机器学习导论
习题三

141180016, 丁俊峰, 141180016@smail.nju.edu.cn

2017 年 4 月 26 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)反证法：假设对于不含冲突数据的训练集，存在与训练集不一致的决策树，则在某个节点上的多个数据会出现无法分开的情况，这与不含冲突数据的训练集假设不符，所以假设不成立。必定存在于训练集一致的决策树。
(2) 为了最小化训练误差，决策树会对训练数据过拟合，造成泛化能力较差，不适用新的情况。

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。

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

Solution.

训练集D的信息熵为$Ent(D)=-(\frac{3}{6}log_2\frac{3}{6}+\frac{3}{6}log_2\frac{3}{6})=1$

对于特征A,$Ent(D^0)=-(\frac{2}{3}log_2\frac{2}{3}+\frac{1}{3}log_2\frac{1}{3})=0.918$,$Ent(D^1)=-(\frac{2}{3}log_2\frac{2}{3}+\frac{1}{3}log_2\frac{1}{3})=0.918$

对于特征B,$Ent(D^1)=-(\frac{1}{2}log_2\frac{1}{2}+\frac{1}{2}log_2\frac{1}{2})=1$,$Ent(D^1)=-(\frac{1}{2}log_2\frac{1}{2}+\frac{1}{2}log_2\frac{1}{2})=1$

对于特征C,$Ent(D^2)=-(\frac{2}{3}log_2\frac{2}{3}+\frac{1}{3}log_2\frac{1}{3})=0.918$,$Ent(D^1)=-(\frac{2}{3}log_2\frac{2}{3}+\frac{1}{3}log_2\frac{1}{3})=0.918$

属性A的信息增益为$Gain(D,A)=Ent(D)-\sum\limits_{v=0}^{1}\frac{|D^v|}{|D|}Ent(D^v)=1-(\frac{1}{2}\times0.918+\frac{1}{2}\times0.918)=0.082$

属性B的信息增益为$Gain(D,B)=Ent(D)-\sum\limits_{v=0}^{1}\frac{|D^v|}{|D|}Ent(D^v)=1-(\frac{1}{3}\times1+\frac{2}{3}\times1)=0$

属性C的信息增益为$Gain(D,C)=Ent(D)-\sum\limits_{v=0}^{1}\frac{|D^v|}{|D|}Ent(D^v)=1-(\frac{1}{2}\times0.918+\frac{1}{2}\times0.918)=0.082$

可以看到A和C的信息增益均优于B，下面对A和C作进一步讨论。

（1）若选择A为划分属性，分为（1，3，5）和（2，4，6）两个子集。

对于A为0的子集，$Ent(D^0)=-(\frac{2}{3}log_2\frac{2}{3}+\frac{1}{3}log_2\frac{1}{3})=0.918$

对于特征B,$Ent(D^0)=0$,$Ent(D^1)=1$

对于特征C,$Ent(D^0)=1$,$Ent(D^1)=0$

属性B的信息增益为$Gain(D^0,B)=Ent(D^0)-\sum\limits_{v=0}^{1}\frac{|D^v|}{|D|}Ent(D^v)=1-(\frac{1}{3}\times0+\frac{2}{3}\times1)=0.257$

属性C的信息增益为$Gain(D^0,C)=Ent(D^0)-\sum\limits_{v=0}^{1}\frac{|D^v|}{|D|}Ent(D^v)=1-(\frac{1}{3}\times0+\frac{2}{3}\times1)=0.257$

若选择B为划分属性

对于A为1的子集，$Ent(D^1)=-(\frac{2}{3}log_2\frac{2}{3}+\frac{1}{3}log_2\frac{1}{3})=0.918$

对于特征B,$Ent(D^0)=-(\frac{1}{2}log_2\frac{1}{2}+\frac{1}{2}log_2\frac{1}{2})=1$,$Ent(D^1)=-(\frac{1}{2}log_2\frac{1}{2}+\frac{1}{2}log_2\frac{1}{2})=1$

对于特征C,$Ent(D^1)=-(\frac{2}{3}log_2\frac{2}{3}+\frac{1}{3}log_2\frac{1}{3})=0.918$,$Ent(D^1)=-(\frac{2}{3}log_2\frac{2}{3}+\frac{1}{3}log_2\frac{1}{3})=0.918$

属性B的信息增益为$Gain(D^1,B)=Ent(D^1)-\sum\limits_{v=0}^{1}\frac{|D^v|}{|D|}Ent(D^v)=1-(\frac{1}{3}\times1+\frac{2}{3}\times1)=0$

属性C的信息增益为$Gain(D^1,C)=Ent(D^1)-\sum\limits_{v=0}^{1}\frac{|D^v|}{|D|}Ent(D^v)=1-(\frac{1}{2}\times0.918+\frac{1}{2}\times0.918)=0.082$

所以$Gain(D_{A=0},B)=0.251$,$Gain(D_{A=0},C)=0.251$，同理$Gain(D_{A=1},B)=0.251$,$Gain(D_{A=1},C)=0.251$。且子集还需要再划分。

