Q1

$TP = 83 \quad TN = 112 \quad FP = 58 \quad FN = 21$

$P = 104 \quad N = 170 \quad All = 274$

$\text{Sensitivity} = \frac{TP}{P} = \frac{83}{104} \approx 0.80$

$\text{Specificity} = \frac{TN}{N} = \frac{112}{170} \approx 0.66$

$\text{Precision} = \frac{TP}{TP+FP} = \frac{83}{83+58} \approx 0.59$

$\text{Accuracy} = \frac{(TP+TN)}{All} = \frac{83+112}{274} \approx 71\%$

$\text{Recall} = \frac{TP}{TP+FN} = \frac{83}{83+21} \approx 0.80$

$\text{F-meassure} = \frac{2 \times \text{Precision} \cdot \text{Recall}}{\text{precision} + \text{recall}} = \frac{2 \times 0.59 \cdot 0.8}{0.59 + 0.8} \approx 0.68$

$F_\beta = F_2 = \frac{(1+2^2) \cdot 0.59 \cdot 0.8}{2^2 \cdot 0.59 \cdot 0.8} \approx 0.75$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

$$= \frac{83 \times 112 - 58 \cdot 21}{\sqrt{(83+58)(83+21)(112+58)(112+21)}}$$

≈ 0.44

Q2 $\overline{err}(M1) = (0.08 + 0.12 + 0.10 + 0.11 + 0.10 + 0.16 + 0.17 + 0.12 + 0.11 + 0.21) \div 10$

$= 0.128$

$\overline{err}(M2) = (0.13 + 0.1 + 0.12 + 0.12 + 0.15 + 0.13 + 0.09 + 0.21 + 0.22 + 0.17) \div 10$

$= 0.144$

$Var(M1) = \dfrac{(-0.048)^2 + (-0.008)^2 + (-0.028)^2 + (-0.018)^2 + (-0.028)^2 + (0.032)^2 + (0.042)^2 + (-0.008)^2 + (-0.018)^2 + (0.082)^2}{9}$

$= 0.00157$

$Var(M2) = \dfrac{(0.13-0.144)^2 + (0.1-0.144)^2 + (0.12-0.144)^2 + (0.12-0.144)^2 + (0.15-0.144)^2 + (0.13-0.144)^2 + (0.09-0.144)^2 + (0.21-0.144)^2 + (0.22-0.144)^2 + (0.17-0.144)^2}{9}$

$= 0.00191$

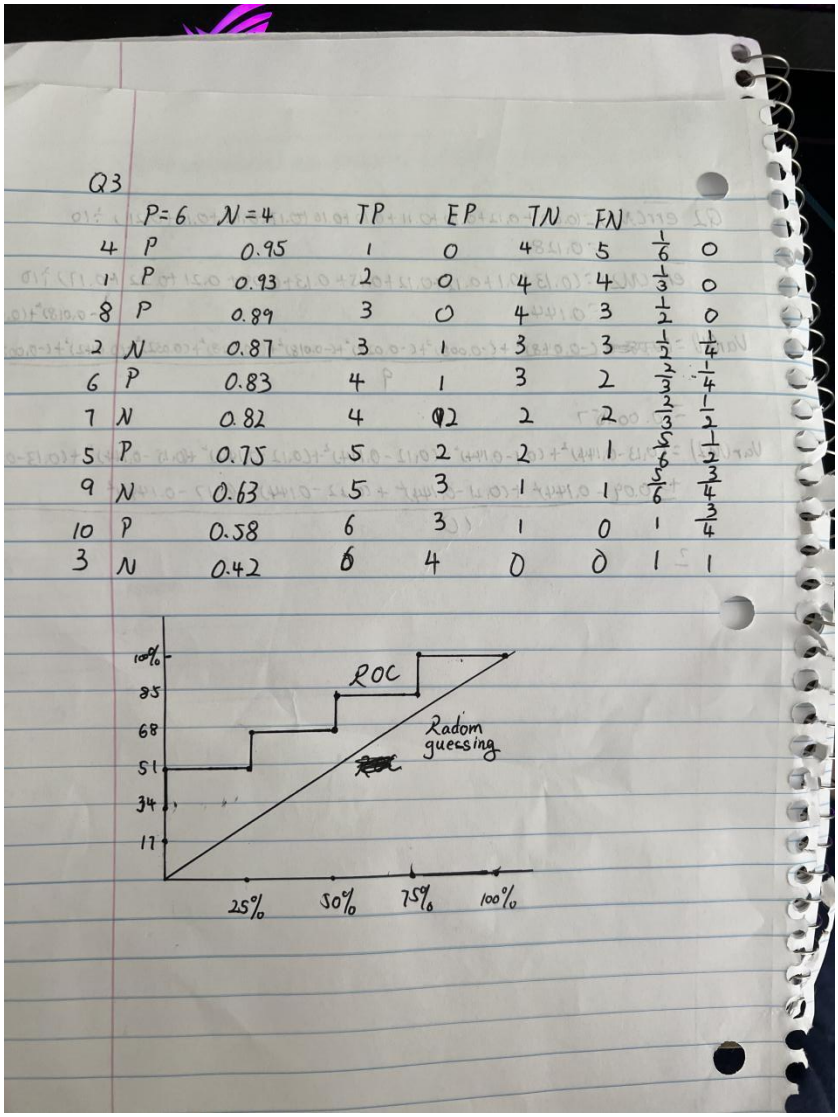
$t_0 = \dfrac{0.128 - 0.144}{\sqrt{\dfrac{0.00157 - 0.00191}{10}}}$

