Ctree 准确

testPred no yes

no 1175 111

yes 20 34

rpart

> table(data.test$y,pred)

pred

no yes

no 1165 30

yes 96 49

pruned tree

> table(data.test$y,pred)

pred

no yes

no 1159 36

yes 96 49

> set.seed(1234)

> ind <- sample(2, nrow(data), replace=TRUE, prob=c(0.7, 0.3))

> data.train <- data[ind==1,]

> data.test <- data[ind==2,]

>myFormula <-

y~ age+job+marital+education+default+balance+housing+loan+contact+day+month+duration+campaign+pdays+previous+poutcome

建立公式

> fit <- rpart(myFormula, method="class", data=data.train)

> printcp(fit)

Classification tree:

rpart(formula = myFormula, data = data.train, method = "class")

Variables actually used in tree construction:

[1] day duration education job marital month pdays poutcome

Root node error: 376/3181 = 0.1182

n= 3181

CP nsplit rel error xerror xstd

1 0.046099 0 1.00000 1.00000 0.048427

2 0.029255 3 0.86170 0.98138 0.048034

3 0.018617 4 0.83245 0.90426 0.046345

4 0.013963 5 0.81383 0.90691 0.046405

5 0.013298 9 0.75798 0.88830 0.045983

6 0.010638 11 0.73138 0.88298 0.045861

7 0.010000 13 0.71011 0.89894 0.046225

利用training data 计算cp和split的关系

> plotcp(fit) 画图

> post(fit,file="Dropbox/datamining/assignment/Split\_cp.ps",title="split and cp")

画出decision tree

> summary(fit)

Call:

rpart(formula = myFormula, data = data.train, method = "class")

n= 3181

CP nsplit rel error xerror xstd

1 0.04609929 0 1.0000000 1.0000000 0.04842735

2 0.02925532 3 0.8617021 0.9813830 0.04803427

3 0.01861702 4 0.8324468 0.9042553 0.04634527

4 0.01396277 5 0.8138298 0.9069149 0.04640520

5 0.01329787 9 0.7579787 0.8882979 0.04598300

6 0.01063830 11 0.7313830 0.8829787 0.04586122

7 0.01000000 13 0.7101064 0.8989362 0.04622502

Variable importance

duration poutcome month pdays day job education previous marital balance

48 25 9 7 2 2 2 1 1 1

Node number 1: 3181 observations, complexity param=0.04609929

predicted class=no expected loss=0.1182018 P(node) =1

class counts: 2805 376

probabilities: 0.882 0.118

left son=2 (2927 obs) right son=3 (254 obs)

Primary splits:

duration < 645.5 to the left, improve=78.82595, (0 missing)

poutcome splits as LLRL, improve=53.20516, (0 missing)

month splits as LLRLLLLRLLRR, improve=29.69298, (0 missing)

previous < 1.5 to the left, improve=20.14035, (0 missing)

pdays < 33 to the left, improve=19.73144, (0 missing)

Node number 2: 2927 observations, complexity param=0.04609929

predicted class=no expected loss=0.08541168 P(node) =0.9201509

class counts: 2677 250

probabilities: 0.915 0.085

left son=4 (2844 obs) right son=5 (83 obs)

Primary splits:

poutcome splits as LLRL, improve=56.92623, (0 missing)

month splits as LLRLLLLRLLRR, improve=31.09248, (0 missing)

duration < 220.5 to the left, improve=30.48090, (0 missing)

pdays < 33 to the left, improve=19.39218, (0 missing)

previous < 0.5 to the left, improve=19.02966, (0 missing)

Node number 3: 254 observations, complexity param=0.04609929

predicted class=no expected loss=0.496063 P(node) =0.0798491

class counts: 128 126

probabilities: 0.504 0.496

left son=6 (93 obs) right son=7 (161 obs)

Primary splits:

duration < 766.5 to the left, improve=5.852475, (0 missing)

marital splits as RLR, improve=4.932727, (0 missing)

previous < 1.5 to the left, improve=3.405919, (0 missing)

contact splits as RLL, improve=2.800931, (0 missing)

education splits as RRRL, improve=2.733173, (0 missing)