（2）若一开始选择C为划分属性，分为（3，4，5）和（1，2，6）两个子集。

对于C为0的子集,$Ent(D_{C=0})=-\sum_{k=0}^{1}p_k log_2 p_k=-(\frac{1}{3}log_2\frac{1}{3}+\frac{2}{3}log_2\frac{2}{3})=0.918$

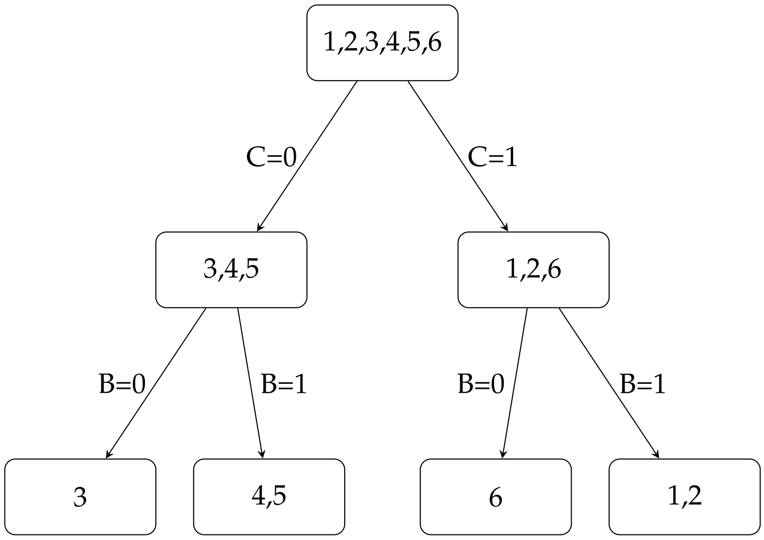
对于特征A,$Ent(D^0_{C=0})=-(\frac{1}{2}log_2\frac{1}{2}+\frac{1}{2}log_2\frac{1}{2})=1$,$Ent(D^1_{C=0})=-(1\times log_2 1+0\times log_2 0)=0$。

对于特征B,$Ent(D^0_{C=0})=-(1\times log_2 1+0\times log_2 0)=0$，$Ent(D^1_{C=0})=-(1\times log_2 1+0\times log_2 0)=0$

特征A的信息增益为$Gain(D_{C=0},A)=Ent(D_{C=0})-\sum_{v=0}^{1}\frac{|D^v_{C=0}|}{|D_{C=0}|}Ent(D^v_{C=0})=1-(\frac{2}{3}\times Ent(D^0_{C=0})+\frac{1}{3}\times Ent(D^1_{C=0}))=0.252$

特征B的信息增益为$Gain(D_{C=1},B)=Ent(D_{C=1})-\sum_{v=0}^{1}\frac{|D^v_{C=1}|}{|D_{C=1}|}Ent(D^v_{C=1})=1-(\frac{1}{3}\times$

$Ent(D^0_{C=1}) + \frac{2}{3} \times Ent(D^1_{C=1})) = 0.918$
选择特征B，且子集不需要再划分。
对于集合C为1的子集,$Ent(D_{C=1}) = -\sum_{k=0}^{1} p_k log_2 p_k = -(\frac{1}{3}log_2\frac{1}{3} + \frac{2}{3}log_2\frac{2}{3}) = 0.918$
对于特征A，$Ent(D^0_{C=1}) = -(1 \times log_2 1 + 0 \times log_2 0) = 0$，$Ent(D^1_{C=1}) = -(\frac{1}{2}log_2\frac{1}{2} + \frac{1}{2}log_2\frac{1}{2}) = 1$
对于特征B，$Ent(D^0_{C=1}) = -(1 \times log_2 1 + 0 \times log_2 0) = 0$，$Ent(D^1_{C=1}) = -(1 \times log_2 1 + 0 \times log_2 0) = 0$
特征A的信息增益为$Gain(D_{C=1}, A) = Ent(D_{C=0}) - \sum_{v=0}^{1} \frac{|D^v_{C=0}|}{|D_{C=0}|} Ent(D^v_{C=0}) = 1 - (\frac{2}{3} \times Ent(D^0_{C=0}) + \frac{1}{3} \times Ent(D^1_{C=0})) = 0.252$
特征B的信息增益为$Gain(D_{C=1}, B) = Ent(D_{C=1}) - \sum_{v=0}^{1} \frac{|D^v_{C=1}|}{|D_{C=1}|} Ent(D^v_{C=1}) = 1 - (\frac{1}{3} \times Ent(D^0_{C=1}) + \frac{2}{3} \times Ent(D^1_{C=1})) = 0.918$
选择特征B,且子集不需要再划分。
综上，根据奥卡姆剃刀原则，最优总划分顺序应该为C、B、A。决策树如下：



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.
为了高效计算更新权重，不能使用正则方程，所以BP算法也是用到了梯度下降法，可见梯度下降法在机器学习中的重要地位。而梯度下降这种优化方式决定了BP算法的关键就是求导时链式法则的应用，将误差逐层传播，更新权重。

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.
性能有所提升，迭代次数更少时模型精度更高，原因是增加了隐层数目和单元数，模型的拟合能力更强。实践中在模型训练时使用的标签有问题，无法训练，后来查阅官方文档发现keras接口使用的标签数据需要是one-hot格式，即要将整数类型标签转化为二进制向量。