## 第一章 绪论

1.1通过对绪论的学习，你对机器学习有什么了解？

参考答案：机器学习是这样的一门学科，它致力于研究如何通过计算的手段，利用经验来改善系统本身的性能。

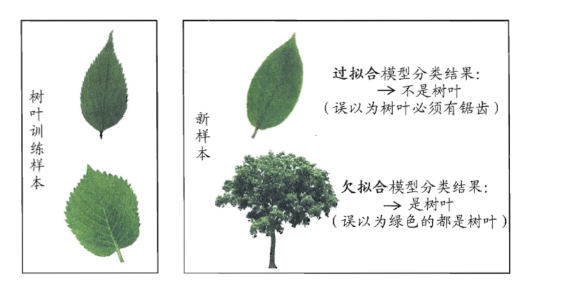
1.2根据预测结果的不同可以将学习任务分为哪两类，其区别是什么？

参考答案：“分类”和“回归”。若我们预测的是离散值，如“好瓜”“坏瓜”，此类学习任务称为分类，若预测的是连续值，如西瓜的成熟度0.95、0.37，此类学习任务称为回归。

1.3简要叙述训练数据集和测试数据集的含义。

参考答案：在类似于机器学习的各个信息科学相关领域中，一组数据被用来发现潜在的预测关系，称为训练数据集。训练数据集是提供给学习者的案例，而试验数据集是用于测试由学习者提出的假设关系的准确度。

1.4结合下图说明简要说明过拟合的含义及其出现的原因。



参考答案：（1） 过拟合：如果一味的去提高训练数据的预测能力，所选模型的复杂度往往会很高，这种现象称为过拟合。其表现就是模型在训练集上的误差很小，但在测试集上误差很大。  
(2) 产生的原因：1. 样本数据的问题。样本数量太少；抽样方法错误；有效样本里的噪音数据干扰过大等。 2. 模型问题。模型复杂度高 、参数太多；决策树模型没有剪枝；权值学习迭代次数过多(Overtraining),拟合了训练数据中的噪声和训练样例中没有代表性的特征。

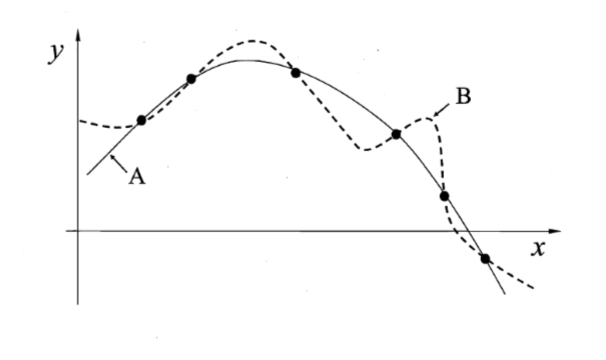
1.5 一个初学机器学习的朋友对房价进行预测。他在一个N=1000个房价数据的数据集上匹配了一个有533个参数的模型，该模型能解释测试数据集上99%的变化，请问该模型能很好地预测来年的房价吗？简要说明原因。

参考答案：虽然该模型在测试数据集上变现很好，但并不一定能很好地预测来年房价的变化，因为其选取的参数过多，可能会拟合数据中的噪声以及一些没有代表性的特征，导致过拟合问题。

1.6 简要叙述什么是有监督学习和无监督学习？

参考答案：有监督学习：对具有标记的训练样本进行学习，以尽可能对训练样本集外的数据进行分类预测。无监督学习：对未标记的样本进行训练学习，以发现这些样本中的结构知识。

1.7 对于一个有限的样本训练集可能存在多条曲线与之一致（如下图所示），我们该如何取舍？其一般性原则是什么？

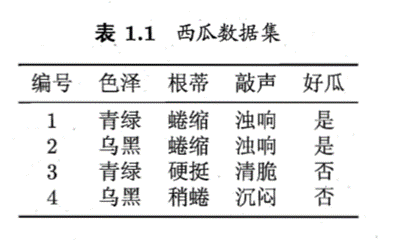


参考答案：任何一个有效的机器学习算法必有其归纳偏好，否则无法产生确定的学习结果，例如，若认为相似的样本必须有相似的输出，则对应的学习算法可能更偏好图中比较“平滑”的曲线A而不是比较“崎岖”的曲线B。“奥卡姆剃刀”，即“若有多个假设与观察一致，则选择最简单的那一个”是一种常用的判断原则。

1.8给你一个数据集，这个数据集有缺失值，且这些缺失值分布在离中值有1个标准偏差的范围内。百分之多少的数据不会受到影响？为什么？

参考答案：约有32%的数据将不受缺失值的影响。因为，由于数据分布在中位数附近，让我们先假设这是一个正态分布。我们知道，在一个正态分布中，约有68%的数据位于跟平均数（或众数、中位数）1个标准差范围内，那么剩下的约32%的数据是不受影响的。因此，约有32%的数据将不受缺失值的影响。

1.9表1.1中若只包含编号为1和4的两个样例，试给出相应的版本空间。



参考答案：精简过的数据集如下

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 编号 | 色泽 | 根蒂 | 敲声 | 好瓜 |
| 1 | 青绿 | 蜷缩 | 浊响 | 是 |
| 2 | 乌黑 | 稍蜷 | 沉闷 | 否 |

