

机器学习导论

作业三

学号, 作者姓名, 邮箱

2018 年 4 月 24 日

1 [15pts] Decision Tree I

- (1) [5pts] 假设一个包含三个布尔属性 X, Y, Z 的空间, 并且目标函数是 $f(x, y, z) = x \text{ XOR } z$, 其中 XOR 为异或运算符。令 H 为基于这三个属性的决策树, 请问: 目标函数 f 可实现吗? 如果可实现, 画出相应的决策树以证明; 如果不可实现, 请论证原因;
- (2) [10pts] 现有如表 1所示数据集:

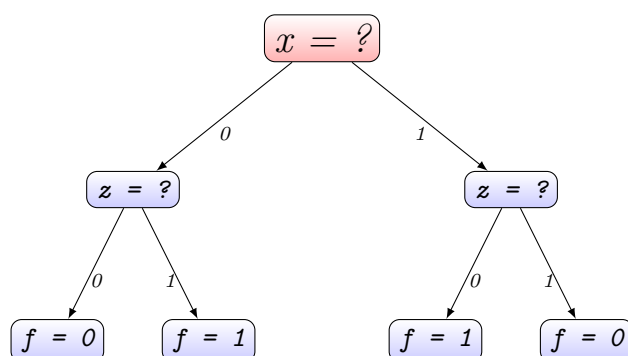
表 1: 样例表

X	Y	Z	f
1	0	1	1
1	1	0	0
0	0	0	0
0	1	1	1
1	0	1	1
0	0	1	0
0	1	1	1
1	1	1	0

请画出由该数据集生成的决策树。划分属性时要求以信息增益 (information gain)为准则。当信息增益 (information gain)相同时, 依据字母顺序选择属性即可。

Solution.

(1) 目标函数 f 是可以实现的，决策树如下所示：



(2) 按照信息增益为准则，划分依据为：

第一层根结点：

如果选择 X 划分：

$$\begin{aligned} Gain(D, X) &= Ent(D) - \sum_{v=1}^2 \frac{|D^v|}{|D|} Ent(D^v) \\ &= 1 - 1 = 0 \end{aligned} \quad (1.1)$$

如果选择 Y 划分：

$$Gain(D, Y) = 1 - 1 = 0 \quad (1.2)$$

如果选择 Z 划分：

$$Gain(D, Z) = 1 - \frac{3}{4} \times 0.918 = 0.6885 \quad (1.3)$$

所以选择 Z 划分。

第二层第一个结点都是同一类，现在考虑第二个结点的划分：

如果选择 X 划分：

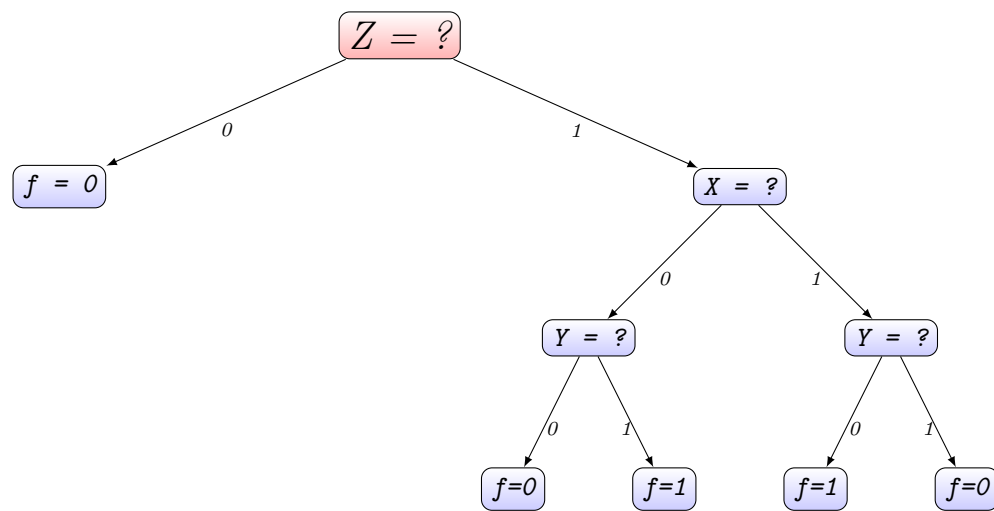
$$Gain(D, X) = 0.918 - 0.918 = 0 \quad (1.4)$$

如果选择 Y 划分：

$$Gain(D, Y) = 0.918 - 0.918 = 0 \quad (1.5)$$

所以按照字母顺序选择 X 划分。

根据数据集生成的决策树如下：



2 [20pts] Decision Tree II

考虑如下矩阵：

$$\begin{bmatrix} 4 & 6 & 9 & 1 & 7 & 5 \\ 1 & 6 & 5 & 2 & 3 & 4 \end{bmatrix}^T$$