Surrogate splits:

job splits as RRRLRRRRRRRR, agree=0.646, adj=0.032, (0 split)

month splits as RRRRRRRRRRL-, agree=0.646, adj=0.032, (0 split)

balance < 2.5 to the left, agree=0.642, adj=0.022, (0 split)

age < 76 to the right, agree=0.638, adj=0.011, (0 split)

day < 30.5 to the right, agree=0.638, adj=0.011, (0 split)

Node number 4: 2844 observations, complexity param=0.01396277

predicted class=no expected loss=0.0685654 P(node) =0.8940585

class counts: 2649 195

probabilities: 0.931 0.069

left son=8 (1842 obs) right son=9 (1002 obs)

Primary splits:

duration < 222.5 to the left, improve=20.872530, (0 missing)

month splits as LLLLLLLRLLRR, improve=20.721750, (0 missing)

age < 70.5 to the left, improve= 8.068449, (0 missing)

pdays < 376 to the left, improve= 7.498162, (0 missing)

contact splits as RRL, improve= 6.816178, (0 missing)

Surrogate splits:

pdays < 401 to the left, agree=0.649, adj=0.003, (0 split)

previous < 11 to the left, agree=0.648, adj=0.002, (0 split)

Node number 5: 83 observations, complexity param=0.02925532

predicted class=yes expected loss=0.3373494 P(node) =0.02609242

class counts: 28 55

probabilities: 0.337 0.663

left son=10 (17 obs) right son=11 (66 obs)

Primary splits:

duration < 163 to the left, improve=10.106650, (0 missing)

job splits as RLRLRRLRLLRL, improve= 3.303072, (0 missing)

month splits as LLRLRLRLLRRL, improve= 2.724376, (0 missing)

pdays < 299 to the right, improve= 2.172344, (0 missing)

previous < 1.5 to the left, improve= 1.571327, (0 missing)

Surrogate splits:

contact splits as RRL, agree=0.807, adj=0.059, (0 split)

Node number 6: 93 observations, complexity param=0.01861702

predicted class=no expected loss=0.3548387 P(node) =0.02923609

class counts: 60 33

probabilities: 0.645 0.355

left son=12 (84 obs) right son=13 (9 obs)

Primary splits:

pdays < 132 to the left, improve=5.683820, (0 missing)

poutcome splits as RRLL, improve=4.736528, (0 missing)

duration < 719 to the right, improve=3.775719, (0 missing)

previous < 0.5 to the left, improve=3.442184, (0 missing)

marital splits as RLR, improve=2.707918, (0 missing)

Surrogate splits:

poutcome splits as RRLL, agree=0.968, adj=0.667, (0 split)

previous < 0.5 to the left, agree=0.957, adj=0.556, (0 split)

job splits as LLLLLLLLLLLR, agree=0.914, adj=0.111, (0 split)

month splits as LLRLLLL-LLL-, agree=0.914, adj=0.111, (0 split)

Node number 7: 161 observations, complexity param=0.01329787

predicted class=yes expected loss=0.4223602 P(node) =0.05061301

class counts: 68 93

probabilities: 0.422 0.578

left son=14 (88 obs) right son=15 (73 obs)

Primary splits:

month splits as LRRLLRRRLL--, improve=3.074884, (0 missing)

education splits as RRRL, improve=3.041125, (0 missing)

day < 17.5 to the left, improve=2.339828, (0 missing)

marital splits as RLR, improve=1.854362, (0 missing)

pdays < 302 to the right, improve=1.807372, (0 missing)

Surrogate splits:

campaign < 3.5 to the left, agree=0.671, adj=0.274, (0 split)

housing splits as RL, agree=0.665, adj=0.260, (0 split)

education splits as RLRL, agree=0.658, adj=0.247, (0 split)

job splits as LLRRRRRLLLLR, agree=0.652, adj=0.233, (0 split)

loan splits as LR, agree=0.596, adj=0.110, (0 split)