数据集共有3个属性，每个属性2种取值，一共 3∗3∗3+1=28种假设，分别为：

1.色泽=青绿 根蒂=蜷缩 敲声=浊响 2.色泽=青绿 根蒂=蜷缩 敲声=沉闷

3.色泽=青绿 根蒂=稍蜷 敲声=浊响 4.色泽=青绿 根蒂=稍蜷 敲声=沉闷

5.色泽=乌黑 根蒂=蜷缩 敲声=浊响 6.色泽=乌黑 根蒂=蜷缩 敲声=沉闷

7.色泽=乌黑 根蒂=稍蜷 敲声=浊响 8.色泽=乌黑 根蒂=稍蜷 敲声=沉闷

9.色泽=青绿 根蒂=蜷缩 敲声=\* 10.色泽=青绿 根蒂=稍蜷 敲声=\*

11.色泽=乌黑 根蒂=蜷缩 敲声=\* 12.色泽=乌黑 根蒂=稍蜷 敲声=\*

13.色泽=青绿 根蒂=\* 敲声=浊响 14.色泽=青绿 根蒂=\* 敲声=沉闷

15.色泽=乌黑 根蒂=\* 敲声=浊响 16.色泽=乌黑 根蒂=\* 敲声=沉闷

17.色泽=\* 根蒂=蜷缩 敲声=浊响 18.色泽=\* 根蒂=蜷缩 敲声=沉闷

19.色泽=\* 根蒂=稍蜷 敲声=浊响 20.色泽=\* 根蒂=稍蜷 敲声=沉闷

21.色泽=青绿 根蒂=\* 敲声=\* 22.色泽=乌黑 根蒂=\* 敲声=\*

23.色泽=\* 根蒂=蜷缩 敲声=\* 24.色泽=\* 根蒂=稍蜷 敲声=\*

25.色泽=\* 根蒂=\* 敲声=浊响 26.色泽=\* 根蒂=\* 敲声=沉闷

27.色泽=\* 根蒂=\* 敲声=\* 28.空集Ø（即好瓜这个概念根本不成立）   
根据数据集可以删除假设 2−8，10−12，14−16，18−20，22，24，26，27，28。  
所以版本空间为:

1.色泽=青绿 根蒂=蜷缩 敲声=浊响 9.色泽=青绿 根蒂=蜷缩 敲声=\*

13.色泽=青绿 根蒂=\* 敲声=浊响 17.色泽=\* 根蒂=蜷缩 敲声=浊响

21.色泽=青绿 根蒂=\* 敲声=\* 23.色泽=\* 根蒂=蜷缩 敲声=\*

25.色泽=\* 根蒂=\* 敲声=浊响

1.10 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=\*)) V（（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西瓜分类问题的假设空间，试估算共有多少种可能的假设（不考虑沉余情况）？

参考答案：表1.1包含4个样例，3种属性，所以假设空间中有3∗4∗4+1=49种假设。在不考虑沉余的情况下，最多包含k个合取式来表达假设空间，显然k的最大值是49，每次从中选出k个来组成析合式，共ΣCk49=249 种可能。但是其中包含了很多沉余的情况(至少存在一个合取式被剩余的析合式完全包含<空集除外>)。

1.11 请简要说说一个完整机器学习项目的流程。

参考答案：抽象成数学问题 2.获取数据 3.特征预处理与特征选择 4.训练模型与调优 5.模型诊断6.模型融合 7.上线运行

1.12对所有优化问题来说, 有没有可能找到比现在已知算法更好的算法？（结合“没有免费的午餐”定理作答）

参考答案：根据“没有免费的午餐”定理：对于训练样本，不同的算法A/B在不同的测试样本中有不同的表现，这表示：对于一个学习算法A，若它在某些问题上比学习算法 B更好，则必然存在一些问题，在那里B比A好。也就是说：对于所有问题，无论学习算法A多聪明，学习算法 B多笨拙，它们的期望性能相同。但是：没有免费午餐定力假设所有问题出现几率相同，实际应用中，不同的场景，会有不同的问题分布，所以，在优化算法时，针对具体问题进行分析，是算法优化的核心所在。

1.13若数据包含噪声，则假设空间中有可能不存在与所有训练样本都一致的假设，在此情形下，试设计一种归纳偏好用于假设选择。

参考答案：通常认为两个数据的属性越相近，则更倾向于将他们分为同一类。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。也可以考虑同时去掉所有具有相同属性而不同分类的数据，留下的数据就是没误差的数据，但是可能会丢失部分信息。

1.14数据挖掘和机器学习的区别是什么？

参考答案：机器语言是指在没有明确的程序指令的情况下，给予计算机学习能力，使它能自主的学习、设计和扩展相关算法。数据挖掘则是一种从非结构化数据里面提取知识或者未知的、人们感兴趣的图片。在这个过程中应用了机器学习算法。

1.15试简述机器学习在互联网搜索的哪些环节起什么作用？

参考答案：

1.最常见的，消息推送，比如淘宝经常会说某些商品你可能会感兴趣；  
2.网站相关度排行，通过点击量，对网页内容进行综合分析、排序等 ；  
3.通过搜索预测天气，流感等等。