≈ 2.743

$t_{0.005, 9} = 3.25$

$2.743 < 3.25$

So M_1 is Better



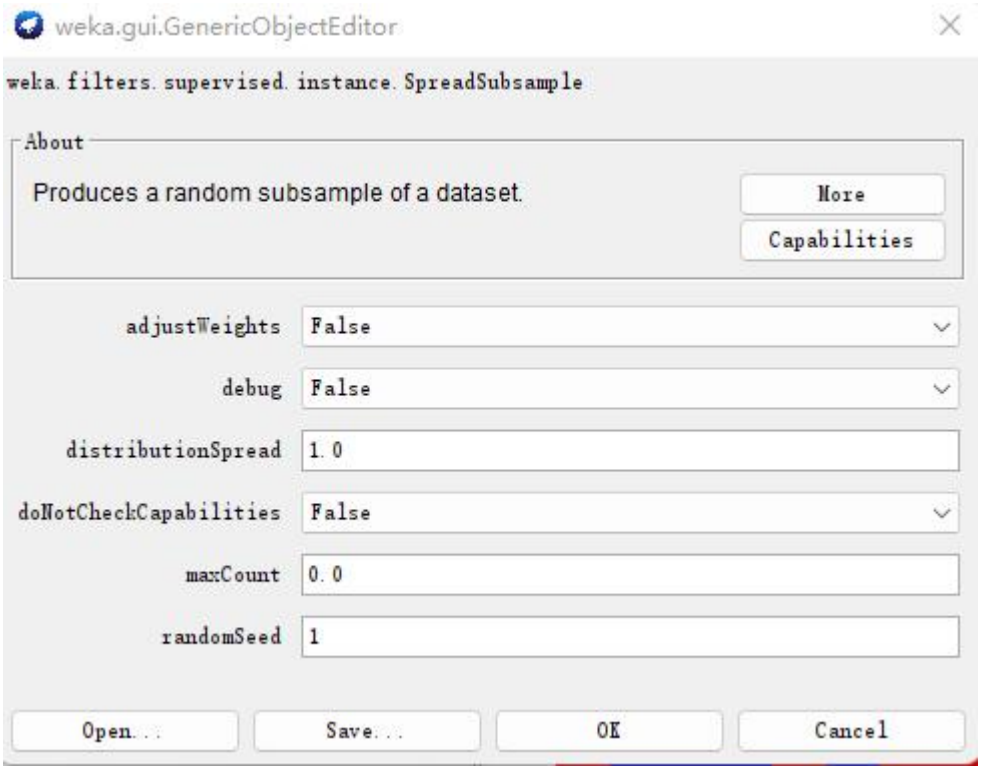
Q4

Original

```
=== Summary ===

Correctly Classified Instances        1373               89.3299 %
Incorrectly Classified Instances       164               10.6701 %
Kappa statistic                          0.3483
Mean absolute error                      0.158
Root mean squared error                  0.2958
Relative absolute error                 77.4244 %
Root relative squared error             92.6749 %
Total Number of Instances             1537

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.969    0.689    0.915      0.969    0.941      0.367    0.794     0.952     no
                 0.311    0.031    0.567      0.311    0.401      0.367    0.794     0.353     yes
Weighted Avg.    0.893    0.613    0.875      0.893    0.879      0.367    0.794     0.883

=== Confusion Matrix ===

    a    b   <-- classified as
 1318   42 |    a = no
  122   55 |    b = yes
```

Undersampled

