相近，则更倾向于他们分为同一类。若相同属性属性出现了两种不同的分类，则认为它属于与他最临近的几个数据的属性，这个跟聚类的思想有点类似，就是找最近的邻居的属性。

2. 可以考虑把这些具有属性相同而不同分类的数据，留下的这些数据都将是没有误差的数据了，这种方法在数据量多而噪声比较少的情况下可以使用，不然可能会丢失部分信息。

**1.4 本章1.4节在论述“没有免费的午餐”定理时，默认使用了“分类错误率”作为性能度量来对分类器进行评估。若换用其他性能度量l,试证明没有免费的午餐”定理仍成立.**

答：NFL首先要保证真目标函数f均匀分布。对于X个训练数据的二分类问题，显然f共有http://images2015.cnblogs.com/blog/1180557/201707/1180557-20170712205955790-886171935.png种情况，其中一半是与假设一致的，也就是P(f(x) = h(x)) = 0.5.

因此，http://images2015.cnblogs.com/blog/1180557/201707/1180557-20170712210329931-1160238508.png

其中http://images2015.cnblogs.com/blog/1180557/201707/1180557-20170712210348087-936350073.png应为常数，如果性能度量为错误率，二者各为0.5，则该值为1，如果为其他性能度量， http://images2015.cnblogs.com/blog/1180557/201707/1180557-20170712211125431-1644465005.png，应当隐含这样的充分条件。

1.5**试述机器学习能在互联网搜索的哪些环节起什么作用。**

答：互联网搜索中能够根据每个人的搜索习惯来进行定点定向投送广告（百度就是这么干的），根据搜索量进行网站排名，智能语义识别搜索等等。