数据挖掘作业二

高阳 15331089

1. 模型的性能度量

（1）对 M1，取阈值为 0.5，计算准确率，查准率，查全率（真正例率，TPR），假正例率（FPR）和 F-measure。

由表画出混淆矩阵对应的表格：

M1(阈值= 0.5)	PREDICTED CLASS		
ACTUAL CLASS		TRUE CLASS = +	TRUE CLASS = -
	TRUE CLASS = +	2(TP)	2(FN)
	TRUE CLASS = -	2(FP)	4(TN)

Accuracy = (TP + TN) / (TP + TN + FP + FN) = 6 / 10 = 0.6

Precision =TP / (TP + FP) = 2 / 4 = 0.5

Recall = TPR = TP / (TP + FN) = 2 / 4 = 0.5

FPR = FP / (FP + TN) = 2 / 6 = 0.33

F-measure = 2 * Precision * Recall / (Precision + Recall) = 0.5

（2）对 M2，按照（1）中要求进行同样计算。

由表画出混淆矩阵对应的表格：

M2(阈值= 0.5)	PREDICTED CLASS		
ACTUAL CLASS		TRUE CLASS = +	TRUE CLASS = -
	TRUE CLASS = +	1(TP)	3(FN)
	TRUE CLASS = -	1(FP)	5(TN)

Accuracy = (TP + TN) / (TP + TN + FP + FN) = 6 / 10 = 0.6

Precision =TP / (TP + FP) = 1 / 2 = 0.5

Recall = TPR = TP / (TP + FN) = 1 / 4 = 0.25

FPR = FP / (FP + TN) = 1 / 6 = 0.17

G-measure = 2 * Precision * Recall / (Precision + Recall) = 0.33

（3）对 M1，取阈值为 0.2，分别进行上述计算。并讨论阈值为 0.2 或 0.5 时，

哪个对 M1 的分类结果更好。

M1(阈值= 0.2)	PREDICTED CLASS		
ACTUAL CLASS		TRUE CLASS = +	TRUE CLASS = -
	TRUE CLASS = +	4(TP)	0(FN)
	TRUE CLASS = -	4(FP)	2(TN)

Accuracy = (TP + TN) / (TP + TN + FP + FN) = 6 / 10 = 0.6

Precision =TP / (TP + FP) = 4 / 8 = 0.5

Recall = TPR = TP / (TP + FN) = 4 / 4 = 1.0

FPR = FP / (FP + TN) = 4 / 6 = 0.67

F-measure = 2 * Precision * Recall / (Precision + Recall) = 0.67

对 M1(阈值= 0.5)有(FPR, TPR) = (0.33, 0.5)

对 M1(阈值= 0.2)有(FPR, TPR) = (0.67, 1.0)

若要分类效果越好，则(FPR, TPR) -> (0, 1)，对比上面两点到(0, 1)的距离知阈值为 0.5 更好。

（4）是否存在更好阈值？若存在求出最优阈值。

存在。

可以设置多个阈值参数画出 ROC 曲线进行比较。

对 M1，从题目表中看出 TRUE CLASS = +的最小 scores 为 0.45，而 TRUE CLASS = -的最大 scores 为 0.67，且有两个 score 大于 0.45，所以不可能找到一个阈值使得样本完美分类。所以我们取这几个阈值进行比较：

A:0.46(0.45-0.47)

B:0.5(0.47-0.55)

C:0.6(0.55-0.67)

对 A:(FPR, TPR) = (0.33, 0.75)

对 B:(FPR, TPR) = (0.33, 0.5)

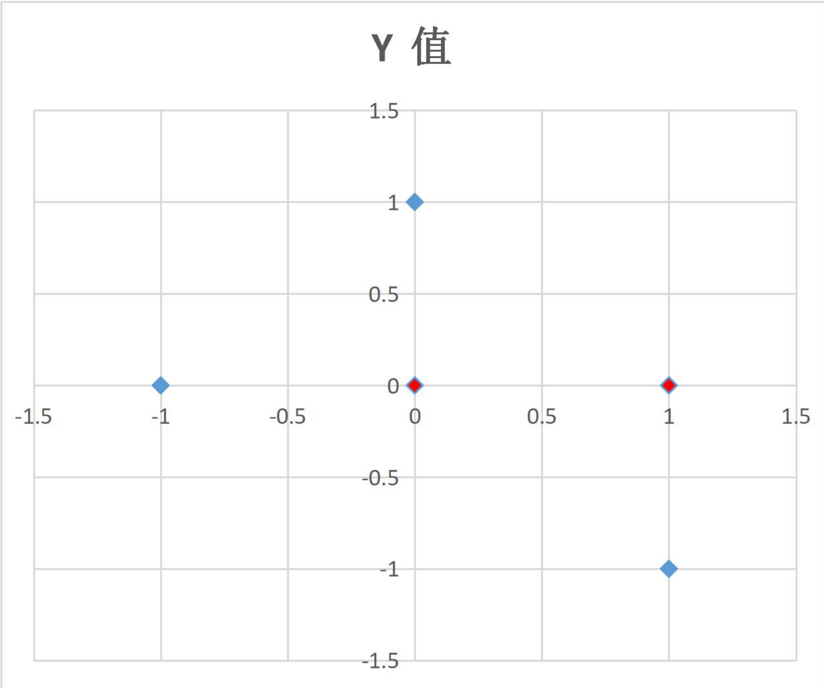
对 C:(FPR, TPR) = (0.17, 0.5)