该矩阵代表了6个样本数据，每个样本都包含2个特征 f_1 和 f_2 。这6个样本数据对应的标签如下：

$$\begin{bmatrix} 1 & 0 & 1 & 0 & 1 & 0 \end{bmatrix}^T$$

在这个问题中，我们要构造一个深度为2的树进行分类任务。

- (1) [5pts] 请计算根结点 (root) 的熵值 (entropy)；
- (2) [10pts] 请给出第一次划分的规则，例如 $f_1 \geq 4, f_2 \geq 3$ 。对于第一次划分后产生的两个结点，请给出下一次划分的规则；
提示：可以直观判断，不必计算熵。
- (3) [5pts] 现在回到根结点 (root)，并且假设我们是建树的新手。是否存在一种划分使得根结点 (root) 的信息增益 (information gain) 为0？

Solution.

- (1) 根结点的熵值为

$$Ent(D) = -\left(\frac{3}{6} \log_2 \frac{3}{6} + \frac{3}{6} \log_2 \frac{3}{6}\right) = 1.000 \quad (2.1)$$

- (2) 规则如图1所示。第一次划分的规则为 $f_1 \leq 6$ 。不满足 $f_1 \leq 6$ 的结点的所有样本的标签都是1，所以这一部分不需要再次划分。第一次划分之后，满足 $f_1 \leq 6$ 的节点的第二次划分的规则为 $f_2 \leq 1$ ，其中满足 $f_2 \leq 1$ 的结点的样本的标签都是1，不满足的都是0，所以不需要再次划分。

- (3) 规则如图2所示。按照 $f_2 \leq 2$ 来划分，将 D 分成了 D^1 和 D^2 两个部分。其中 D^1 包含第一个和第四个样例。由此可得

$$Ent(D^1) = -\left(\frac{1}{2} \log_2 \frac{1}{2} + \frac{1}{2} \log_2 \frac{1}{2}\right) = 1.000 \quad (2.2)$$

$$Ent(D^2) = -\left(\frac{2}{4} \log_2 \frac{2}{4} + \frac{2}{4} \log_2 \frac{2}{4}\right) = 1.000 \quad (2.3)$$

所以可得信息增益为

$$\begin{aligned} Gain(D, f_2 \leq 2) &= Ent(D) - \sum_{v=1}^2 \frac{|D^v|}{|D|} Ent(D^v) \\ &= 1 - \left(\frac{2}{6} \times 1 + \frac{4}{6} \times 1\right) \\ &= 0 \end{aligned} \quad (2.4)$$

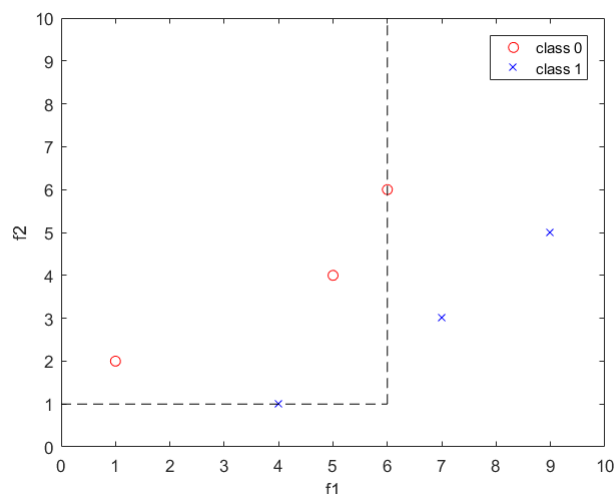


图 1: 划分规则

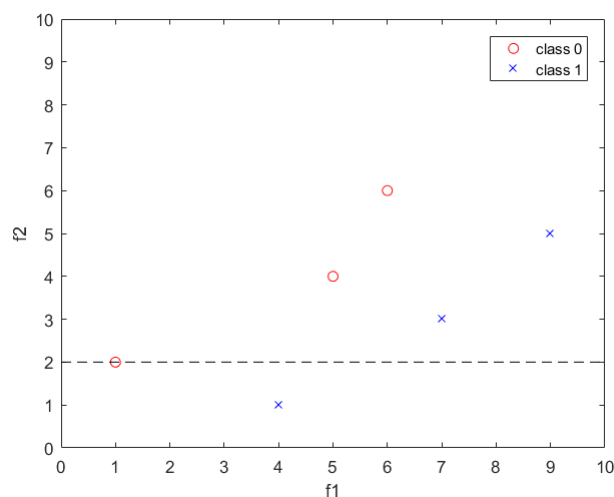


图 2: 划分规则

3 [25pts] Universal Approximator

已知函数 $f : [-1, 1]^n \mapsto [-1, 1]$ 满足 ρ -Lipschitz 性质。给定误差 $\epsilon > 0$, 请构造一个激活函数为 $\text{sgn}(\mathbf{x})$ 的神经网络 $\mathcal{N} : [-1, 1]^n \mapsto [-1, 1]$, 使得对于任意的输入样本 $\mathbf{x} \in [-1, 1]^n$, 有 $|f(\mathbf{x}) - \mathcal{N}(\mathbf{x})| \leq \epsilon$ 。

