# 机器学习作业 3

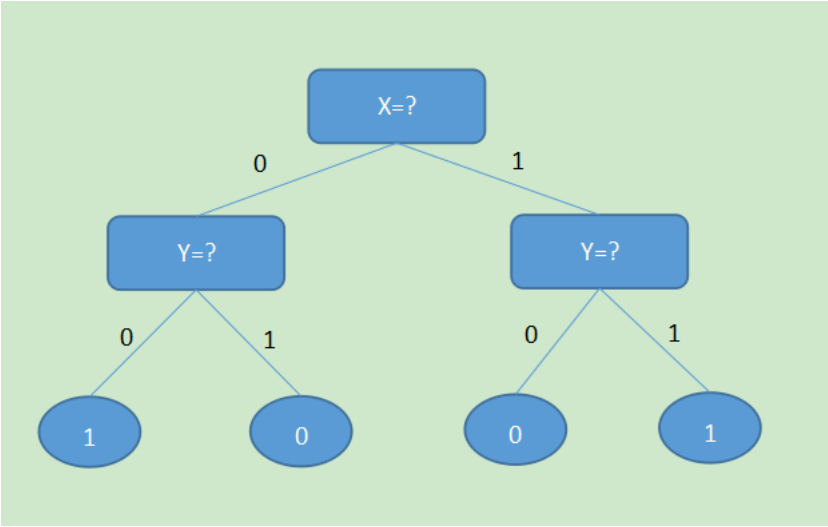
2020 年 11 月 17 日

181860152 周宇翔

## 1.[20pts]Decision Tree

### (1)

f 是 realizable 的




### (2)

$p_0 = 0.4, p_1 = 0.6$

$Ent(D) = -\sum_{k=0}^{1} p_k log_2 p_k$

$= -0.4log_20.4 - 0.6log_20.6 = 0.971$

设两个 Feature 分别为 $F_1, F_2$($F_1, F_2$ 就是 $x_1, x_2$ 我后来才发现但是懒得改了 ovo)

如果把 $F_1, F_2$ 按照离散值处理: $Gain(D, F_1) = Ent(D) - 10 * \frac{1}{10} * 0 = 0.971$

$IV(a) = log_210 = 3.32$

$Gain\_ratio(D, F_1) = 0.292$

类似的 $Gain\_ratio(D, F_2) = 0.292$

但是如果把每个 feature 当作离散值的一个分类处理, 会导致虽然每个节点的纯度很高, 但是决策树的泛化能力很弱, 无法对新样本进行有效预测

因此应该把这些值当作连续值处理

划分点选取: 首先选取 $F_1$ 作为划分属性, 连续值排序后为 22,23,24,25,32,43,48,52,52,53

候选划分点为 {22.5,23.5,24.5,28.5,37.5,45.5,50,52,52.5}

$Gain(D, F_1, 22.5) = 0.971 - 0 - \frac{9}{10}(-\frac{4}{9}log_2\frac{4}{9} - \frac{5}{9}log_2\frac{5}{9}) = 0.079$

$Gain(D, F_1, 23.5) = 0.971 + 0.2 * 2 * 0.5log_20.5 + 0.8 * (\frac{3}{8}log_2\frac{3}{8} + \frac{5}{8}log_2\frac{5}{8}) = 7.45 * 10^{-3}$

$Gain(D, F_1, 24.5) = 0.971 + 0.3(\frac{1}{3}log_2\frac{1}{3} + \frac{2}{3}log_2\frac{2}{3}) + 0.7(\frac{4}{7}log_2\frac{4}{7} + \frac{3}{7}log_2\frac{3}{7}) = 5.85 * 10^{-3}$

$Gain(D, F_1, 28.5) = 0.971 + 0.4(\frac{1}{4}log_2\frac{1}{4} + \frac{3}{4}log_2\frac{3}{4}) + 0.6(2 * \frac{1}{2}log_2\frac{1}{2}) = 0.0465$

$Gain(D, F_1, 37.5) = 0.971 + 0.5(\frac{1}{5}log_2\frac{1}{5} + \frac{4}{5}log_2\frac{4}{5}) + 0.5(\frac{2}{5}log_2\frac{2}{5} + \frac{3}{5}log_2\frac{3}{5}) = 0.124$

$Gain(D, F_1, 45.5) = 0.971 + 0.6(\frac{1}{3}log_2\frac{1}{3} + \frac{2}{3}log_2\frac{2}{3}) + 0.4 * 2 * \frac{1}{2} * log_2\frac{1}{2} = 0.0200$

$Gain(D, F_1, 50) = 0.971 + 0.7(\frac{2}{7}log_2\frac{2}{7} + \frac{5}{7}log_2\frac{5}{7}) + 0.3(\frac{1}{3}log_2\frac{1}{3} + \frac{2}{3}log_2\frac{2}{3}) = 0.0913$