分别计算它们到(0, 1)的距离得 0.6 更好。即最优阈值在(0.55-0.67)之间。

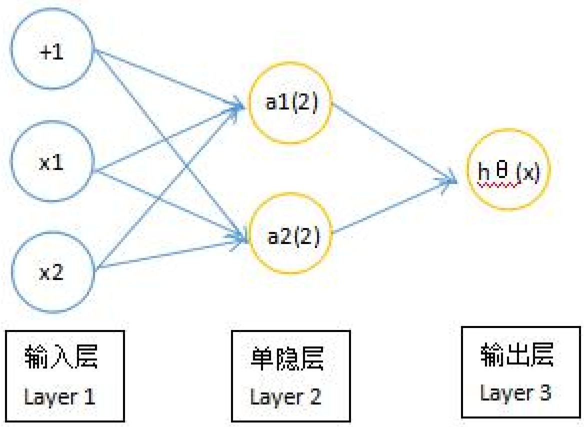
2. 神经网络

（1）如图：

线性不可分。



（2）分析将 Sigmoid 激活函数换成线性函数的缺陷。



Sigmoid 函数值域在 0-1 之间，符合任何概率模型，且关于 y 轴呈中心对称。若将其换成线性函数，则必须确定值域范围，且结果必须进行归一化才能更好地比较。同时做为神经网络的激活函数，需经过多次迭代，会产生很多额外开销。

（3）令初始参数全为 0，运用前馈(feedfoward)算法计算在初始化参数下此三层神经网络的输出；运用反向传播(backpropagation)算法，计算代价函数对所有参

数的偏导数，讨论将参数初始化全部设为 0 所带来的问题。

令：

g(z) = 1 / (1 + e ^ (-z))

X = [X0 X1 X2]T

θ1(1) = [0 0 0]

θ2(1) = [0 0 0]

θ1(2) = [0 0 0]

A(2) = [a0(2) a1(2) a2(2)]

前馈算法：

a0(2) = 1

a1(2) = g(θ10(1) * x0 + θ11(1) * x1 + θ12(1) * x2) = g(θ1(1) * X) = g(0) = 0.5

a2(2) = g(θ20(1) * x0 + θ21(1) * x1 + θ22(1) * x2) = g(θ2(1) * X) = g(0) = 0.5

则 A(2) = [1 0.5 0.5]

则：

hθ(x) = a1(3)

= g(θ10(2) * a0(2) + θ11(2) * a1(2) + θ12(2) * θa2(2))

= g(θ1(2) * A(2))

= g(0) = 0.5

反向传播算法：

利用前面第一次迭代得出的结果 0.5 进行反向传播。

设βj(l)为 l 层上面节点 j 输出值与真实值的误差,则：

β(3) = hθ(x) - y

β1(2) = (θ1(2))T * β(3) .* g'(θ1(1) * X) = (θ1(2))T * β(3) .* X .* (1 -X)

β2(2) = (θ2(2))T * β(3) .* g'(θ2(1) * X) = (θ2(2))T * β(3) .* X .* (1 -X)

则开销函数 J(θ)关于每一系数θ的偏导数：

J'(θij(l)) = aj(l) * βi(l + 1)

利用 matlab 进行辅助计算：

x1,x2,x3 为三个输入值，y 为输出值，w 为到第一层的权重(2X3)，v 为到第二层的权重(1X3)

```
x1 = [1,1,1,1,1];
x2 = [0,1,0,-1,1];
x3 = [0,0,1,0,-1];
y = [1,1,0,0,0];
data = [x1;x2;x3;y];
[m, n] = size(data);
w = [0,0,0;0,0,0];
v = [0,0,0];
```

BP 算法：

```
%BackPropagation algorithm
delta3 = o - y;
for i = 1:3
    delta2(i) = hidden1(i) * (1 - hidden1(i)) * delta3 * v(:, i);
end
```

更新权重：

```
%renew weight
for i = 1:2
    for j = 1:3
        w(i, j) = w(i, j) + delta2(i) * x(j);
    end
end

for i = 1:3
    v(i) = v(i) + delta3 * hidden1(1);
end
```

