Classification Trees

Best off-the shelf classifier?

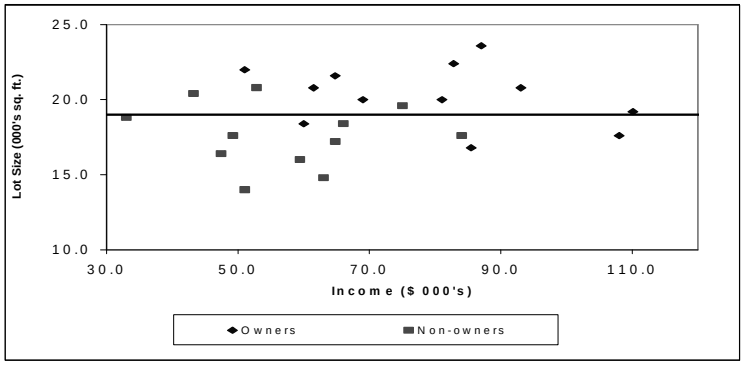
- (Arguably) Classification Trees
  - Robust to outliers, handles missing data well, easy interpretation, requires little knowledge of statistics
- Recursive partitioning
  - Split criteria (e.g. Gini, entropy, likelihood)
  - Number of branches at a split
  - Stop growing logic
    - Lower limit on #cases at node, Maximum depth of node, Chi-squared test (CHAID)
  - Pruning Logic
    - Use validation data, node penalty (CART: Brieman, Friedman, Olshen, Stone 1983)

Classification Trees

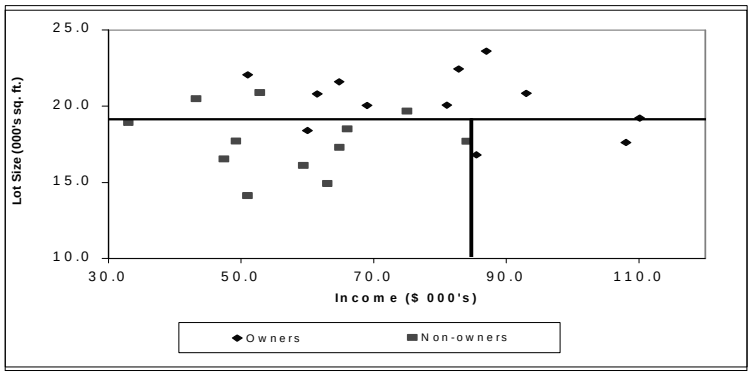
Riding Lawn Mowers Data (Johnson and Wichern)

Observation	Income ($ 000's)	Lot Size (000's sq. ft.)	Owners=1, Non-owners=2
1	60	18.4	1
2	85.5	16.8	1
3	64.8	21.6	1
4	61.5	20.8	1
5	87	23.6	1
6	110.1	19.2	1
7	108	17.6	1
8	82.8	22.4	1
9	69	20	1
10	93	20.8	1
11	51	22	1
12	81	20	1
13	75	19.6	2
14	52.8	20.8	2
15	64.8	17.2	2
16	43.2	20.4	2
17	84	17.6	2
18	49.2	17.6	2
19	59.4	16	2
20	66	18.4	2
21	47.4	16.4	2
22	33	18.8	2
23	51	14	2
24	63	14.8	2

