

A Brief History of Classification and Regression Trees

Wei-Yin Loh

Department of Statistics

University of Wisconsin–Madison

www.stat.wisc.edu/~loh/

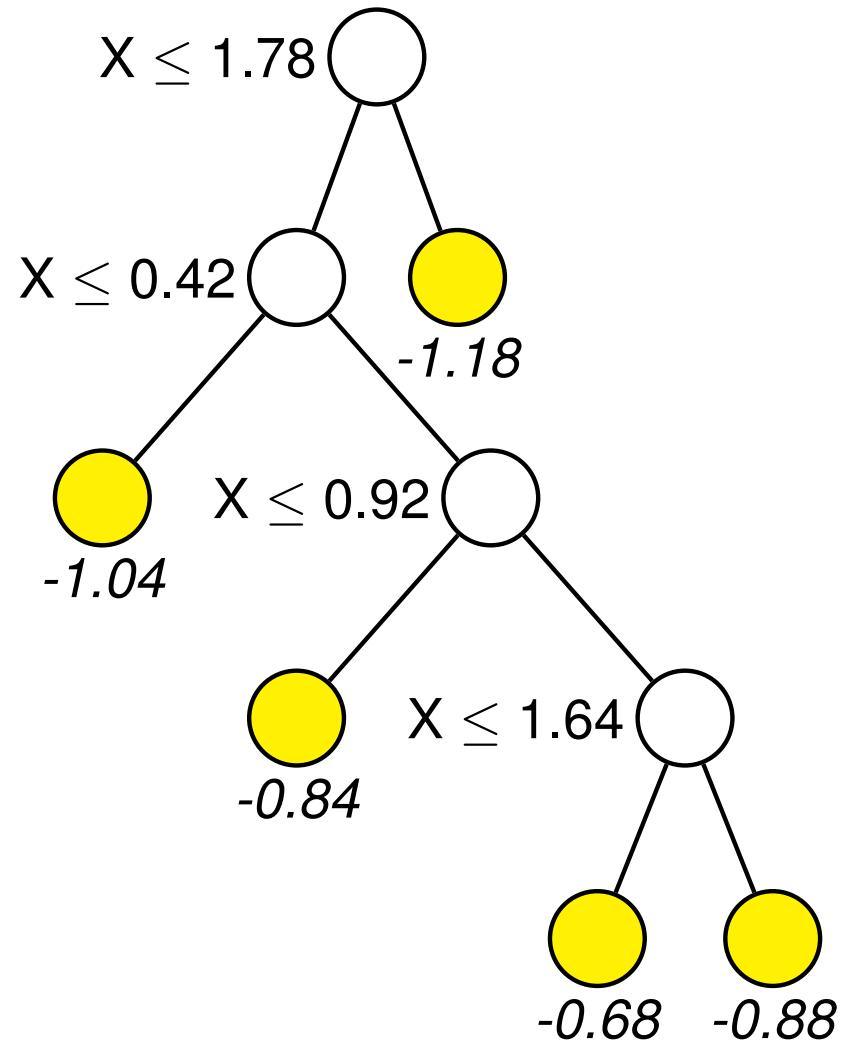
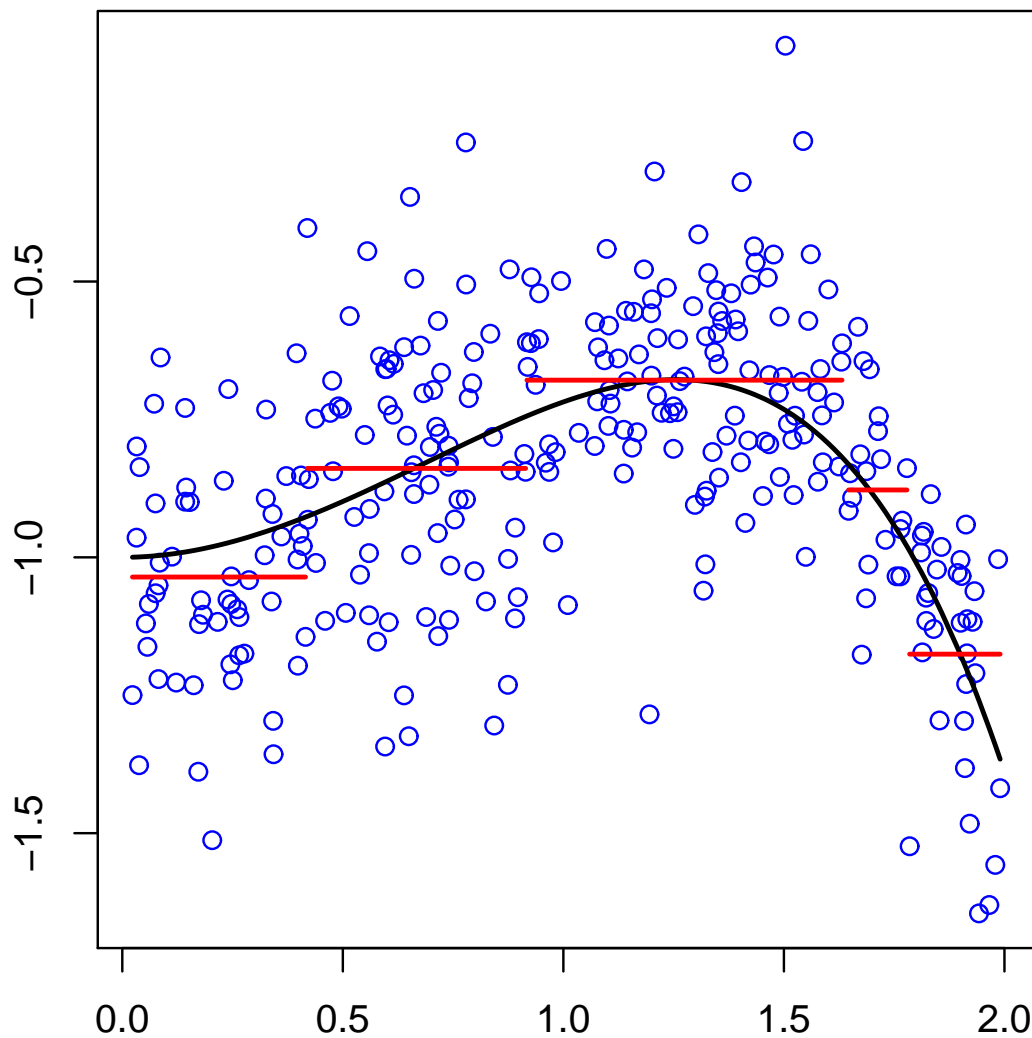
Main paradigms