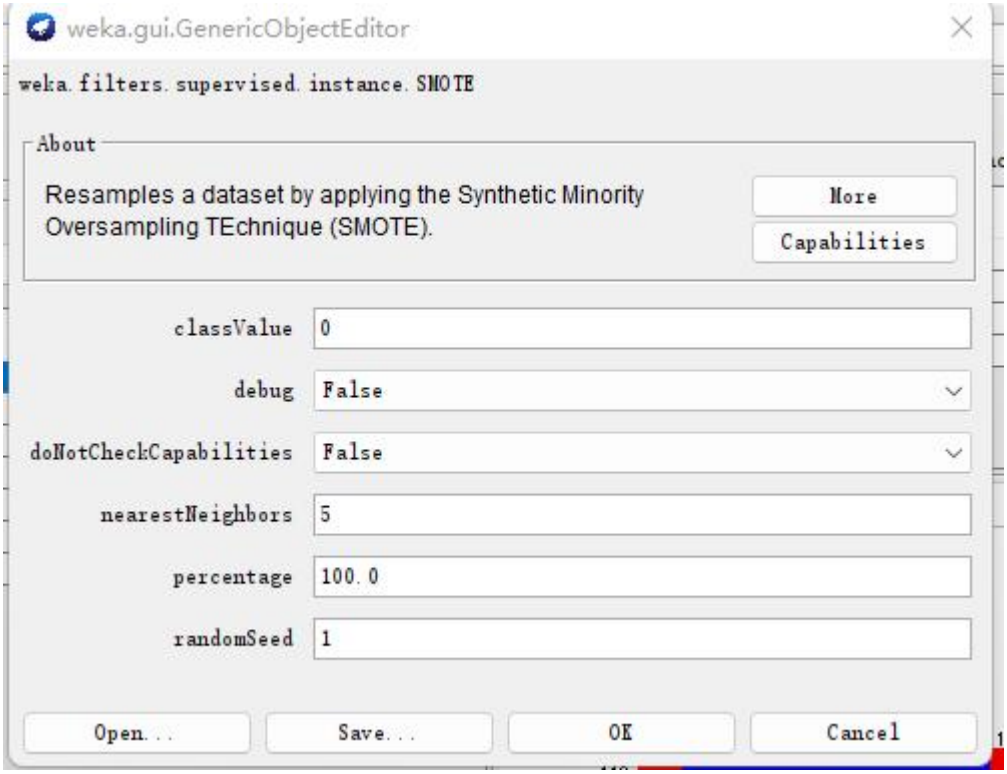
Open train.aff, click choose,click filter,click supervised,click instance choose Spread sub sample.All data detail is showed below




```
=== Summary ===

Correctly Classified Instances        1130               73.5198 %
Incorrectly Classified Instances       407               26.4802 %
Kappa statistic                          0.315
Mean absolute error                      0.2841
Root mean squared error                  0.4558
Relative absolute error                 56.8273 %
Root relative squared error             91.1537 %
Total Number of Instances             1537

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.716    0.119    0.979      0.716    0.827      0.399    0.823     0.968     no
                 0.881    0.284    0.288      0.881    0.434      0.399    0.823     0.287     yes
Weighted Avg.    0.735    0.138    0.899      0.735    0.782      0.399    0.823     0.890

=== Confusion Matrix ===

   a   b   <-- classified as
 974 386 |   a = no
  21 156 |   b = yes
```