经过 5 个测试样例的迭代，最后输出的 w 为：

	1	2	3
1	-0.0535	0.0453	-0.0083
2	-0.0535	0.0453	-0.0083

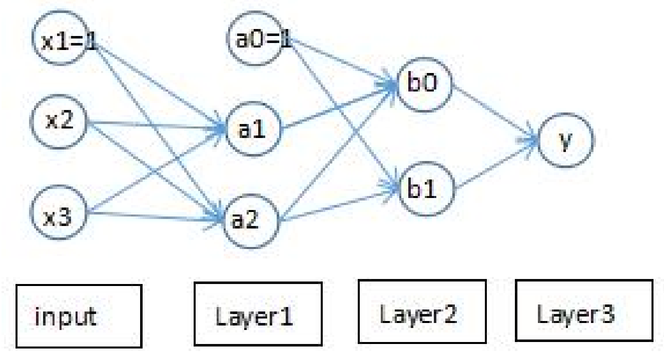
v 为：

	1	2	3
1	-0.1070	-0.1070	-0.1070

将初始化参数全部设为 0，使得第一次迭代的所有结果均不受输入值的影响，而从第二次迭代才开始受影响，所以相当于多进行了一次迭代。

（4）

如图



激活函数仍为 sigmoid 函数。

3. 决策树

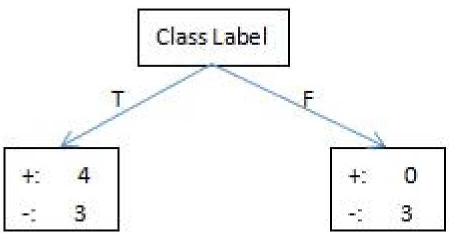
（1）计算以属性 A 或 B 为划分的信息熵增益，并说明决策树学习算法选择哪个属性进行划分。

$$GAIN_{split} = Entropy(p) - \left(\sum_{i=1}^{k} \frac{n_i}{n} Entropy(i) \right)$$

Parent Node, p is split into k partitions;
n_i is number of records in partition i

Entropy(p) = -0.4 * log2(0.4) - 0.6 * log2(0.6) = 0.528 + 0.442 = 0.97

对 A：

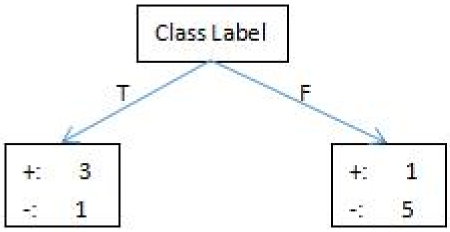


Entropy(T) = -0.57 * log2(0.57) - 0.43 * log2(0.43) = 0.46 + 0.52 = 0.98

Entropy(F) = 0

GAIN(A) = Entropy(p) - (0.7 * Entropy(T) + 0.3 * Entropy(F)) = 0.97 - 0.7 * 0.98 = 0.284

对 B：



Entropy(T) = -0.75 * log2(0.75) - 0.25 * long2(0.25) = 0.31 + 0.5 = 0.81

Entropy(F) = -0.167 * log2(0.167) - 0.833 * log2(0.833) = 0.43 + 0.22 = 0.65

GAIN(B) = Entropy(p) - (0.4 * Entropy(T) + 0.6 * Entropy(F)) = 0.97 - 0.324 - 0.39 = 0.256

（2）Gini 增益

对 A：

GINI(T) = 1 - 0.32 - 0.18 = 0.5

GINI(F) = 1 - 1 = 0

GAIN(A) = 0.7 * GINI(T) + 0.3 * GINI(F) = 0.7 * 0.5 = 0.35

对 B：

GINI(T) = 1 - 0.56 - 0.0625 = 0.375

GINI(F) = 1 - 0.027889 - 0.693889 = 0.278

GAIN(B) = 0.4 * GINI(T) + 0.6 * GINI(F) = 0.15 + 0.1668 = 0.32

（3）分类误差增益

对 A：

ERROR(T) = 1 - MAX(0.57, 0.43) = 1 - 0.57 = 0.43

ERROR(F) = 1 - MAX(0, 1) = 0

GAIN(A) = 0.7 * ERROR(T) + 0.3 * ERROR(F) = 0.301

对 B:

ERROR(T) = 1 - MAX(0.75, 0.25) = 1 - 0.75 = 0.25

ERROR(F) = 1 - MAX(0.167, 0.833) = 0..167

GAIN(B) = 0.4 * 0.25 + 0.6 * 0.167 = 0.2

（4）

答：由上面的结果可以看到，信息熵增益的结果较小，GINI 增益的结果较大，而分类误差增益的结果波动较大。

信息熵增益对分类率取 log，所以分类率越小结果的绝对值越大，影响就越大。

GINI 增益对分类率取平方，所以当分类率越小时，结果也越小，影响越小。

分类误差增益只考虑分类样本数多的属性，所以误差较大。