- 1st gen.** AID (Morgan and Sonquist, 1963), THAID (Messenger and Mandell, 1972), CHAID (Kass, 1980)
- 2nd gen.** **CART** (Breiman et al., 1984) , RECPAM (Ciampi et al., 1988), Segal (1988, 1992), LeBlanc and Crowley (1992), Alexander and Grimshaw (1996), Zhang (1998), MVPART (De'ath, 2002), Su et al. (2004); **ID3** (Quinlan, 1986), M5 (Quinlan, 1992), C4.5 (Quinlan, 1993); **FACT** (Loh and Vanichsetakul, 1988)
- 3rd gen.** QUEST (Loh and Shih, 1997), CRUISE (Kim and Loh, 2001, 2003), Bayesian CART (Chipman et al., 1998; Denison et al., 1998)
- 4th gen.** GUIDE (Loh, 2002, 2009; Loh and Zheng, 2013; Loh et al., 2015), CTREE (Hothorn et al., 2006), MOB (Zeileis et al., 2008); Random forest (Breiman, 2001), TARGET (Fan and Gray, 2005; Gray and Fan, 2008), BART (Chipman et al., 2010)

Automatic Interaction Detection (AID) (Morgan and Sonquist, 1963)

- The first regression tree algorithm
- Fit a piecewise-constant model by recursively splitting data into two subsets (nodes), with splits of form “ $X \leq c$ ” or “ $X \in A$ ”
- For each node t , define node **impurity** $\phi(t) = \sum_{i \in t} (y_i - \bar{y})^2$
- Choose split to minimize sum of impurities (greedy search)
- Stop splitting if reduction in node impurity not high enough

Example of piecewise-constant regression tree



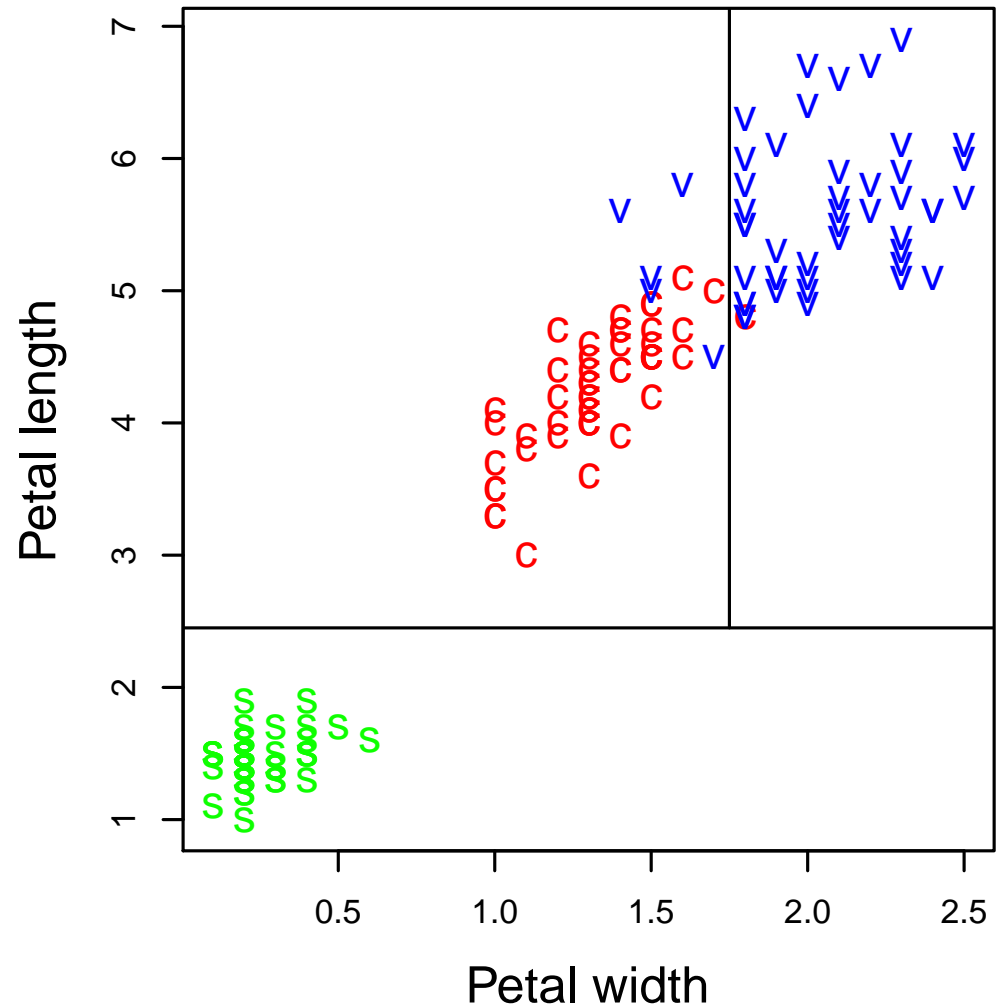
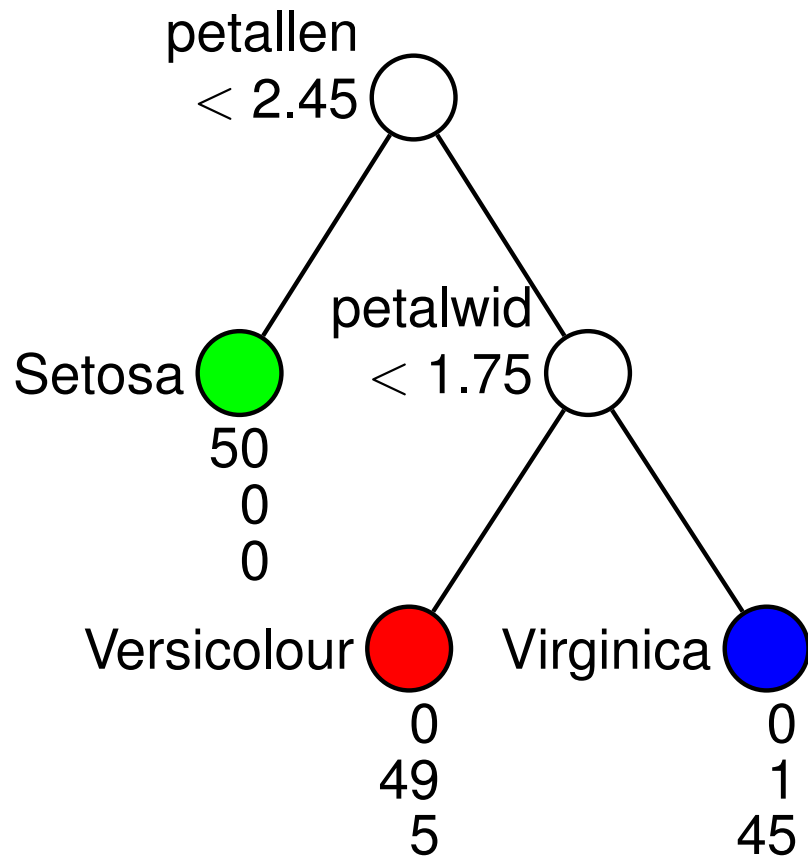
THAID (Messenger and Mandell, 1972)