$Gain(D, F_1, 52) = Gain(D, F_1, 52.5) = 0.971 + 0.9(\frac{2}{3}log_2\frac{2}{3} + \frac{1}{3}log_2\frac{1}{3}) = 0.145$

选取 $F_2$ 作为划分属性, 连续值排序后为 25,27,38,40,44,48,52,65,77,110

候选划分点为 {26,32.5,39,42,46,50,58.5,71,93.5}

$Gain(D, F_2, 26) = 0.971 + 0.9(\frac{2}{3}log_2\frac{2}{3} + \frac{1}{3}log_2\frac{1}{3}) = 0.0527$

$Gain(D, F_2, 32.5) = 0.971 + 0.8(\frac{1}{4}log_2\frac{1}{4} + \frac{3}{4}log_2\frac{3}{4}) = 0.322$

$Gain(D, F_2, 39) = 0.971 + 0.3(\frac{1}{3}log_2\frac{1}{3} + \frac{2}{3}log_2\frac{2}{3}) + 0.7(\frac{5}{7}log_2\frac{5}{7} + \frac{2}{7}log_2\frac{2}{7}) = 0.0913$

$Gain(D, F_2, 42) = 0.971 + 0.4 * 2 * \frac{1}{2}log_2\frac{1}{2} + 0.6(\frac{1}{3}log_2\frac{1}{3} + \frac{2}{3}log_2\frac{2}{3}) = 0.0200$

$Gain(D, F_2, 46) = 0.971 + 0.5(\frac{2}{5}log_2\frac{2}{5} + \frac{3}{5}log_2\frac{3}{5}) + 0.5(\frac{1}{5}log_2\frac{1}{5} + \frac{4}{5}log_2\frac{4}{5}) =$

0.124

$Gain(D, F_2, 50) = 0.971 + 0.6(\frac{1}{2}log_2\frac{1}{2} * 2) + 0.4(\frac{1}{4}log_2\frac{1}{4} + \frac{3}{4}log_2\frac{3}{4}) = 0.146$

$Gain(D, F_2, 58.5) = 0.971 + 0.7(\frac{3}{7}log_2\frac{3}{7} + \frac{4}{7}log_2\frac{4}{7}) = 0.281$

$Gain(D, F_2, 71) = 0.971 + 0.8(2 * \frac{1}{2}\frac{1}{2}) = 0.171$

$Gain(D, F_2, 93.5) = 0.971 + 0.9(\frac{4}{9}log_2\frac{4}{9} + \frac{5}{9}log_2\frac{5}{9}) = 0.079$

综合选取使得 Gain(D,a,t) 为 Gain(D,$F_2$,32.5), 也即第一步采用 $F_2$ 划分, 分界点为 32.5

再根据 $F_1$ 进行划分, 对于 $F_2$ <= 32.5 的情况已经无需再划分, 对于 $F_2$ > 32.5 的情况,$F_1$ 的取值排序后为 22,24,25,32,43,48,52,53

候选划分点为 {23,24.5,28.5,37.5,45.5,50,52.5}

$Gain(D, F_1, 23) - Ent(D) = \frac{7}{8}(\frac{5}{7}log_2\frac{5}{7} + \frac{2}{7}log_2\frac{2}{7}) = -0.755$

$Gain(D, F_1, 24.5) - Ent(D) = \frac{3}{4}(\frac{1}{3}log_2\frac{1}{3} + \frac{2}{3}log_2\frac{2}{3}) = -0.68$

$Gain(D, F_1, 28.5) - Ent(D) = \frac{5}{8}(\frac{2}{5}log_2\frac{2}{5} + \frac{3}{5}log_2\frac{3}{5}) = -0.607$

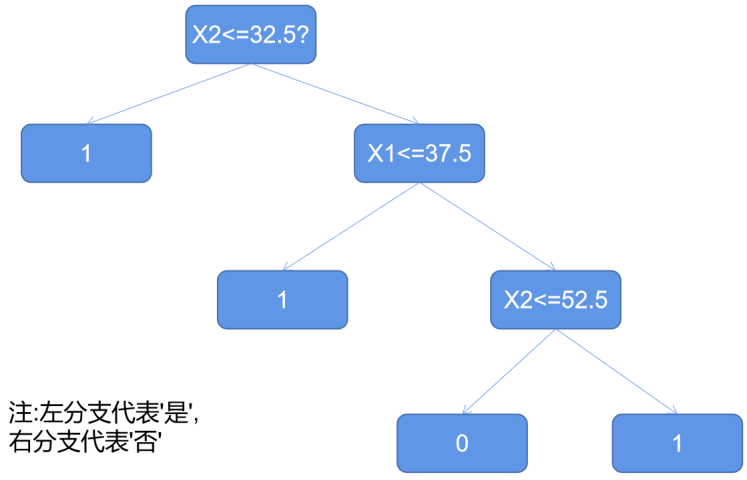
$Gain(D, F_1, 37.5) - Ent(D) = \frac{1}{2}(\frac{1}{4}log_2\frac{1}{4} + \frac{3}{4}log_2\frac{3}{4}) = -0.406$