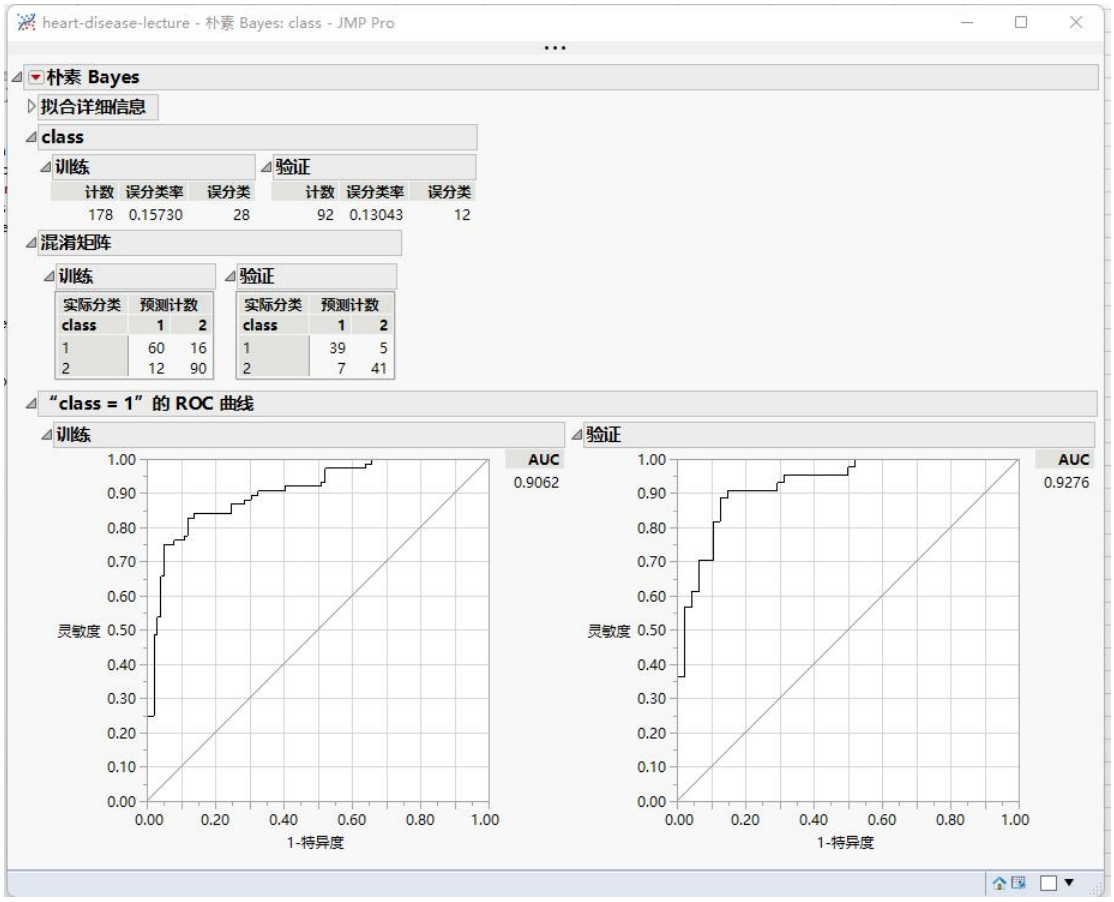
Oversampled

Open train.aff, click choose,click filter,click supervised,click instance choose smote .All data detail is showed below




```
Correctly Classified Instances        1332               86.6623 %
Incorrectly Classified Instances       205               13.3377 %
Kappa statistic                          0.3503
Mean absolute error                      0.164
Root mean squared error                  0.3211
Relative absolute error                 59.7727 %
Root relative squared error             96.6788 %
Total Number of Instances             1537

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.924    0.571    0.926      0.924    0.925      0.350    0.788     0.948     no
                 0.429    0.076    0.422      0.429    0.426      0.350    0.788     0.362     yes
Weighted Avg.    0.867    0.514    0.868      0.867    0.867      0.350    0.788     0.881

=== Confusion Matrix ===

    a    b   <-- classified as
 1256  104 |    a = no
  101   76 |    b = yes
```