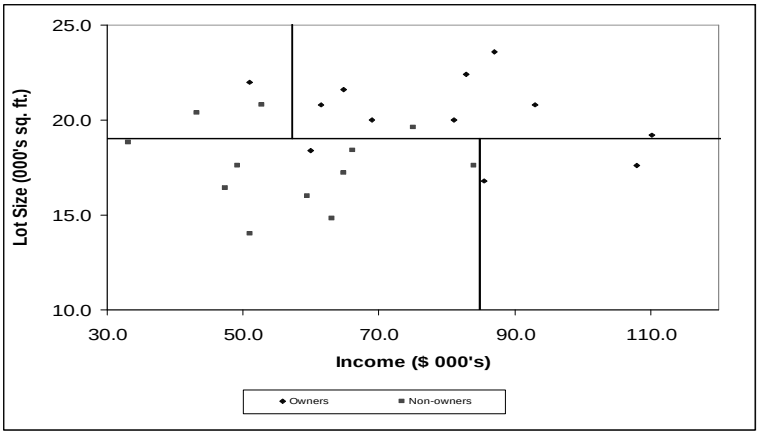
Partition 1



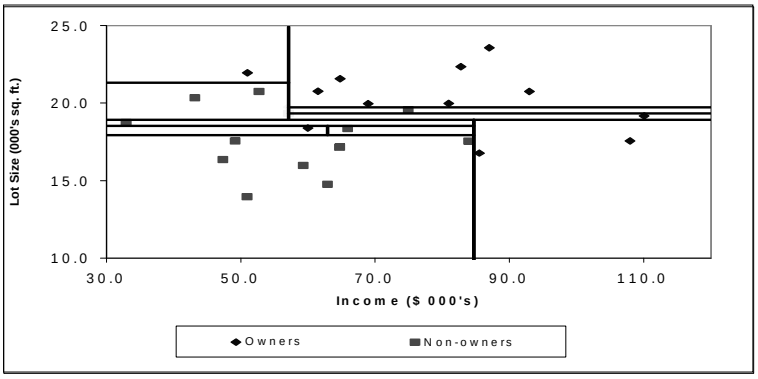
Partition 2



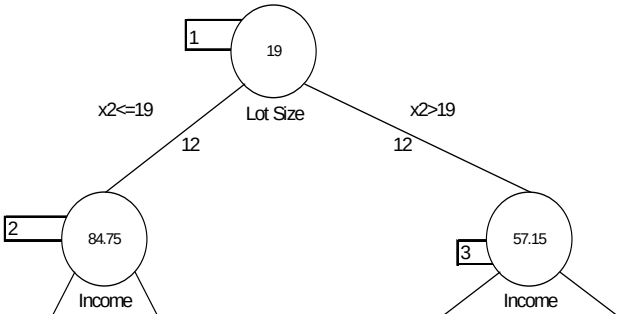
Partition 3



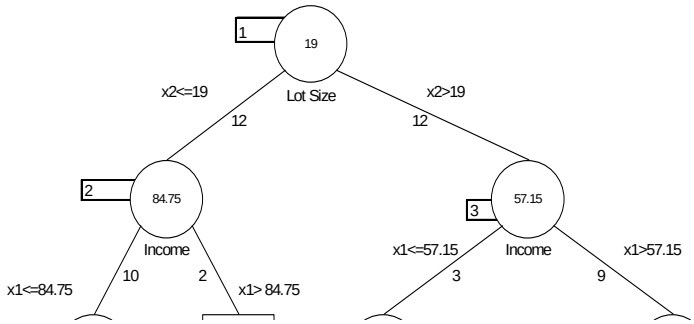
Final Partitioning



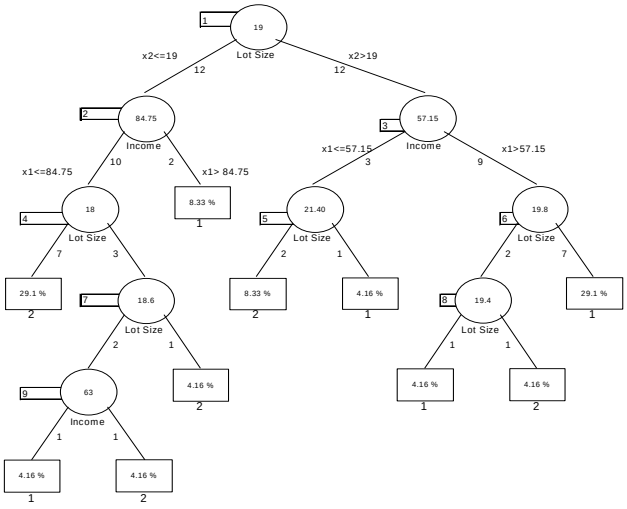
Tree Diagram: First Split



Tree Diagram: First 3 Splits



Tree Diagram: Full Tree



Boston Housing Data

CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT	MEDV	Valclass
0.00632	18	2.31	0	0.538	6.575	65.2	4.09	1	296	15.3	396.9	4.98	24	2
0.02731	0	7.07	0	0.469	6.421	78.9	4.9671	2	242	17.8	396.9	9.14	21.6	2
0.02729	0	7.07	0	0.469	7.185	61.1	4.9671	2	242	17.8	392.83	4.03	34.7	3
0.03237	0	2.18	0	0.458	6.998	45.8	6.0622	3	222	18.7	394.63	2.94	33.4	3
0.06905	0	2.18	0	0.458	7.147	54.2	6.0622	3	222	18.7	396.9	5.33	36.2	3
0.02985	0	2.18	0	0.458	6.43	58.7	6.0622	3	222	18.7	394.12	5.21	28.7	2
0.08829	12.5	7.87	0	0.524	6.012	66.6	5.5605	5	311	15.2	395.6	12.43	22.9	2
0.14455	12.5	7.87	0	0.524	6.172	96.1	5.9505	5	311	15.2	396.9	19.15	27.1	2
0.21124	12.5	7.87	0	0.524	5.631	100	6.0821	5	311	15.2	386.63	29.93	16.5	2
0.17004	12.5	7.87	0	0.524	6.004	85.9	6.5921	5	311	15.2	386.71	17.1	18.9	2
0.22489	12.5	7.87	0	0.524	6.377	94.3	6.3467	5	311	15.2	392.52	20.45	15	2
0.11747	12.5	7.87	0	0.524	6.009	82.9	6.2267	5	311	15.2	396.9	13.27	18.9	2
0.09378	12.5	7.87	0	0.524	5.889	39	5.4509	5	311	15.2	390.5	15.71	21.7	2
0.62976	0	8.14	0	0.538	5.949	61.8	4.7075	4	307	21	396.9	8.26	20.4	2
0.63796	0	8.14	0	0.538	6.096	84.5	4.4619	4	307	21	380.02	10.26	18.2	2
