机器学习导论
习题三

参考答案

2017 年 5 月 3 日

1 [30pts] Decision Tree Analysis

决策树是一类常见的机器学习方法，但是在训练过程中会遇到一些问题。

(1) **[15pts]** 试证明对于不含冲突数据(即特征向量完全相同但标记不同)的训练集，必存在与训练集一致(即训练误差为0)的决策树;

(2) **[15pts]** 试分析使用“最小训练误差”作为决策树划分选择的缺陷。

Solution.

(1) 可通过反证法，直接构造等方式证明。（言之有理即可）

(2) 过拟合问题，计算存储开销大等（言之有理即可）

需注意事项：第一问的证明中，很多同学的讨论局限在了“特征是离散变量”，建议对连续性变量也进行讨论（提示：可以将连续值离散化，见书4.4.1）。第二问的证明中，有些同学的讨论是基于第一问自己提出的（如暴力穷举的）模型进行的，这样做是片面的，应该去讨论“最小训练误差”作为准则的决策树的缺陷，而不是某一种特例模型的缺陷。

2 [30pts] Training a Decision Tree

考虑下面的训练集：共计6个训练样本，每个训练样本有三个维度的特征属性和标记信息。详细信息如表1所示。

请通过训练集中的数据训练一棵决策树，要求通过“信息增益”(information gain)为准则来选择划分属性。请参考书中图4.4，给出详细的计算过程并画出最终的决策树。

Solution. 本题目主要考察对课本中决策树生成过程的理解，尤其是当划分指标在两个特征上取值相同时，要用一个一致的指标进行选择。最终的决策树有多种形态，如下。

表 1: 训练集信息

序号	特征 **A**	特征 **B**	特征 **C**	标记
1	0	1	1	0
2	1	1	1	0
3	0	0	0	0
4	1	1	0	1
5	0	1	0	1
6	1	0	1	1

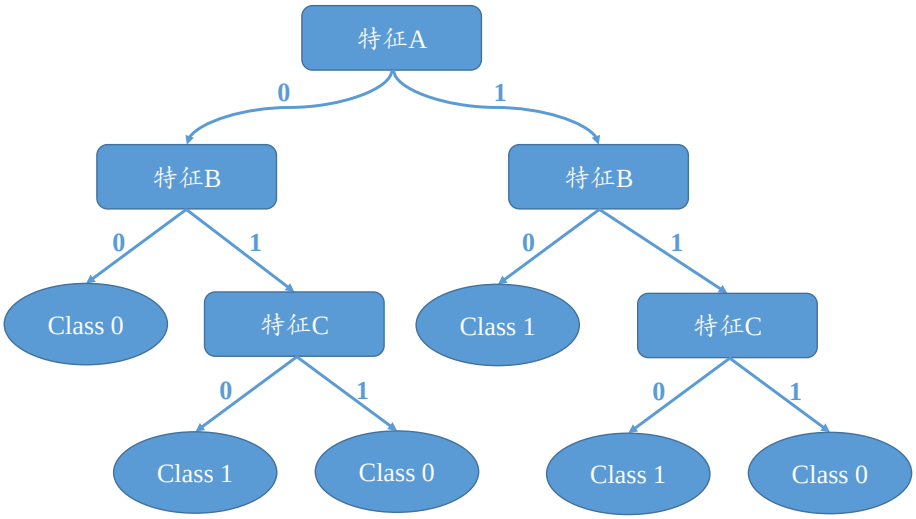

图 1: 生成的决策树-1

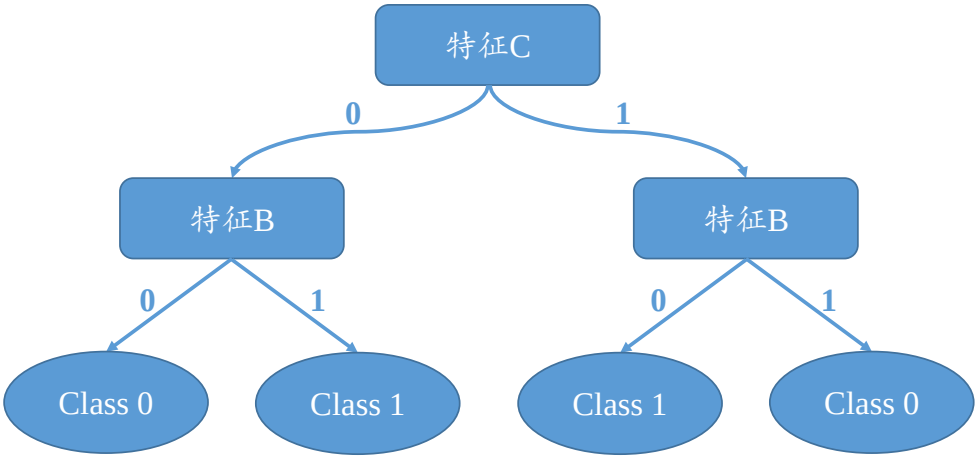

图 2: 生成的决策树-2

3 [40pts] Back Propagation

单隐层前馈神经网络的误差逆传播(error BackPropagation，简称BP)算法是实际工程实践中非常重要的基础，也是理解神经网络的关键。

请编程实现BP算法，算法流程如课本图5.8所示。详细编程题指南请参见链接：`http://lamda.nju.edu.cn/ml2017/PS3/ML3_programming.html`

在实现之后，你对BP算法有什么新的认识吗？请简要谈谈。

Solution. 编程题没有标准答案，只要按照课本公式和算法实现，通过检测就算正确。能够加深对BP算法和神经网络的认识，本题的目的就达到了。综合测试集精度、运行时间、可读性，我们选出优秀代码公布于 `http://lamda.nju.edu.cn/ml2017/ml_faq.html`中供同学们参考。

附加题 [30pts] Neural Network in Practice

在实际工程实现中，通常我们会使用已有的开源库，这样会减少搭建原有模块的时间。因此，请使用现有神经网络库，编程实现更复杂的神经网络。详细编程题指南请参见链接：`http://lamda.nju.edu.cn/ml2017/PS3/ML3_programming.html`

和上一题相比，模型性能有变化吗？如果有，你认为可能是什么原因。同时，在实践过程中你遇到了什么问题，是如何解决的？

Solution. 使用Keras完成任务比较简单，MatConvNet的编程则较为困难，我们选出了优秀代码公布于`http://lamda.nju.edu.cn/ml2017/ml_faq.html`中供同学们参考。

通过这两道编程题，我们希望大家体会到，神经网络模型的性能与许多因素有关，与上一题相比，性能有可能提升，也有可能变差。使用库函数还会遇到安装、配置、调试方面的奇特问题，这些是自己从零开始编程不会遇到的。当然，正确使用库函数可以提高编程速度和运行速度，将来处理实际问题时，应当以使用库为主。