Q5


heart-disease-lecture - 朴素 Bayes: class - JMP Pro
朴素 Bayes
拟合详细信息
class
训练
计数 误分类率 误分类
178 0.15730 28
验证
计数 误分类率 误分类
92 0.13043 12
混淆矩阵
训练
实际分类 预测计数
class 1 2
1 60 16
2 12 90
验证
实际分类 预测计数
class 1 2
1 39 5
2 7 41
“class = 1” 的 ROC 曲线
训练
AUC
0.9062
灵敏度
1-特异度
验证
AUC
0.9276
灵敏度
1-特异度

Naïve

	1	2
1	39	5
2	7	41

	2	1
2	41	7
1	5	39.

Class 1

TP: 39 FN: 5 FP: 7 TN: 41

Precision : $\frac{39}{46} \approx 0.84$

Recall : $\frac{39}{44} \approx 0.87$

$F = \frac{2 \cdot 0.84 \cdot 0.87}{0.84 + 0.87} \approx 0.85$

$MCC = \frac{(39 \cdot 41) - (5 \times 7)}{\sqrt{46 \cdot 44 \cdot 48 \cdot 46}}$

≈ 0.74

Class 2

TP: 41 FN: 7 FP: 5 TN: 39

Precision : $\frac{41}{46} \approx 0.89$

Recall = $\frac{41}{48} \approx 0.85$

$F = \frac{2 \cdot 0.89 \cdot 0.85}{0.89 + 0.85} \approx 0.87.$

$MCC = \frac{41 \cdot 39 - 35}{\sqrt{46 \cdot 48 \cdot 44 \cdot 46}}$

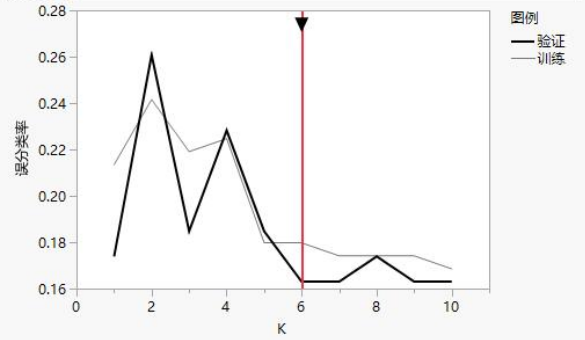
≈ 0.74

heart-disease-lecture - K 最近邻: class - JMP Pro

K 最近邻

class

模型选择



训练

K	计数	误分类率	误分类	
1	178	0.21348	38	
2	178	0.24157	43	
3	178	0.21910	39	
4	178	0.22472	40	
5	178	0.17978	32	
6	178	0.17978	32	
7	178	0.17416	31	
8	178	0.17416	31	
9	178	0.17416	31	
10	178	0.16854	30	*

验证

K	计数	误分类率	误分类	
1	92	0.17391	16	
2	92	0.26087	24	
3	92	0.18478	17	
4	92	0.22826	21	
5	92	0.18478	17	
6	92	0.16304	15	*
7	92	0.16304	15	
8	92	0.17391	16	
9	92	0.16304	15	
10	92	0.16304	15	