Node number 8: 1842 observations

predicted class=no expected loss=0.02388708 P(node) =0.5790632

class counts: 1798 44

probabilities: 0.976 0.024

Node number 9: 1002 observations, complexity param=0.01396277

predicted class=no expected loss=0.1506986 P(node) =0.3149953

class counts: 851 151

probabilities: 0.849 0.151

left son=18 (954 obs) right son=19 (48 obs)

Primary splits:

month splits as LLRLLLLRLLRR, improve=17.098820, (0 missing)

pdays < 392.5 to the left, improve=11.342250, (0 missing)

contact splits as RRL, improve=10.494610, (0 missing)

job splits as RLLLRRLLRLLR, improve= 9.097817, (0 missing)

previous < 1.5 to the left, improve= 6.368135, (0 missing)

Node number 10: 17 observations

predicted class=no expected loss=0.1764706 P(node) =0.005344231

class counts: 14 3

probabilities: 0.824 0.176

Node number 11: 66 observations

predicted class=yes expected loss=0.2121212 P(node) =0.02074819

class counts: 14 52

probabilities: 0.212 0.788

Node number 12: 84 observations

predicted class=no expected loss=0.297619 P(node) =0.02640679

class counts: 59 25

probabilities: 0.702 0.298

Node number 13: 9 observations

predicted class=yes expected loss=0.1111111 P(node) =0.002829299

class counts: 1 8

probabilities: 0.111 0.889

Node number 14: 88 observations, complexity param=0.01329787

predicted class=no expected loss=0.4886364 P(node) =0.02766426

class counts: 45 43

probabilities: 0.511 0.489

left son=28 (16 obs) right son=29 (72 obs)

Primary splits:

education splits as LRRL, improve=3.546717, (0 missing)

marital splits as RLR, improve=3.250755, (0 missing)

previous < 3.5 to the left, improve=2.187529, (0 missing)

balance < 68.5 to the right, improve=1.434416, (0 missing)

job splits as RLL-LRLLLRL-, improve=1.111201, (0 missing)

Node number 15: 73 observations

predicted class=yes expected loss=0.3150685 P(node) =0.02294876

class counts: 23 50

probabilities: 0.315 0.685

Node number 18: 954 observations, complexity param=0.01396277

predicted class=no expected loss=0.129979 P(node) =0.2999057

class counts: 830 124

probabilities: 0.870 0.130

left son=36 (945 obs) right son=37 (9 obs)

Primary splits:

pdays < 392.5 to the left, improve=10.465730, (0 missing)

month splits as RR-RLLL-LL--, improve= 7.522166, (0 missing)

contact splits as RRL, improve= 7.086055, (0 missing)

previous < 1.5 to the left, improve= 5.745292, (0 missing)

job splits as RLRRRRRLRLLR, improve= 5.589310, (0 missing)

Node number 19: 48 observations, complexity param=0.01396277

predicted class=yes expected loss=0.4375 P(node) =0.01508959

class counts: 21 27

probabilities: 0.437 0.562

left son=38 (26 obs) right son=39 (22 obs)

Primary splits:

day < 16.5 to the left, improve=5.310315, (0 missing)

balance < 1346 to the right, improve=3.921296, (0 missing)

duration < 324 to the left, improve=1.911713, (0 missing)

housing splits as LR, improve=1.425000, (0 missing)

age < 53.5 to the right, improve=1.196157, (0 missing)

Surrogate splits:

balance < 214 to the right, agree=0.688, adj=0.318, (0 split)

month splits as --L----L--RL, agree=0.667, adj=0.273, (0 split)

duration < 251.5 to the right, agree=0.667, adj=0.273, (0 split)

pdays < 100.5 to the left, agree=0.667, adj=0.273, (0 split)

job splits as RLR-LLL-RLLL, agree=0.646, adj=0.227, (0 split)