0.62739	0	8.14	0	0.538	5.834	56.5	4.4986	4	307	21	395.62	8.47	19.9	2
1.05393	0	8.14	0	0.538	5.935	29.3	4.4986	4	307	21	386.85	6.58	23.1	2
0.7842	0	8.14	0	0.538	5.99	81.7	4.2579	4	307	21	386.75	14.67	17.5	2
0.80271	0	8.14	0	0.538	5.456	36.6	3.7965	4	307	21	288.99	11.69	20.2	2
0.7258	0	8.14	0	0.538	5.727	69.5	3.7965	4	307	21	390.95	11.28	18.2	2
1.25179	0	8.14	0	0.538	5.57	98.1	3.7979	4	307	21	376.57	21.02	13.6	1

Data Partition

# training rows	304
# validation rows	202

Selected variables

CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT	MEDV	Valclass
0.2498	0	21.89	0	0.624	5.857	98.2	1.6686	4	437	21.2	392.04	21.32	13.3	1
14.2362	0	18.1	0	0.693	6.343	100	1.5741	24	666	20.2	396.9	20.32	7.2	1
0.43571	0	10.59	1	0.489	5.344	100	3.875	4	277	18.6	396.9	23.09	20	2
8.98296	0	18.1	1	0.77	6.212	97.4	2.1222	24	666	20.2	377.73	17.6	17.8	2
41.5292	0	18.1	0	0.693	5.531	85.4	1.6074	24	666	20.2	329.46	27.38	8.5	1
0.05789	12.5	6.07	0	0.409	5.878	21.4	6.498	4	345	18.9	396.21	8.1	22	2
20.7162	0	18.1	0	0.659	4.138	100	1.1781	24	666	20.2	370.22	23.34	11.9	1

XLMiner: Training log

Training Log

Growing the Tree	
#Nodes	Error
0	36.18
1	15.64
2	5.75
3	3.29
4	2.94
5	1.88
6	1.42
7	1.26
8	1.2
9	0.63
10	0.59
11	0.49
12	0.42
13	0.35
14	0.34
15	0.32
16	0.25
17	0.22
18	0.21
19	0.15
20	0.09
21	0.09
22	0.09
23	0.08
24	0.05
25	0.03
26	0.03
27	0.02
28	0.01
29	0
30	0

Training Data Misclassifications

Training Misclassification Summary

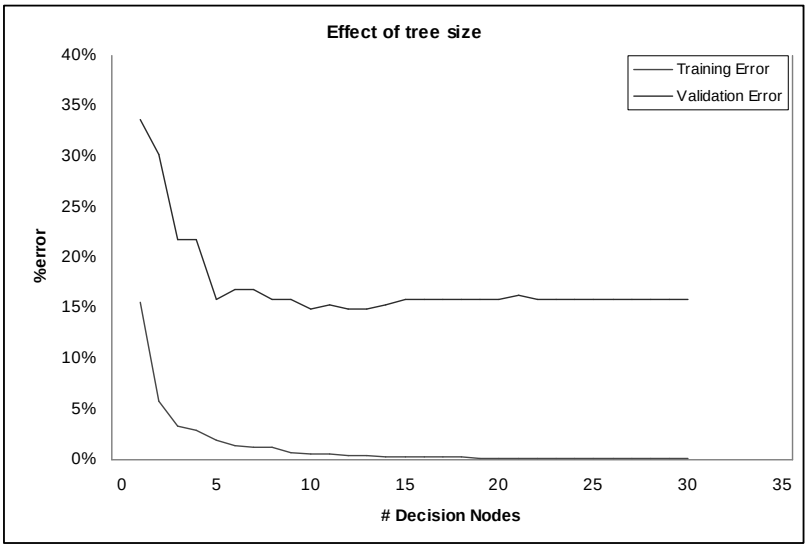
Classification Confusion Matrix			
	Predicted Class		
Actual Class	1	2	3
1	59	0	0
2	0	194	0
3	0	0	51