最佳 K=6 的混淆矩阵

训练

实际分类	预测计数	
class	1	2
1	59	17
2	15	87

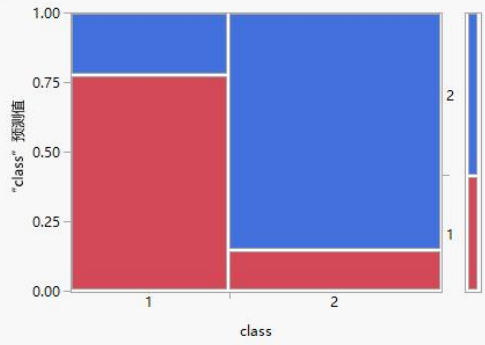
验证

实际分类	预测计数	
class	1	2
1	34	10
2	5	43

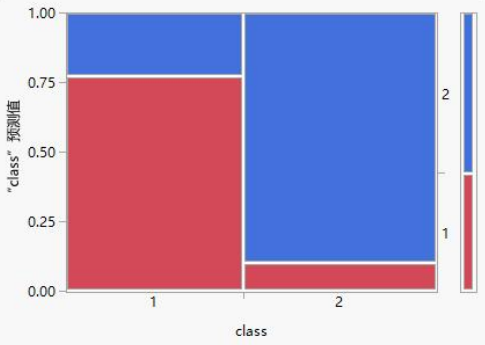
马赛克图

马赛克图，K=6

训练



验证



knn

Best k value is k=6

	1	2
1	34	10
2	5	43

	2	1
2	43	5
1	10	34

Class 1 TP: 34 FN: 10 FP: 5 TN: 43

Precision: $\frac{34}{39} \approx 0.87$

Recall: $\frac{34}{44} \approx 0.77$

$F = \frac{0.87 \cdot 0.77 \cdot 2}{0.87 + 0.77} \approx 0.82$

$MCC = \frac{34 \cdot 43 - 50}{\sqrt{39 \cdot 44 \cdot 48 \cdot 53}}$

≈ 0.68

Class 2 TP: 43 FN: 5 FP: 10 TN: 34

Precision: $\frac{43}{53} \approx 0.81$

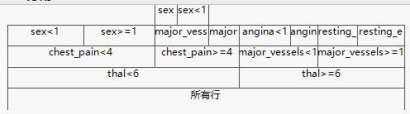
Recall: $\frac{43}{48} \approx 0.90$

$F = \frac{2 \cdot 0.81 \cdot 0.9}{1.71} \approx 0.85$

$MCC = \frac{43 \cdot 34 - 50}{\sqrt{39 \cdot 44 \cdot 48 \cdot 53}}$

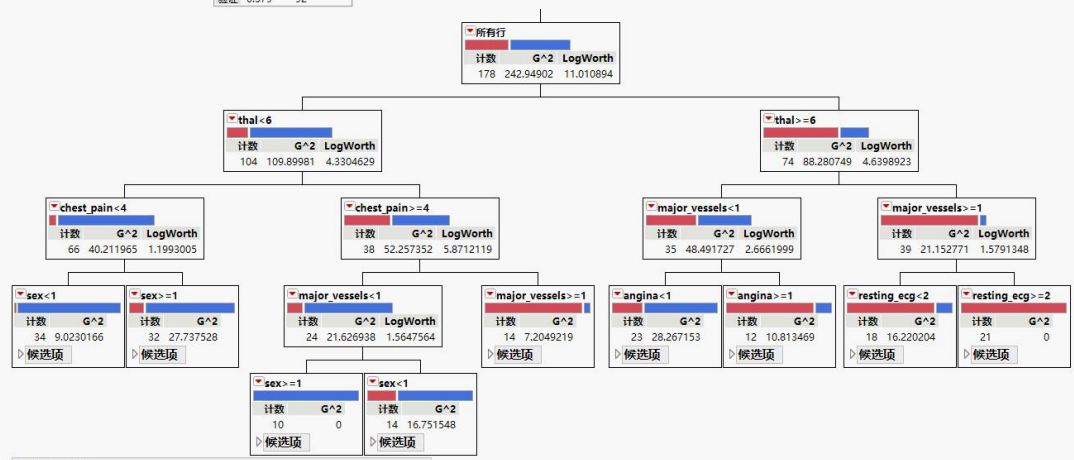
≈ 0.68

“class” 分割

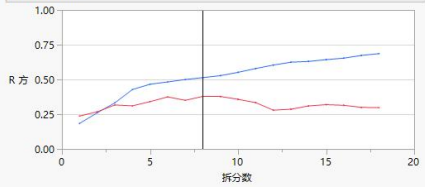


拆分 | 剪除 | 执行 | 点着色

	R 方	数目	拆分数
训练	0.514	178	8
验证	0.379	92	



拆分历史记录



以红色显示验证数据

拟合详细信息

测度	训练	验证	定义
熵 R 方	0.5138	0.3790	1-Loglike(model)/Loglike(0)
广义 R 方	0.6770	0.5447	(1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
-Log p 均值	0.3318	0.4299	∑ -Log(p[j])/n
RASE	0.3186	0.3701	√ ∑(y[j]-p[j])²/n
绝对偏差的均值	0.2150	0.2484	∑ |y[j]-p[j]|/n
误分类率	0.1292	0.1739	∑ (p[j]≠pMax)/n
数目	178	92	n