Node number 28: 16 observations

predicted class=no expected loss=0.1875 P(node) =0.005029865

class counts: 13 3

probabilities: 0.812 0.188

Node number 29: 72 observations, complexity param=0.0106383

predicted class=yes expected loss=0.4444444 P(node) =0.02263439

class counts: 32 40

probabilities: 0.444 0.556

left son=58 (63 obs) right son=59 (9 obs)

Primary splits:

marital splits as RLL, improve=2.285714, (0 missing)

job splits as RLR-LRLLLRR-, improve=2.005556, (0 missing)

previous < 1.5 to the left, improve=1.791025, (0 missing)

duration < 962 to the right, improve=1.269841, (0 missing)

age < 40.5 to the left, improve=1.078775, (0 missing)

Surrogate splits:

job splits as LLL-LRLLLLL-, agree=0.889, adj=0.111, (0 split)

Node number 36: 945 observations

predicted class=no expected loss=0.1227513 P(node) =0.2970764

class counts: 829 116

probabilities: 0.877 0.123

Node number 37: 9 observations

predicted class=yes expected loss=0.1111111 P(node) =0.002829299

class counts: 1 8

probabilities: 0.111 0.889

Node number 38: 26 observations

predicted class=no expected loss=0.3461538 P(node) =0.00817353

class counts: 17 9

probabilities: 0.654 0.346

Node number 39: 22 observations

predicted class=yes expected loss=0.1818182 P(node) =0.006916064

class counts: 4 18

probabilities: 0.182 0.818

Node number 58: 63 observations, complexity param=0.0106383

predicted class=yes expected loss=0.4920635 P(node) =0.01980509

class counts: 31 32

probabilities: 0.492 0.508

left son=116 (36 obs) right son=117 (27 obs)

Primary splits:

job splits as RLR-L-LLLRR-, improve=2.380952, (0 missing)

previous < 1.5 to the left, improve=2.027913, (0 missing)

duration < 913 to the right, improve=1.115123, (0 missing)

day < 23.5 to the left, improve=1.113602, (0 missing)

age < 31.5 to the left, improve=1.027913, (0 missing)

Surrogate splits:

day < 27.5 to the left, agree=0.667, adj=0.222, (0 split)

duration < 894 to the right, agree=0.619, adj=0.111, (0 split)

campaign < 1.5 to the right, agree=0.619, adj=0.111, (0 split)

age < 26.5 to the right, agree=0.603, adj=0.074, (0 split)

month splits as L--RR---LL--, agree=0.603, adj=0.074, (0 split)

Node number 59: 9 observations

predicted class=yes expected loss=0.1111111 P(node) =0.002829299

class counts: 1 8

probabilities: 0.111 0.889

Node number 116: 36 observations

predicted class=no expected loss=0.3888889 P(node) =0.0113172

class counts: 22 14

probabilities: 0.611 0.389

Node number 117: 27 observations

predicted class=yes expected loss=0.3333333 P(node) =0.008487897

class counts: 9 18

probabilities: 0.333 0.667

> print(data\_rpart)

n= 3181

node), split, n, loss, yval, (yprob)

\* denotes terminal node

1) root 3181 376 no (0.88179818 0.11820182)

2) duration< 645.5 2927 250 no (0.91458832 0.08541168)

4) poutcome=failure,other,unknown 2844 195 no (0.93143460 0.06856540)

8) duration< 222.5 1842 44 no (0.97611292 0.02388708) \*

9) duration>=222.5 1002 151 no (0.84930140 0.15069860)

18) month=apr,aug,feb,jan,jul,jun,may,nov 954 124 no (0.87002096 0.12997904)

36) pdays< 392.5 945 116 no (0.87724868 0.12275132) \*

37) pdays>=392.5 9 1 yes (0.11111111 0.88888889) \*

19) month=dec,mar,oct,sep 48 21 yes (0.43750000 0.56250000)

38) day< 16.5 26 9 no (0.65384615 0.34615385) \*

39) day>=16.5 22 4 yes (0.18181818 0.81818182) \*

5) poutcome=success 83 28 yes (0.33734940 0.66265060)