- The first classification tree algorithm
- Split data to maximize sum of #cases in modal category
- Predicted class is the mode

Example: Fisher's iris data

- 3 classes (Setosa, Versicolour, Virginica)
- 50 observations per class (150 total)
- 4 predictor variables:
 1. petal length
 2. petal width
 3. sepal length
 4. sepal width

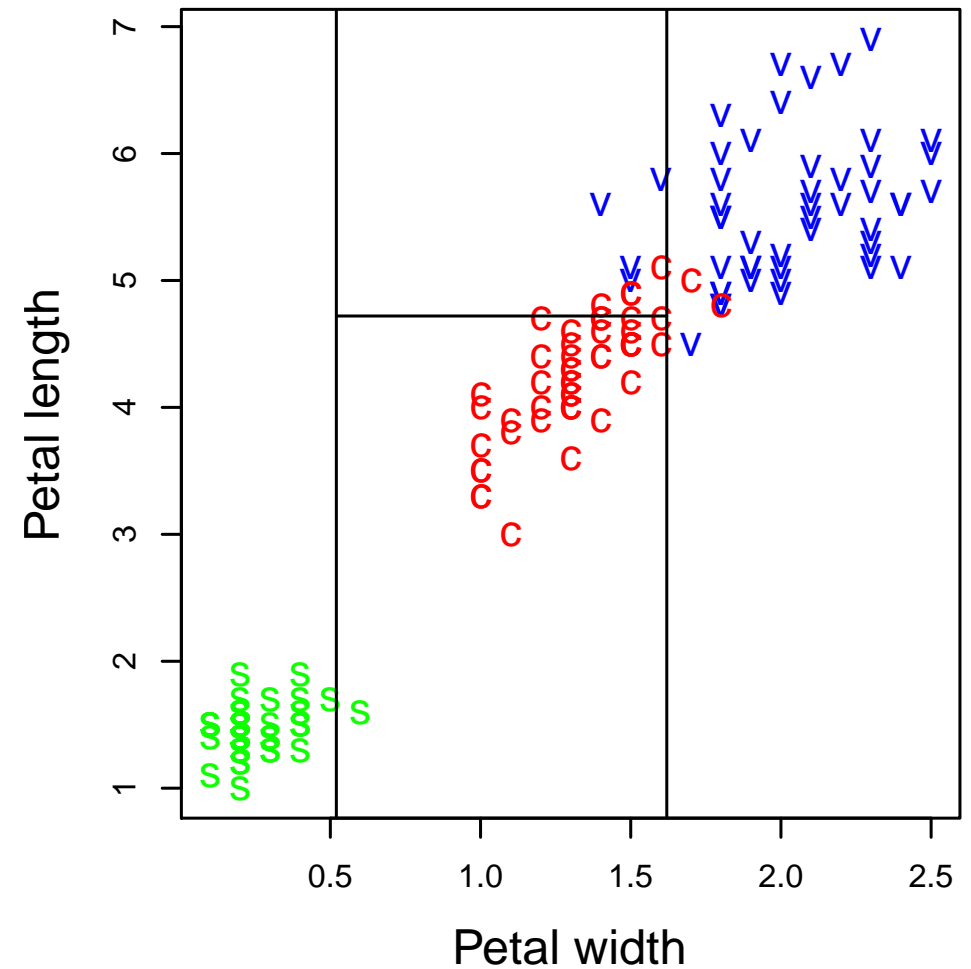
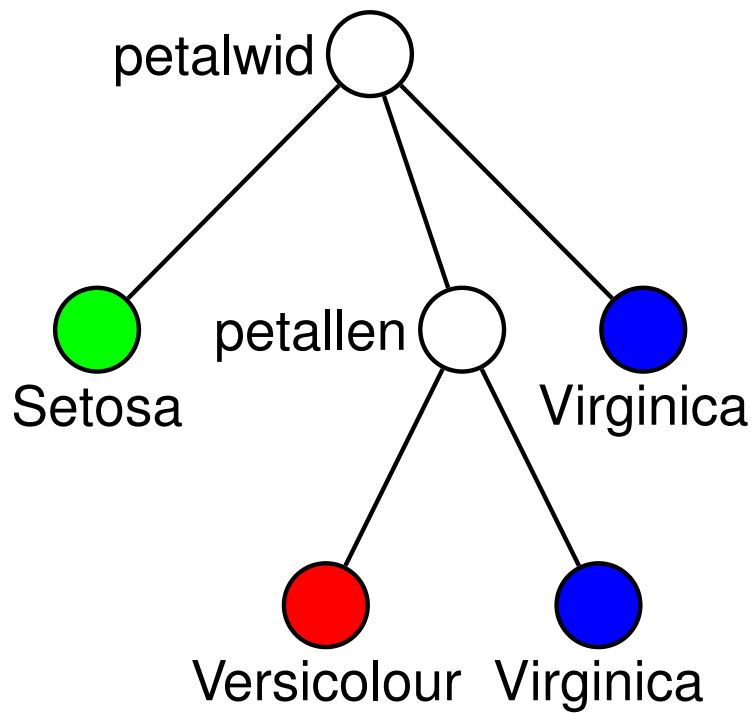
RPART classification tree for iris data



CHAID (Kass, 1980)

- Split each categorical X into one subnode for each category
- Split each ordered X into 10 equal-sized intervals
- Test pairs of subnodes for significant differences; merge those not significant
- Use Bonferroni corrections to control multiplicity of tests
- Choose most significant split
- **Strength:** computational speed
- **Weaknesses:** inaccurate split points and over-conservative Bonferroni corrections

CHAID tree for iris data



CART (Breiman et al., 1984)

Uses greedy search of AID and THAID with these additions:

1. Trees pruned instead of stopping rules
2. Trees selected by cross-validation
3. Unequal class priors and misclassification costs allowed
4. Missing values handled by surrogate splits
5. Variable importance scores used to detect masking
6. Linear splits $\sum_i a_i x_i \leq c$ obtained by random search

