Surrogates

Rscript: mtcars CART.R

CART

Based on Chew C. H. (2020) textbook: AI, Analytics and Data Science. Vol 1., Chap 8.

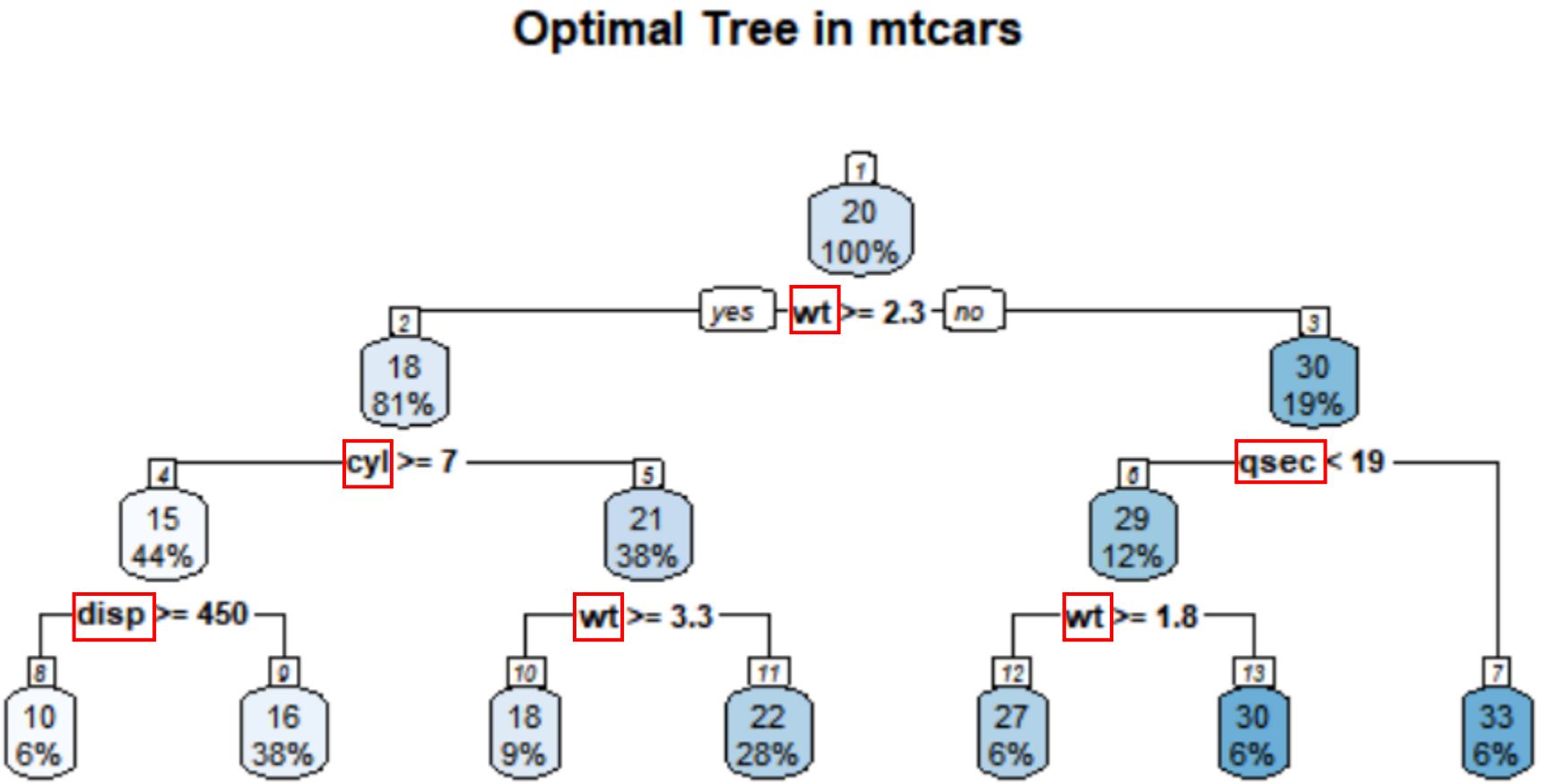
Base R dataset: mtcars

32 cases, 11 columns.