Error Report			
Class	# Cases	# Errors	% Error
1	59	0	0.00
2	194	0	0.00
3	51	0	0.00
Overall	304	0	0.00

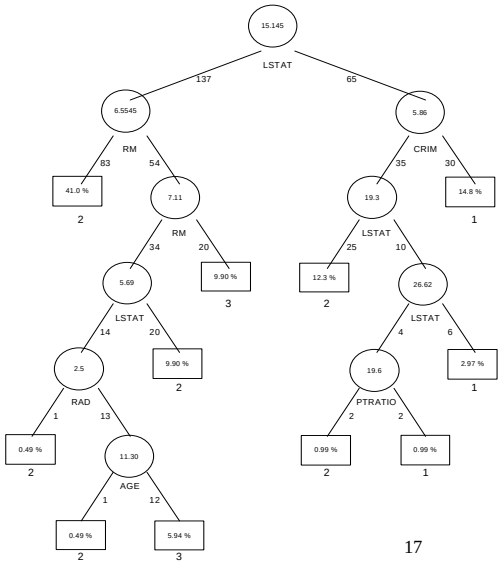
These are cases in the training data

XLMiner: Prune Log

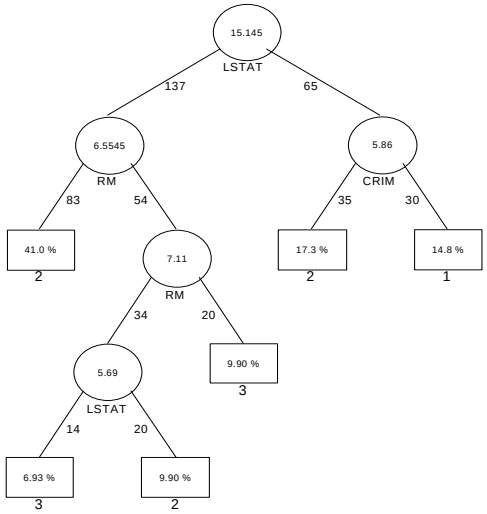
# Decision Nodes	Training Error	Validation Error			
30	0.00%	15.84%			
29	0.00%	15.84%			
28	0.01%	15.84%			
27	0.02%	15.84%			
26	0.03%	15.84%			
25	0.03%	15.84%			
24	0.05%	15.84%			
23	0.08%	15.84%			
22	0.09%	15.84%			
21	0.09%	16.34%			
20	0.09%	15.84%			
19	0.15%	15.84%			
18	0.21%	15.84%			
17	0.22%	15.84%			
16	0.25%	15.84%			
15	0.32%	15.84%			
14	0.34%	15.35%			
13	0.35%	14.85%			
12	0.42%	14.85%			
11	0.49%	15.35%			
10	0.59%	14.85%	<-- Minimum Error Prune	Std. Err.	0.02501957
9	0.63%	15.84%			
8	1.20%	15.84%			
7	1.26%	16.83%			
6	1.42%	16.83%			
5	1.88%	15.84%	<-- Best Prune		
4	2.94%	21.78%			
3	3.29%	21.78%			
2	5.75%	30.20%			
1	15.64%	33.66%			



Minimum Error Tree



Best Pruned Tree



Confusion Table
(Best Pruned Tree)

Validation Misclassification Summary

Classification Confusion Matrix			
	Predicted Class		
Actual Class	1	2	3
1	25	10	0
2	5	120	9
3	0	8	25