$Gain(D, F_1, 45.5) - Ent(D) = \frac{3}{8}(\frac{1}{3}log_2\frac{1}{3} + \frac{2}{3}log_2\frac{2}{3}) + \frac{5}{8}(\frac{1}{5}log_2\frac{1}{5} + \frac{4}{5}log_2\frac{4}{5}) = -0.796$

$Gain(D, F_1, 50) - Ent(D) = \frac{1}{4}(\frac{1}{2}log_2\frac{1}{2} * 2) + \frac{3}{4}(\frac{5}{6}log_2\frac{5}{6} + \frac{1}{6}log_2\frac{1}{6}) = -0.738$

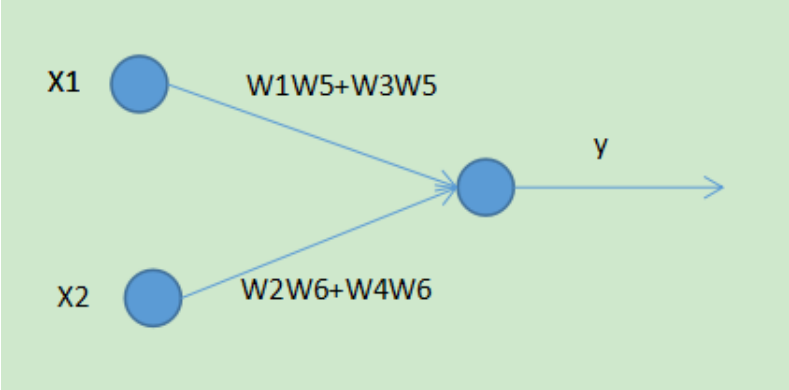
$Gain(D, F_1, 52.5) - Ent(D) = \frac{7}{8}(\frac{6}{7}log_2\frac{6}{7} + \frac{1}{7}log_2\frac{1}{7}) = -0.518$

综上选取 $F_1 = 37.5$ 作为一个划分点; 对于 $F_1$ <= 37.5 的情况已无需再划分, 对于 $F_1 > 37.5$ 的情况易观察到 $F_2 = 52.5$ 是一个最佳划分点, 综上决策树如下



## 2.[20pts]Neural Network

### (1)




### (2)

可以, 因为在这个计算过程中始终只会有输入变量 $X_1, ..., X_d$ 的一次项和常数项存在, 不会产生交叉项, 高次项或者非线性项, 因此可以不用隐层表示

### (3)

对于 Logistic Regression, 输入 $\mathbf{X} = X_1, ..., X_d$
输出为 $z = \mathbf{w}^T X + b$, 并用 sigmoid 函数 $y = \frac{1}{1+e^{-z}}$ 来将 z 值转化为一个接近 0 或 1 的 y 值
对于含有隐层的神经网络, 设置隐层同样含有 d 个神经元, 输出记为 $H_1, H_2, ..., H_d$, 第 i 个输入 $X_i$ 和第 j 个隐层神经元 $H_j$ 之间的权重设为 $a_{ij}$, 隐层采用线性激励函数 $f(x) = x$, 此时有如下关系

$$\begin{bmatrix} a_{11} & ... & a_{d1} \\ ... & & ... \\ a_{1d} & ... & a_{dd} \end{bmatrix} \Big[X_1, ..., X_d\Big]^T = \Big[H_1, ..., H_d\Big]^T \tag{1}$$