RPART (Therneau et al., 2014) is an R implementation of CART

CART extensions to censored data

Node impurity measures

Gordon and Olshen (1985)	Minimum Wasserstein distance between Kaplan-Meier curve and a point mass
Ciampi et al. (1988)	Proportional hazards likelihood ratio
Segal (1988)	Logrank test statistic
Davis and Anderson (1989)	Exponential loglikelihood
LeBlanc and Crowley (1992)	Proportional hazards loglikelihood

CART extensions in other directions

Ciampi (1991)	Generalized linear models
Segal (1992)	Longitudinal data
Alexander and Grimshaw (1996)	Simple linear regression
Zhang (1998)	Multiple binary response variables
Yu and Lambert (1999)	Longitudinal data
De'ath (2002), Lee (2005)	Multivariate responses
Sela and Simonoff (2012)	Random effects tree

Selection biases and computational problems of CART

- An ordered X variable with n unique values yields $(n - 1)$ splits of the form $X \leq c$
- A categorical variable with m unique values yields $(2^{m-1} - 1)$ splits of the form $X \in A$

As a result, CART is **biased** toward selecting

1. split variables that allow **more** splits (Loh and Shih, 1997)
2. split variables with **more** missing values
3. surrogate variables with **fewer** missing values (Kim & Loh, 2001)

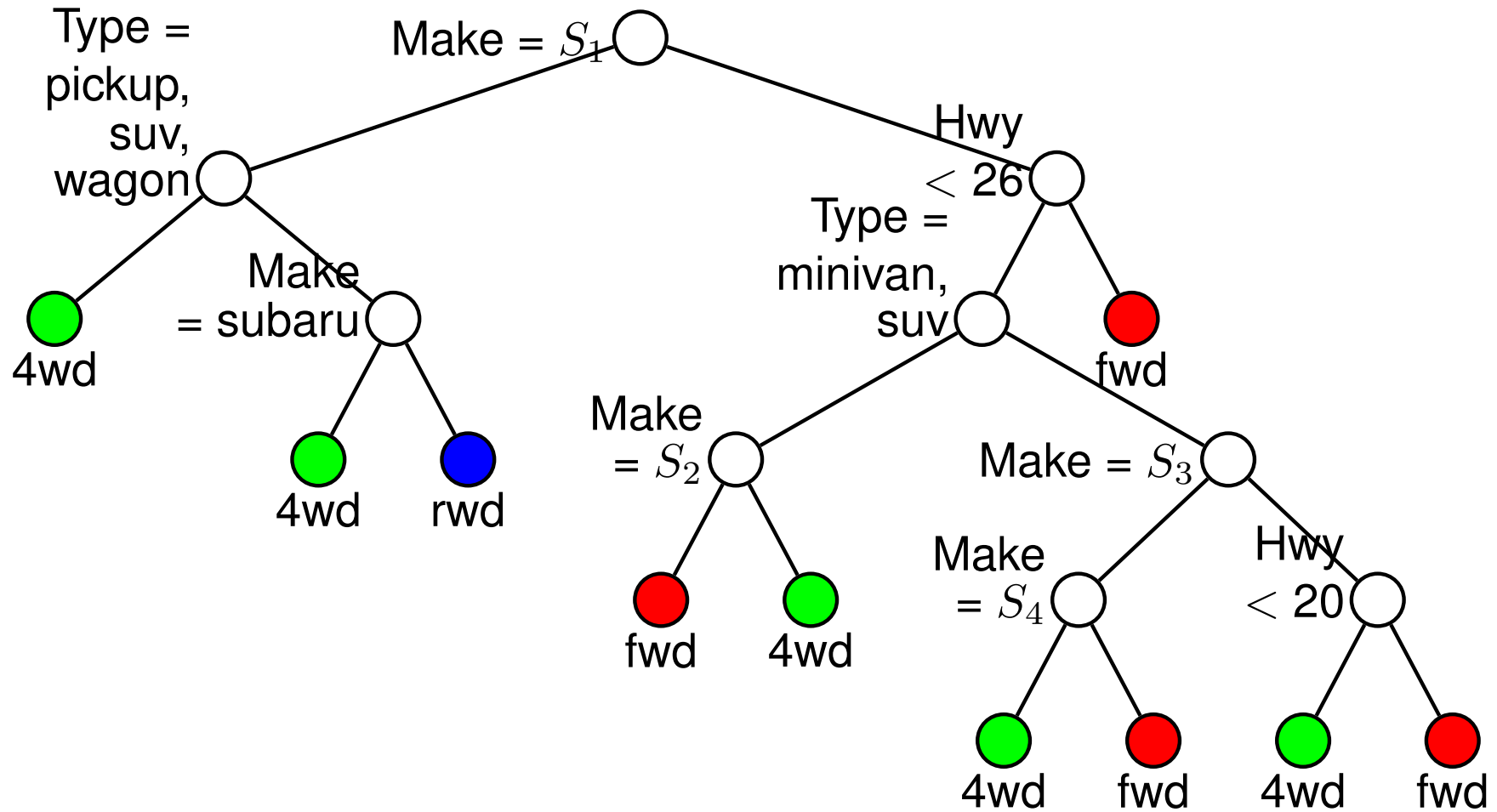
Example: Predicting drive train for model year 2004 cars

- 428 cars and 13 variables (2 categorical, 11 ordered)
- Drive train takes three values:
 - 94 (22%) four-wheel (4wd)
 - 224 (52%) front-wheel (fwd)
 - 110 (25%) rear-wheel (rwd)

Predictor variables

Variable	Description	Variable	Description
Make	Make of car (38 values)	City	City miles/gallon
Type	Type of car (6 values)	Hwy	Highway miles/gallon
Rprice	Suggested retail price	Weight	Weight (pounds)
Dcost	Dealer cost	Whlbase	Wheel base (in.)
Enginsz	Engine size (liters)	Length	Length (in.)
Cylin	Number of cylinders	Width	Width (in.)
Hp	Horsepower		

RPART tree for car data (took 22 cpu hrs!)