混淆矩阵

训练

实际分类	预测计数	
class	1	2
1	59	17
2	6	96

验证

实际分类	预测计数	
class	1	2
1	36	8
2	8	40

	1	2
1	39	5
2	7	41

	2	1
2	41	7
1	5	39

Class 1
: 39 FN: 5
Precision -

Partition

	1	2
1	36	8
2	8	40

	2	1
2	40	8
1	8	36

Class 1

TP: 36 FN: 8 FP: 8 TN: 40

Precision: $\frac{36}{44} \approx 0.82$

Recall: $\frac{36}{44} \approx 0.82$

$F = \frac{0.82 \cdot 0.82 \cdot 2}{1.64} \approx 0.82$

$MCC = \frac{36 \cdot 40 - 64}{\sqrt{44 \cdot 44 \cdot 48 \cdot 48}} \approx 0.65$

Class 2

TP: 40 FN: 8 FP: 8 TN: ~~36~~

Precision: ~~$\frac{36}{44} \approx 0.82$~~ $\frac{40}{48} \approx 0.83$

Recall: ~~$\frac{36}{44} \approx 0.82$~~ $\frac{40}{48} \approx 0.83$

F ≈ ~~0.82~~ $\frac{2 \cdot 0.83 \cdot 0.83}{0.83 + 0.83} \approx 0.83$

$MCC \approx \frac{40 \cdot 36 - 64}{\sqrt{44 \cdot 44 \cdot 48 \cdot 48}} \approx 0.65$

“class”的提升树

规格

目标	class	训练行数:	178
验证列:	Validation	验证行数:	92
层数:	40		
每树拆分数:	3		
学习率:	0.1		
过拟合惩罚:	0.0001		

总体统计量

测度	训练	验证	定义
熵 R 方	0.6220	0.4575	1-Loglike(model)/Loglike(0)
广义 R 方	0.7684	0.6260	(1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
-Log p 均值	0.2580	0.3755	∑ -Log(ρ[j])/n
RASE	0.2679	0.3520	√ ∑(y[j]-ρ[j])²/n
绝对偏差的均值	0.1908	0.2467	∑ |y[j]-ρ[j]|/n
误分类率	0.0843	0.1630	∑ (ρ[j]≠ρMax)/n
数目	178	92	n

混淆矩阵

训练

实际分类	预测计数	
class	1	2
1	65	11
2	4	98

验证

实际分类	预测计数	
class	1	2
1	38	6
2	9	39

1 Naive

	1	2	
1	39		

Boosted tree

	1	2
1	38	6
2	9	39

	2	1
2	39	9
1	6	38

Class 1

TP: 38 FN: 6 FP: 9 TN: 39

Precision: $\frac{38}{47} \approx 0.81$

Recall: $\frac{38}{44} \approx 0.80$

$F = \frac{2 \cdot 0.8 \cdot 0.81}{1.61} \approx 0.8$

$MCC = \frac{38 \cdot 39 - 54}{\sqrt{44 \cdot 47 \cdot 48 \cdot 45}}$

≈ 0.68

Class 2

TP: 39 FN: 9 FP: 6 TN: 38

Precision: $\frac{39}{45} \approx 0.87$

Recall: $\frac{39}{48} \approx 0.81$

$F = \frac{2 \cdot 0.87 \cdot 0.81}{1.68} \approx 0.84$

$MCC = \frac{38 \cdot 39 - 54}{\sqrt{44\ 47 \cdot 48 \cdot 45}}$

≈ 0.68

heart-disease-lecture - 神经: class - JMP Pro

神经

验证列: Validation

模型启动

模型 NTanH(5)