10) duration< 163 17 3 no (0.82352941 0.17647059) \*

11) duration>=163 66 14 yes (0.21212121 0.78787879) \*

3) duration>=645.5 254 126 no (0.50393701 0.49606299)

6) duration< 766.5 93 33 no (0.64516129 0.35483871)

12) pdays< 132 84 25 no (0.70238095 0.29761905) \*

13) pdays>=132 9 1 yes (0.11111111 0.88888889) \*

7) duration>=766.5 161 68 yes (0.42236025 0.57763975)

14) month=apr,feb,jan,may,nov 88 43 no (0.51136364 0.48863636)

28) education=primary,unknown 16 3 no (0.81250000 0.18750000) \*

29) education=secondary,tertiary 72 32 yes (0.44444444 0.55555556)

58) marital=married,single 63 31 yes (0.49206349 0.50793651)

116) job=blue-collar,management,self-employed,services,student 36 14 no (0.61111111 0.38888889)

232) previous< 1.5 30 9 no (0.70000000 0.30000000) \*

233) previous>=1.5 6 1 yes (0.16666667 0.83333333) \*

117) job=admin.,entrepreneur,technician,unemployed 27 9 yes (0.33333333 0.66666667) \*

59) marital=divorced 9 1 yes (0.11111111 0.88888889) \*

15) month=aug,dec,jul,jun,mar 73 23 yes (0.31506849 0.68493151) \*

> plot(data\_rpart)

> text(data\_rpart, use.n=T)

> post(fit,file="Dropbox/datamining/assignment/tree1.ps",title="Decision Tree")

画出decision tree， 截止条件 cp

pred=predict(data\_rpart, data.test, type="class")

> table(data.test$y,pred)

> opt <- which.min(data\_rpart$cptable[,"xerror"])

> cp <- data\_rpart$cptable[opt, "CP"]

> data\_prune <- prune(data\_rpart, cp = cp)

> print(data\_prune)

n= 3181

node), split, n, loss, yval, (yprob)

\* denotes terminal node

1) root 3181 376 no (0.88179818 0.11820182)

2) duration< 645.5 2927 250 no (0.91458832 0.08541168)

4) poutcome=failure,other,unknown 2844 195 no (0.93143460 0.06856540)

8) duration< 222.5 1842 44 no (0.97611292 0.02388708) \*

9) duration>=222.5 1002 151 no (0.84930140 0.15069860)

18) month=apr,aug,feb,jan,jul,jun,may,nov 954 124 no (0.87002096 0.12997904)

36) pdays< 392.5 945 116 no (0.87724868 0.12275132) \*

37) pdays>=392.5 9 1 yes (0.11111111 0.88888889) \*

19) month=dec,mar,oct,sep 48 21 yes (0.43750000 0.56250000)

38) day< 16.5 26 9 no (0.65384615 0.34615385) \*

39) day>=16.5 22 4 yes (0.18181818 0.81818182) \*

5) poutcome=success 83 28 yes (0.33734940 0.66265060)

10) duration< 163 17 3 no (0.82352941 0.17647059) \*

11) duration>=163 66 14 yes (0.21212121 0.78787879) \*

3) duration>=645.5 254 126 no (0.50393701 0.49606299)

6) duration< 766.5 93 33 no (0.64516129 0.35483871)

12) pdays< 132 84 25 no (0.70238095 0.29761905) \*

13) pdays>=132 9 1 yes (0.11111111 0.88888889) \*

7) duration>=766.5 161 68 yes (0.42236025 0.57763975) \*

Prune 树

> plot(data\_prune)

> text(data\_prune, use.n=T)

> post(data\_prune,file="Dropbox/datamining/assignment/ptree.ps",title="Pruned Decision Tree")

根据prune过的树来predict 与data.test作比较

> y\_pred <- predict(data\_prune, newdata=data.test)

> plot(y\_pred ~ y, data=data.test, xlab="Observed", ylab="Predicted")