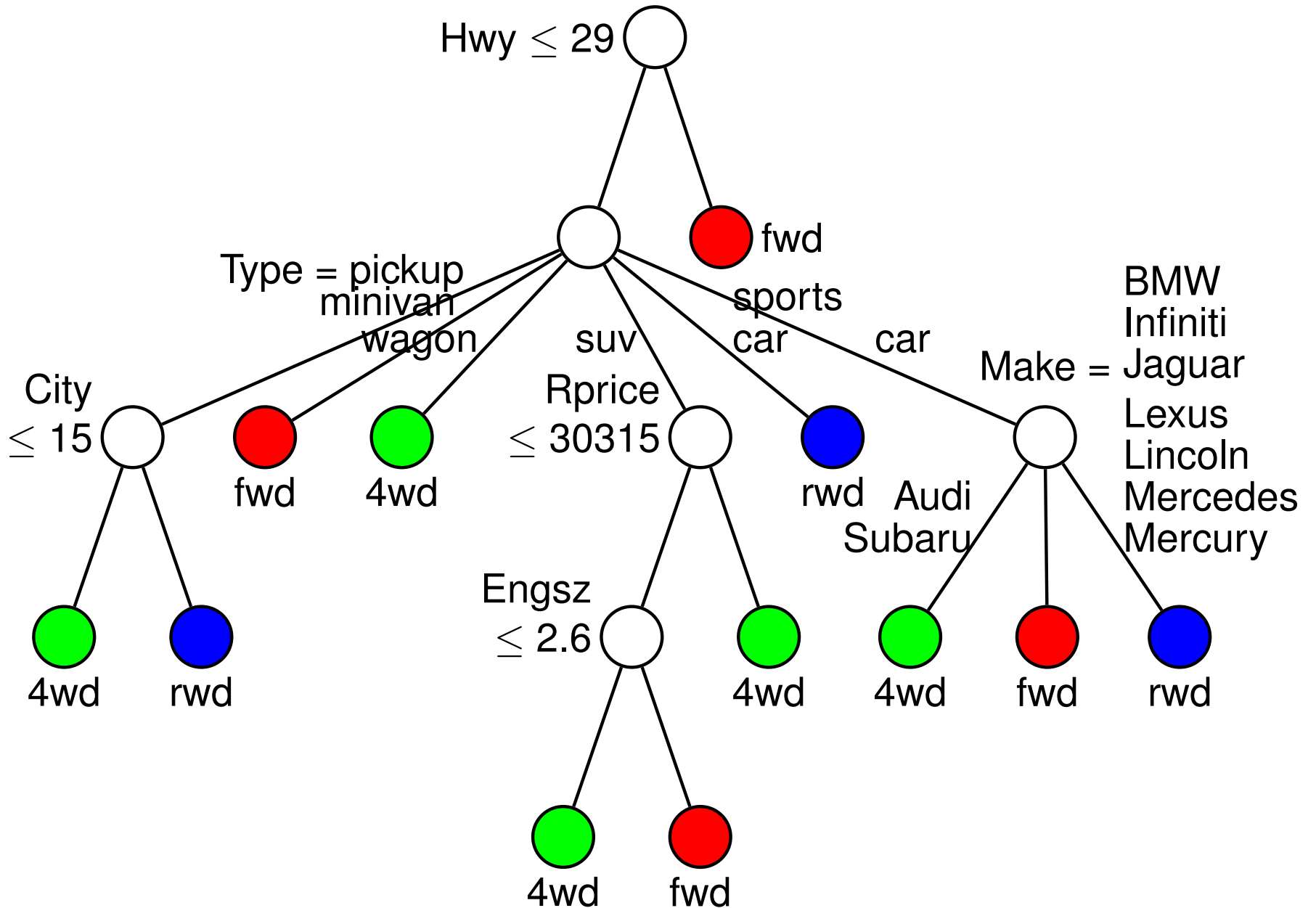


$S_1 = \{\text{BMW, GMC, Hummer, Infiniti, Jaguar, Land-Rover, Lexus, Lincoln, Mazda, Mercedes, Porsche, Subaru}\}$; $S_2 = \{\text{Cadillac, Chrysler, Kia, Mercury, Nissan}\}$;
 $S_3 = \{\text{Audi, Kia, Mitsubishi, Nissan, Pontiac, Volkswagen, Volvo}\}$;
 $S_4 = \{\text{Audi, Nissan, Volvo}\}$

ID3 (Quinlan, 1986) and C4.5 (Quinlan, 1993)

- Entropy-based impurity criterion called gain ratio
- Exhaustive search for splits ($X < c$) if split on ordered X
- One branch for each value ($X = a$) if split on categorical X
- Branches merged if error rate is reduced
- Pruning with error estimates based on binomial confidence bounds
- Case weights to deal with missing values
- Class priors and misclassification costs cannot be specified