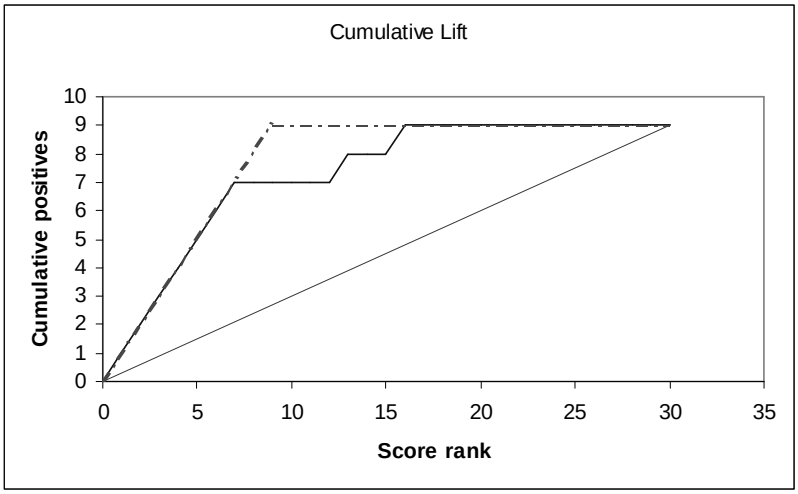
Error Report			
Class	# Cases	# Errors	% Error
1	35	10	28.57
2	134	14	10.45
3	33	8	24.24
Overall	202	32	15.84

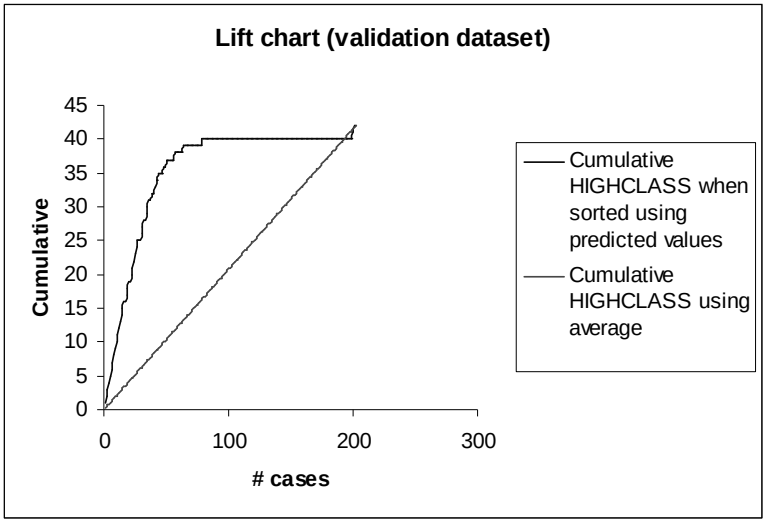
Classifier Performance

	Predicted Log-odds of Success	Predicted Prob. of Success	Actual Value
1	3.599307414	0.97338507	1
2	-6.50733096	0.001490234	0
3	0.406147418	0.600163743	0
4	-14.2910246	6.21565E-07	0
5	4.52734963	0.989306304	1
6	-1.29162444	0.215577985	0
7	-37.6118633	4.62781E-17	0
8	-1.11566321	0.2468166	0
9	-4.32895212	0.013009865	0
10	-24.5364409	2.2078E-11	0
11	-21.6854469	3.82059E-10	0
12	-19.8653678	2.3582E-09	0
13	-13.1040164	2.03703E-06	0
14	4.447239807	0.98842471	1
15	3.529356376	0.971511604	1
16	3.638085074	0.974371436	1
17	-2.68064849	0.064124948	0
18	-0.04018669	0.48995468	0
19	-10.075036	4.21162E-05	0
20	-10.2859002	3.41095E-05	0
21	-14.6084228	4.52525E-07	0
22	8.90164547	0.999863854	1
23	0.087398916	0.521835831	0
24	-6.05898776	0.002331317	1
25	-1.91830242	0.128050988	1
26	-13.2348802	1.78716E-06	0
27	-9.65091205	6.43627E-05	0
28	-13.4562271	1.4323E-06	0
29	-13.9339888	8.8827E-07	0
30	1.725691044	0.848860427	1