设第 i 个隐层神经元和输出神经元之间连接的权重为 $w_i$, 输出激励函数选取为 sigmoid 函数, 阈值记作 b

此时有

$$z = \left[H_1, ..., H_d\right] \left[w_1, ..., w_d\right]^T - b \tag{2}$$

$$y = \frac{1}{1 + e^{-z}} \tag{3}$$

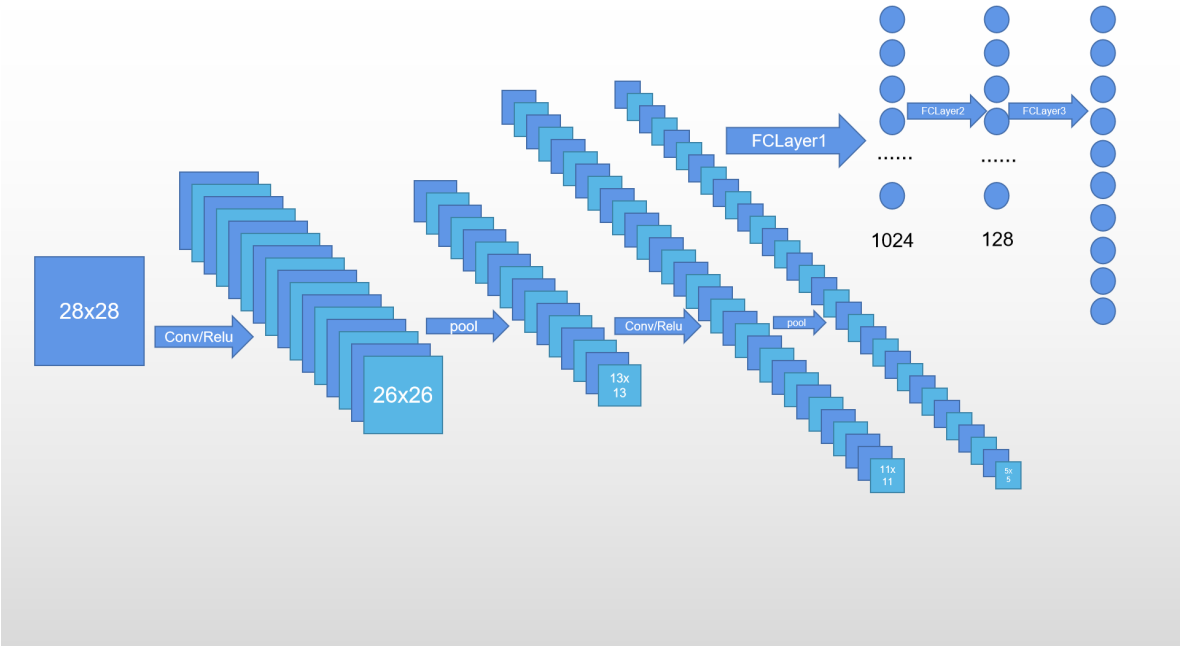
综合 (1) 式 (2) 式有

$$z = \begin{bmatrix} a_{11} & ... & a_{d1} \\ ... & & ... \\ a_{1d} & ... & a_{dd} \end{bmatrix} \left[X_1, ..., X_d\right]^T \left[w_1, ...w_d\right]^T - b = \sum_{i=1}^{d} (\sum_{j=1}^{d} w_j a_{ij} X_i) - b \tag{4}$$

根据看到此时的 z 和我们所要求得对数几率回归中的 z 是同一形式, 在神经网络训练完成后, 只要按照 (4) 式进行计算即可得到对应的对数几率回归的 z 函数的系数

## 3.[60pts]Neural Network in Practice

## (2)

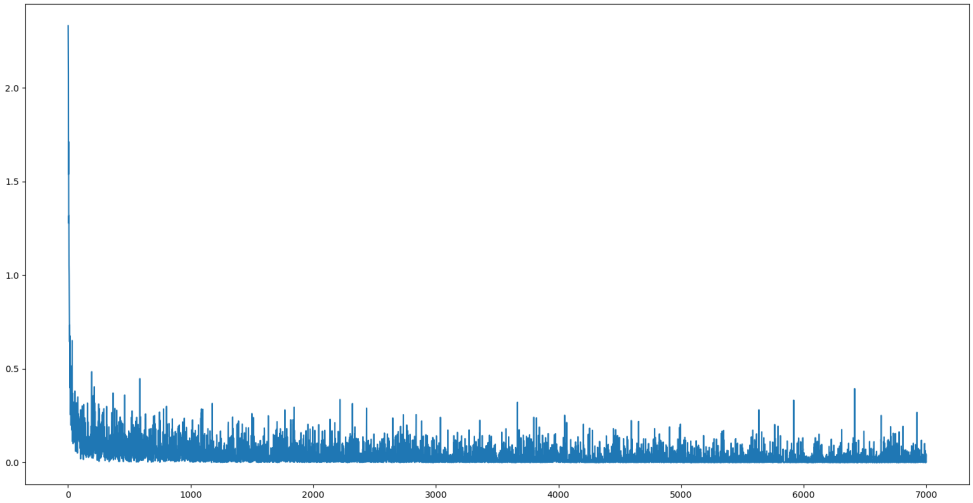



## (3)

选取 epochs∈{1,2,3,4,5,6,7,8},Learning Rate∈{0.001,0.002,0.003,0.004,0.005}
进行测试
具体数据见附表 params2.txt
综合 Accuracy,AverageLoss 以及 Epochs 考虑, 选取
Epochs=7,Learning Rate=0.002 进行作为该模型的参数

## (4)

对于 training_loss 的变化, 我采用对每一批 (60 个) 数据都记录一次 loss 来刻画, 并使用
matplotlib.pylab 来绘制图像
图大致如下, 具体可见附件 training_loss.png


2.0
1.5
1.0
0.5
0.0
0
1000
2000
3000
4000
5000
6000
7000