C4.5 tree for car data

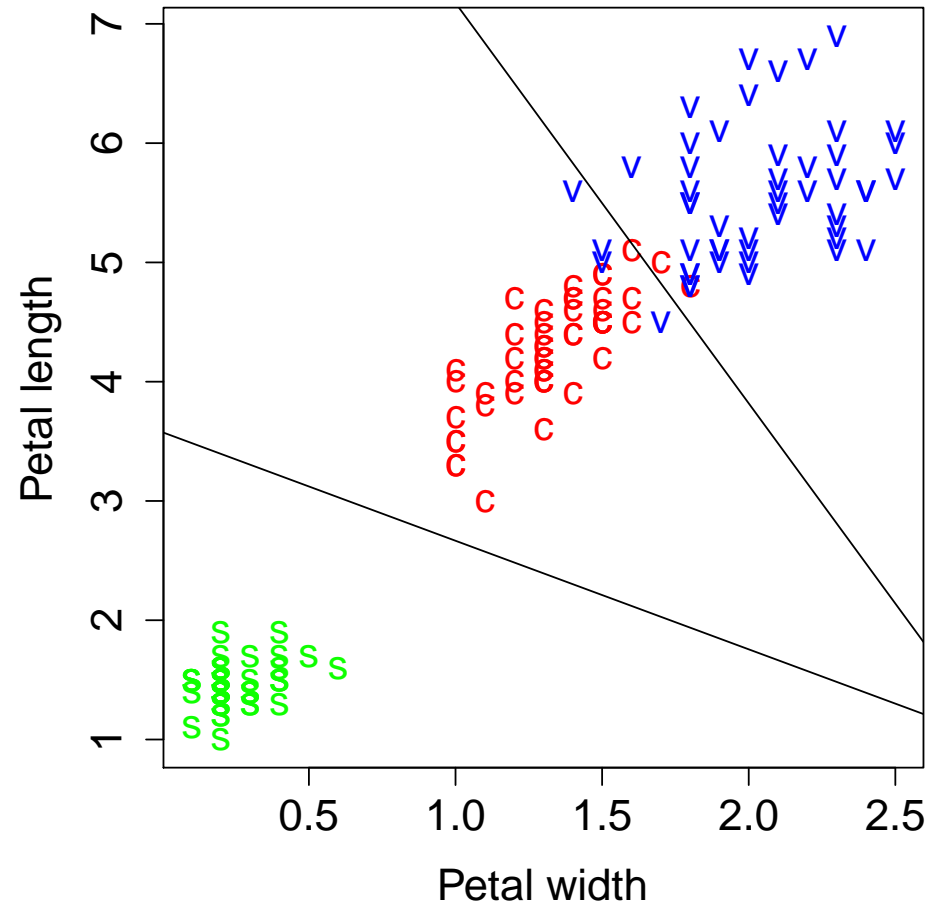
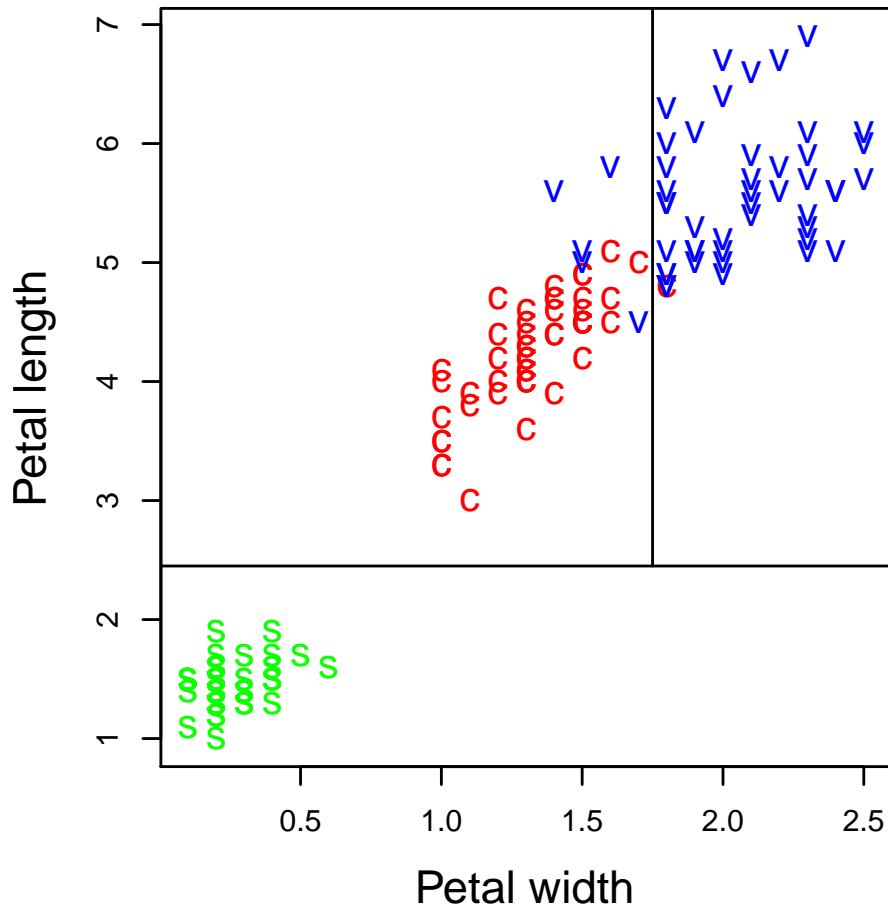


Selection biases of C4.5

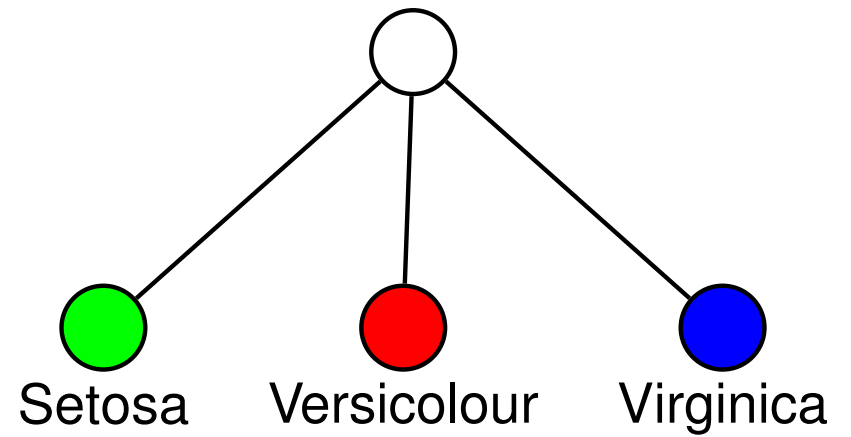
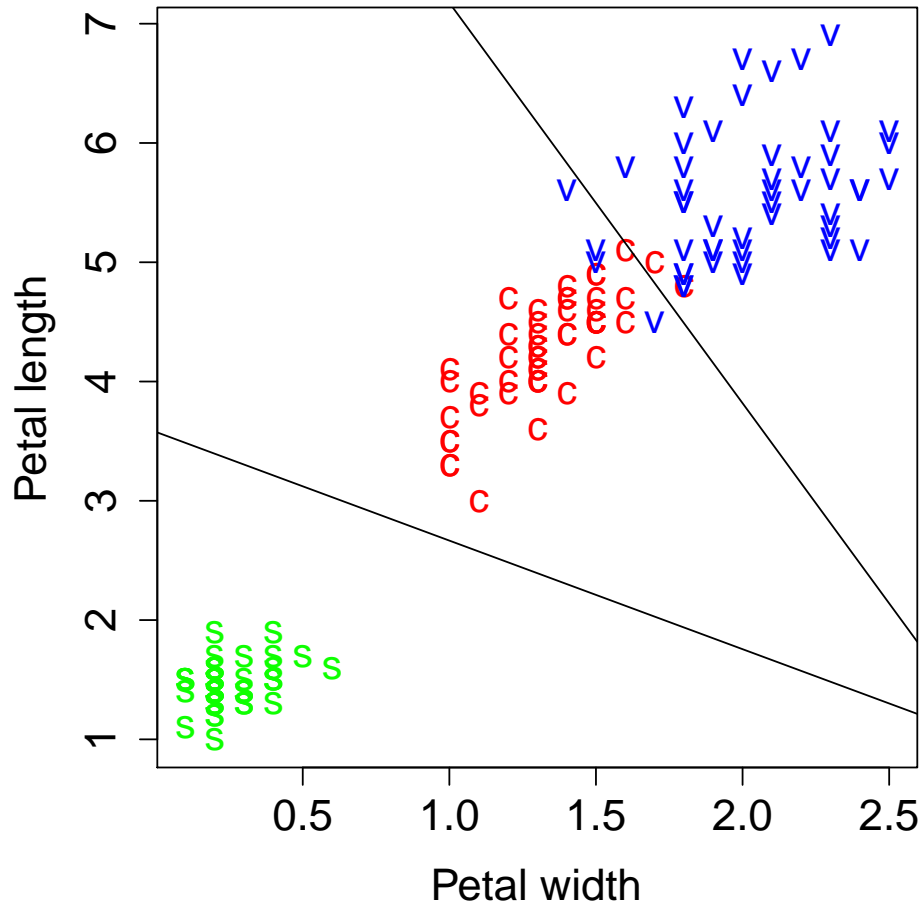
Biased toward selecting