(Lipschitz 条件为: $\forall \mathbf{x}, \mathbf{y} \in [-1, 1]^n, \exists \rho > 0, \text{ s.t. } |f(\mathbf{x}) - f(\mathbf{y})| \leq \rho \|\mathbf{x} - \mathbf{y}\|_2$, 其中 $\text{sgn}(\mathbf{x})$ 的定义参见《机器学习》第98页。)

- (1) [5pts] 请画出构造的神经网络 \mathcal{N} 的示意图;
- (2) [10pts] 请对构造的神经网络进行简要的说明(写清每一层的线性组合形式, 也就是结点间的连接方式和对应的权重);

(3) [10pts] 证明自己构造的神经网络的拟合误差满足要求。

Solution. 此处用于写解答(中英文均可)

4 [40pts] Neural Network in Practice

通过《机器学习》课本第5章的学习，相信大家已经对神经网络有了初步的理解。深度神经网络在某些现实机器学习问题，如图像、自然语言处理等表现优异。本次作业旨在引导大家学习使用一种深度神经网络工具，快速搭建、训练深度神经网络，完成分类任务。

我们选取PyTorch为本次实验的深度神经网络工具，有了基础工具，我们就能如同搭积木一样构建深度神经网络。PyTorch是Facebook开发的一种开源深度学习框架，有安装方便、文档齐全、构架方便、训练效率高等特点。本次作业的首要任务就是安装PyTorch。

目前PyTorch仅支持Linux和MacOS操作系统，所以Window用户需要装一个Linux虚拟机或者直接安装Linux系统。PyTorch安装很方便，只需要在其主页中的Get Start一栏选择对应的环境设置，便能够一键安装。有GPU的同学也可以尝试安装GPU版本的PyTorch。为保证此次作业的公平性，只要求使用CPU进行网络训练，当然有条件的同学也可以尝试使用GPU进行训练。在批改作业时，助教会提供Python 2.7、3.5、3.6三种环境进行实验验证。

我们选取CIFAR10作为本次作业的训练任务。CIFAR10是一个经典的图片分类数据集，数据集中总共有60000张 32×32 的彩色图片，总共有10类，每类6000张图片，其中50000张图片构成训练集，10000张图片构成测试集。PyTorch通过torchvision给用户提供了获取CIFAR10的方法，详细信息可见PyTorch的教程。此外关于CIFAR10分类准确率排行可见此链接。

下面我们将尝试使用PyTorch来解决实际问题：

(1) [15pts] 首先我们跟随PyTorch的教程，用一个简单的卷积神经网络（Convolutional Neural Network, CNN），完成CIFAR10上的分类任务，具体要求如下：

- [7pts] 在代码实现之前，大家可能需要对CNN网络进行一定的了解，请大家自行查阅资料（PyTorch的教程中也有部分介绍CNN网络），并在实验报告中给出对CNN的见解：主要回答什么是卷积层，什么是Pooling层，以及两者的作用分别是什么；
- [8pts] 接下来就是具体的代码实现和训练。教程会手把手教你完成一次训练过程，其中使用SGD作为优化方法，请同学们自行调整epoch的大小和学习率，完成此次训练。另外，请在实验报告中给出必要的参数设置，以及训练结果如最终的loss、在测试集上的准确率等；

(2) [20pts] 显然，这样一个简单的网络在CIFAR10上并不能取得令人满意的结果，我们需要选取一个更为复杂的网络来提升训练效果。在此小题中，我们选取了CIFAR10准确率排行榜上排名第二的结构，具体参见论文链接。为了方便大家实现，我们直接给出了网络结构如图3所示。请大家搭建完成此网络结构，并选择Adam为优化器，自行调整相关参数完成训练和预测，实验结果报告内容同第（1）小题；

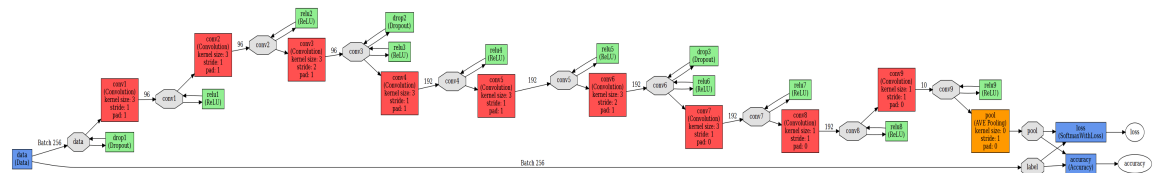


图 3: 待实现网络结构

- (3) [5pts] 通过上一题实验我们可以发现，即使使用现成的网络结构也不一定能达到与其相同的训练效果。请大家分析其中的原因，并谈谈本次实验的感想，以及对深度学习调参的体会。

实验报告.