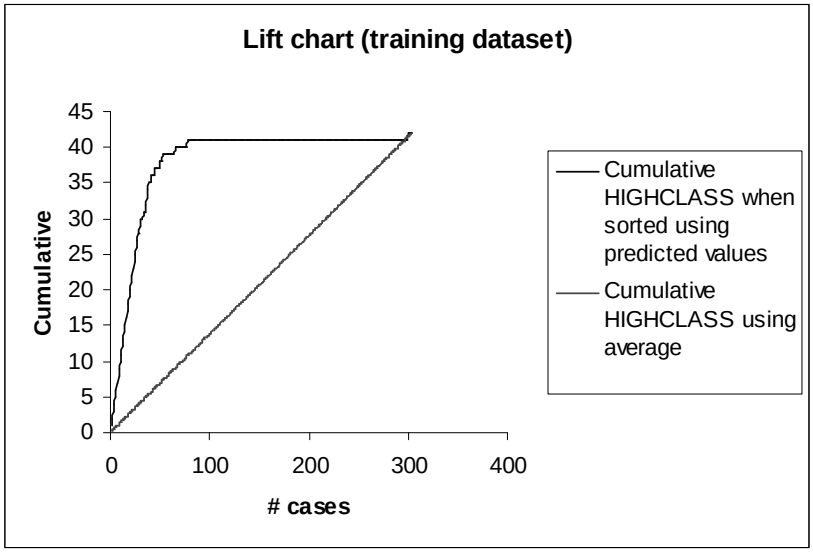
	Case number	Predicted Log-odds of Success	Predicted Prob. of Success	Actual Value of HICLASS
1	22	8.9016	0.9999	1
2	5	4.5273	0.9893	1
3	14	4.4472	0.9884	1
4	16	3.6381	0.9744	1
5	1	3.5993	0.9734	1
6	15	3.5294	0.9715	1
7	30	1.7257	0.8489	1
8	3	0.4061	0.6002	0
9	23	0.0874	0.5218	0
10	18	-0.0402	0.4900	0
11	8	-1.1157	0.2468	0
12	6	-1.2916	0.2156	0
13	25	-1.9183	0.1281	1
14	17	-2.6806	0.0641	0
15	9	-4.3290	0.0130	0
16	24	-6.0590	0.0023	1
17	2	-6.5073	0.0015	0
18	27	-9.6509	0.0001	0
19	19	-10.0750	0.0000	0
20	20	-10.2859	0.0000	0
21	13	-13.1040	0.0000	0
22	26	-13.2349	0.0000	0
23	28	-13.4562	0.0000	0
24	29	-13.9340	0.0000	0
25	4	-14.2910	0.0000	0
26	21	-14.6084	0.0000	0
27	12	-19.8654	0.0000	0
28	11	-21.6854	0.0000	0
29	10	-24.5364	0.0000	0
30	7	-37.6119	0.0000	0

	Probability Rank	Predicted Prob. of Success	Actual Value of HICLASS	cumulative Actual Value
	0			0
1	1	0.9999	1	1
2	2	0.9893	1	2
3	3	0.9884	1	3
4	4	0.9744	1	4
5	5	0.9734	1	5
6	6	0.9715	1	6
7	7	0.8489	1	7
8	8	0.6002	0	7
9	9	0.5218	0	7
10	10	0.4900	0	7
11	11	0.2468	0	7
12	12	0.2156	0	7
13	13	0.1281	1	8
14	14	0.0641	0	8
15	15	0.0130	0	8
16	16	0.0023	1	9
17	17	0.0015	0	9
18	18	0.0001	0	9
19	19	0.0000	0	9
20	20	0.0000	0	9
21	21	0.0000	0	9
22	22	0.0000	0	9
23	23	0.0000	0	9
24	24	0.0000	0	9
25	25	0.0000	0	9
26	26	0.0000	0	9
27	27	0.0000	0	9
28	28	0.0000	0	9
29	29	0.0000	0	9
30	30	0.0000	0	9



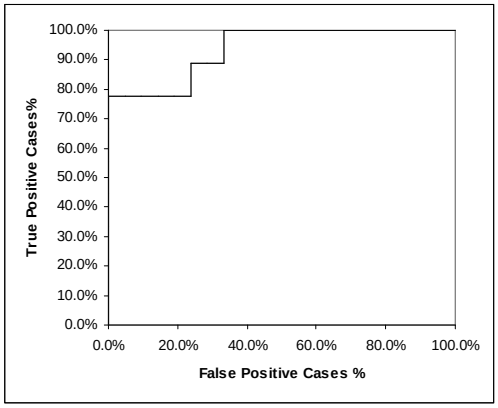


March 2003 23



March 2003 24

ROC curve


100.0%
90.0%
80.0%
70.0%
60.0%
50.0%
40.0%
30.0%
20.0%
10.0%
0.0%
True Positive Cases%
0.0%
20.0%
40.0%
60.0%
80.0%
100.0%
False Positive Cases %