训练

class

测量	值
广义 R 方	0.6787613
熵 R 方	0.5157974
RMSE	0.3231021
绝对偏差的均值	0.212372
误分类率	0.1404494
-对数似然	58.818279
频数总和	178

混淆矩阵

实际分类	预测计数	
class	1	2
1	65	11
2	14	88

混淆率

实际分类	预测率	
class	1	2
1	0.855	0.145
2	0.137	0.863

验证

class

测量	值
广义 R 方	0.6545488
熵 R 方	0.4872319
RMSE	0.3277325
绝对偏差的均值	0.2069902
误分类率	0.1413043
-对数似然	32.654386
频数总和	92

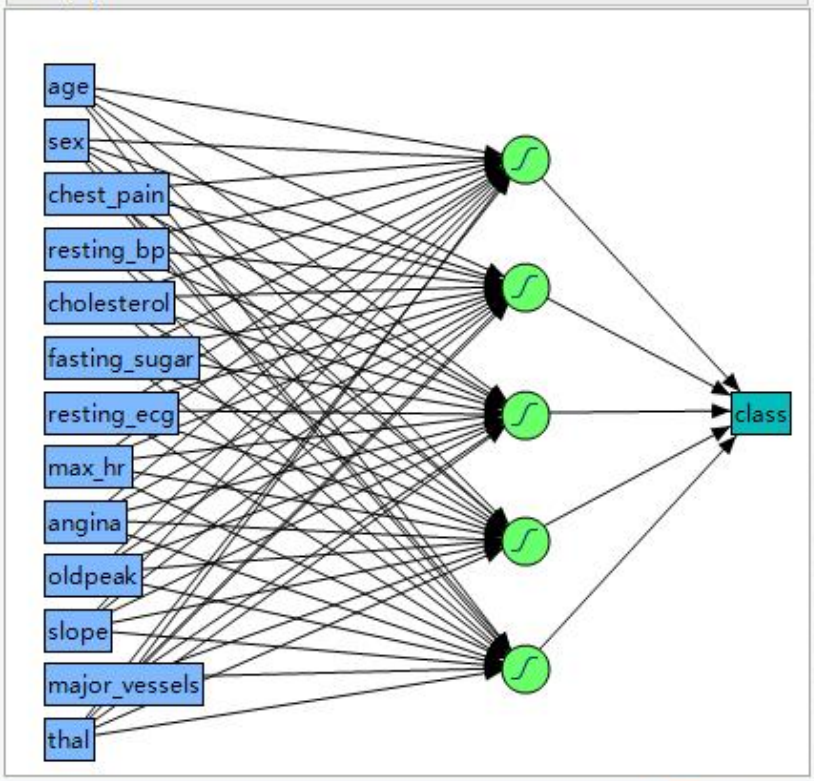
混淆矩阵

实际分类	预测计数	
class	1	2
1	38	6
2	7	41

混淆率

实际分类	预测率	
class	1	2
1	0.864	0.136
2	0.146	0.854

关系图



Neural

	1	2
1	38	6
2	7	41

	2	1
2	41	7
1	6	38

Class 1

TP: 38 FN: 6 FP: 7 TN: 41

Precision : $\frac{38}{45} \approx 0.84$

Recall : $\frac{38}{44} \approx 0.86$

$F = \frac{2 \cdot 0.84 \cdot 0.86}{1.7} \approx$ 0.85

$MCC = \frac{38 \cdot 41 - 42}{\sqrt{45 \cdot 44 \cdot 47 \cdot 48}}$

≈ 0.72

Class

TP: 41 FN: 7 FP: 6 TN: 38

Precision : $\frac{41}{47} \approx 0.87$

Recall : $\frac{41}{48} \approx 0.85$

$F = \frac{2 \cdot 0.87 \cdot 0.85}{1.72} \approx 0.86$

$MCC = \frac{41 \cdot 38 - 42}{\sqrt{45 \cdot 44 \cdot 47 \cdot 48}}$

≈ 0.72

There are 2 hidden layers