1. categorical variables with **more** values
2. ordered variables with **fewer** values
3. variables with **more** missing values

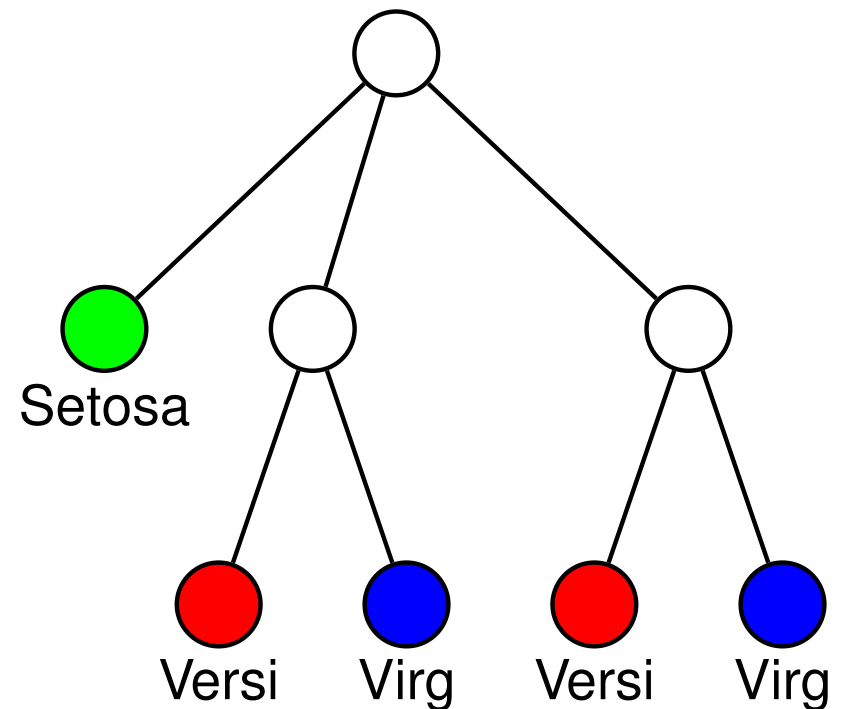
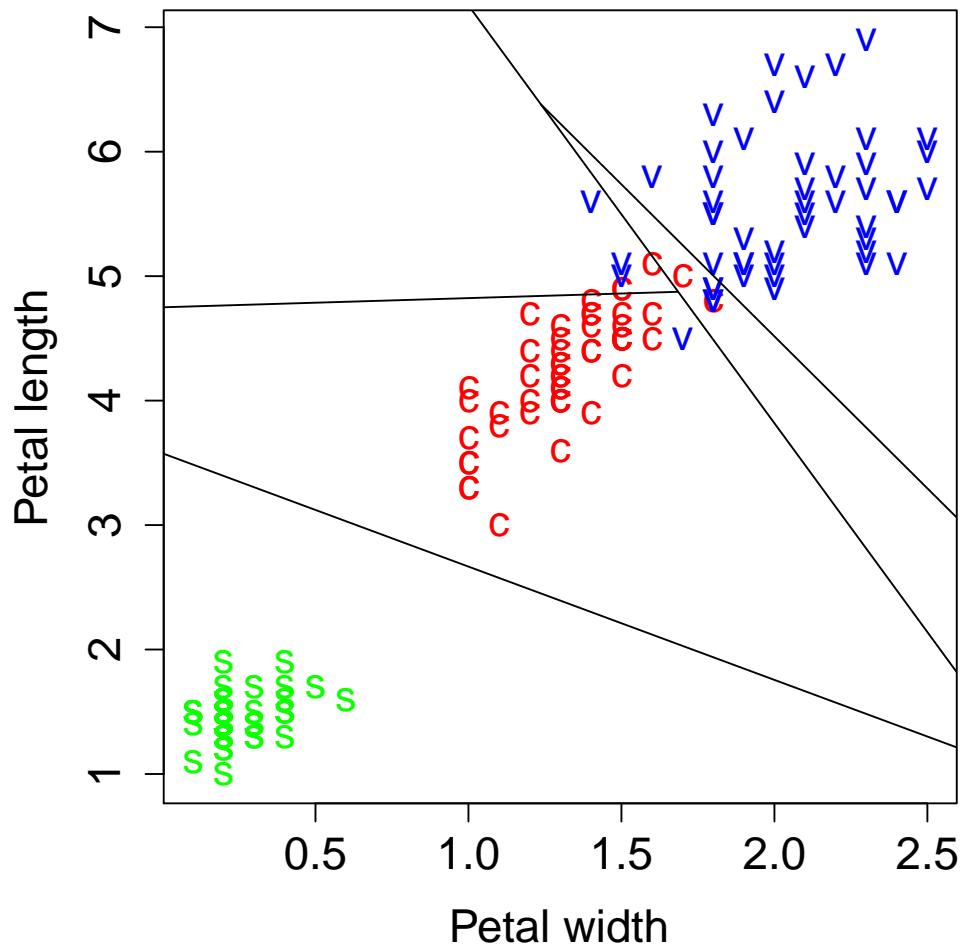
CART vs. linear discriminant analysis (LDA)



LDA as a linear-split tree



FACT (Loh and Vanichsetakul, 1988): Recursive linear discriminant analysis



FACT (Loh and Vanichsetakul, 1988)