Continuous Outcome variable Y: mpg

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21.0	6	160.0	110	3.90	2.620	16.46	0	1	4	4
Mazda RX4 Wag	21.0	6	160.0	110	3.90	2.875	17.02	0	1	4	4
Datsun 710	22.8	4	108.0	93	3.85	2.320	18.61	1	1	4	1
Hornet 4 Drive	21.4	6	258.0	110	3.08	3.215	19.44	1	0	3	1
Hornet Sportabout	18.7	8	360.0	175	3.15	3.440	17.02	0	0	3	2
Valiant	18.1	6	225.0	105	2.76	3.460	20.22	1	0	3	1

CART requires the splitting variable value to be available



Displays surrogates at each internal node via summary(). Example: At node 1 i.e. root node

```
70 # Surrogates shown in summary()
71 summary(cart2)
```

Min CP required to prune at Node 1

```
Node number 1: 32 observations,    complexity param=0.6526612
  mean=20.09062, MSE=35.18897
  left son=2 (26 obs) right son=3 (6 obs)
  Primary splits:
      wt   < 2.26   to the right, improve=0.6526612, (0 missing)
      cyl  < 5      to the right, improve=0.6431252, (0 missing)
      disp < 163.8  to the right, improve=0.6130502, (0 missing)
      hp   < 118    to the right, improve=0.6010712, (0 missing)
      vs   < 0.5    to the left,  improve=0.4409477, (0 missing)
  Surrogate splits:
      disp < 101.55 to the right, agree=0.969, adj=0.833, (0 split)
      hp   < 92     to the right, agree=0.938, adj=0.667, (0 split)
      drat < 4      to the left,  agree=0.906, adj=0.500, (0 split)
      cyl  < 5      to the right, agree=0.844, adj=0.167, (0 split)
```

This surrogate split is 96.9% similar to the best Primary split.

This surrogate split is 83.3% better than using majority rule to send NAs to left or right child.

Surrogates are only activated when there are missing values

- The dataset mtcars has no missing values.

1. Copy the data to be mtcars2.
2. Delete the wt value in first row of mtcars2.
3. Delete wt and disp values in second row of mtcars2.
4. Repeat the CART model on mtcars2.
5. View surrogates activated in node 1.

```
70 # Surrogates shown in summary() ----------------------------
71 summary(cart2)
72
73 # Create missing values in first two rows
74 mtcars2 <- mtcars
75 mtcars2[1,6] <- NA   # first row, 6th col. wt is the 6th col.
76 mtcars2[2,6] <- NA
77 mtcars2[2,3] <- NA   # 3rd column is disp.
```

Check the NAs in rows 1 and 2

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21.0	6	160.0	110	3.90	*NA*	16.46	0	1	4	4
Mazda RX4 Wag	21.0	6	*NA*	110	3.90	*NA*	17.02	0	1	4	4
Datsun 710	22.8	4	108.0	93	3.85	2.320	18.61	1	1	4	1
Hornet 4 Drive	21.4	6	258.0	110	3.08	3.215	19.44	1	0	3	1

- Case 1 has one missing value in wt.
- Case 2 has two missing values:
  - wt
  - disp

```
79  cart3 <- rpart(mpg ~ ., data = mtcars2, method = 'anova', control = rpart.control(minsplit = 2, cp = 0))
80
81  summary(cart3)
```

```
Node number 1: 32 observations,    complexity param=0.6526612
  mean=20.09062, MSE=35.18897
  left son=2 (26 obs) right son=3 (6 obs)
  Primary splits:
      wt   < 2.26   to the right, improve=0.6709400, (2 missing)
      cyl  < 5      to the right, improve=0.6431252, (0 missing)
      disp < 153.35 to the right, improve=0.6305513, (1 missing)
      hp   < 118    to the right, improve=0.6010712, (0 missing)
      vs   < 0.5    to the left,  improve=0.4409477, (0 missing)
  Surrogate splits:
      disp < 101.55 to the right, agree=0.967, adj=0.833, (1 split)
      hp   < 92     to the right, agree=0.933, adj=0.667, (1 split)
      drat < 4      to the left,  agree=0.900, adj=0.500, (0 split)
      cyl  < 5      to the right, agree=0.833, adj=0.167, (0 split)
      am   < 0.5    to the left,  agree=0.833, adj=0.167, (0 split)
```

- Case 1 used disp as surrogate.
- Case 2 used hp as surrogate.

Enhanced version of CART: Random Forest

- Ensemble of 500 CARTs.
- More stable than 1 CART model.
- More accurate than 1 CART model.
- Will not overfit.
- Taught in BC2407 Analytics II.

Summary: Categorical Y vs Continuous Y

	Categorical Y	Continuous Y
rpart() parameter	method = ‘class’	method = ‘anova’
Model Prediction	Majority Category	Mean value of Y
Metric to determine Best Split	Gini index	SSE
Metric to evaluate model performance	Misclassification Error	MSE

Summary of CART

- Phrase 1: Grow Tree to max.
- Phrase 2: Prune Tree to min.
- Optimal Tree selection via 10 fold CV with 1 SE rule.
- Extract Decision Rules (i.e. Predictions) at terminal nodes.
- Variable Importance.
- CART handles missing values automatically via Surrogates.