Linear splits: Recursive LDA

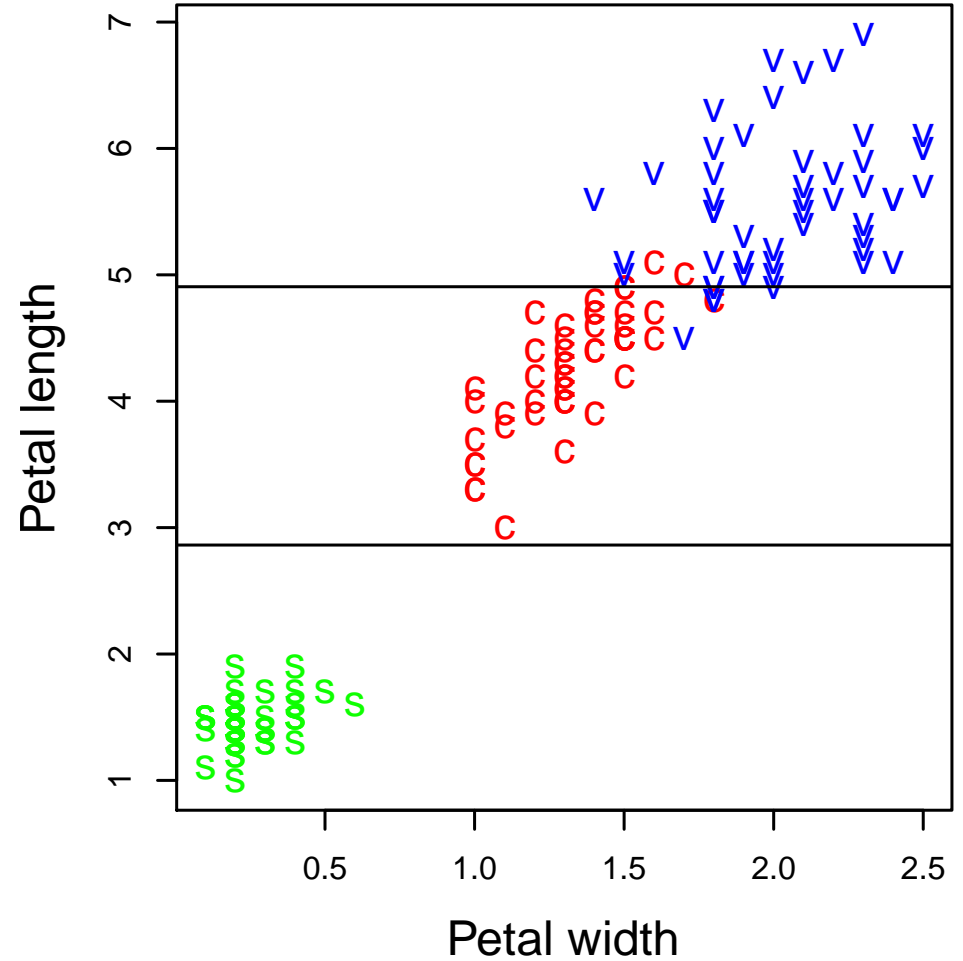
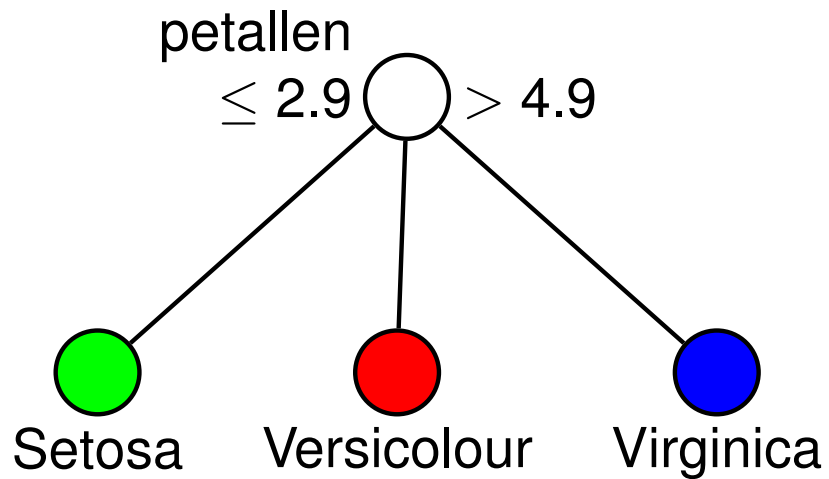
Univariate splits:

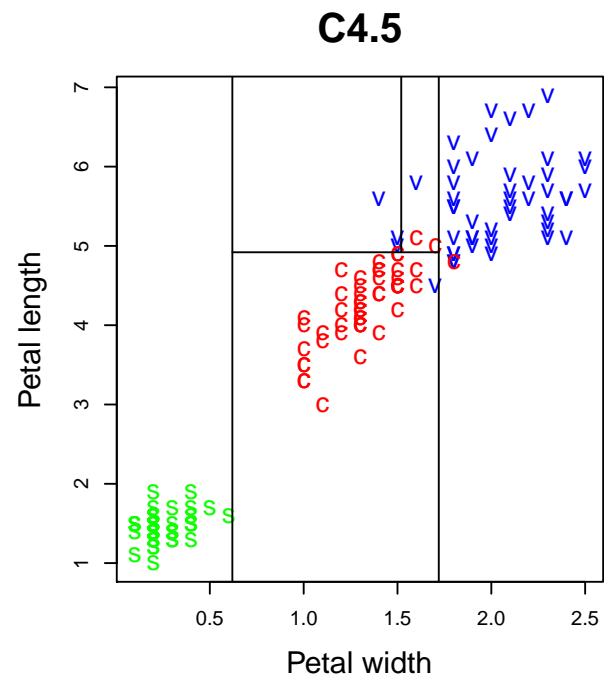
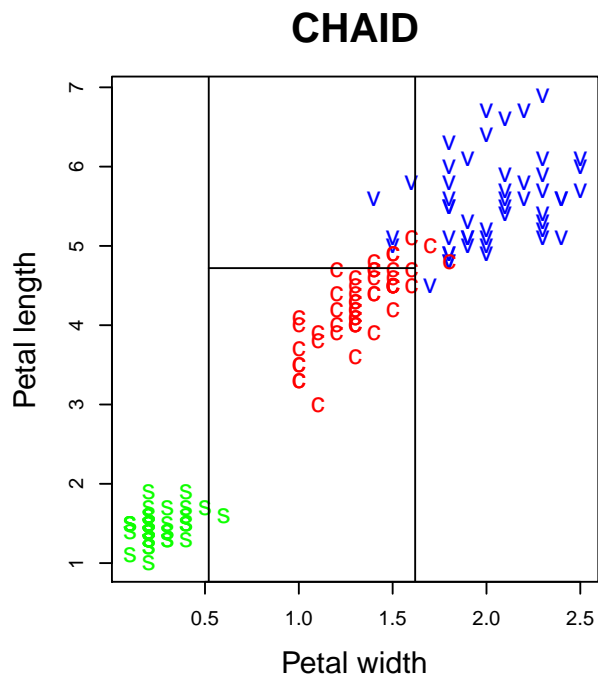
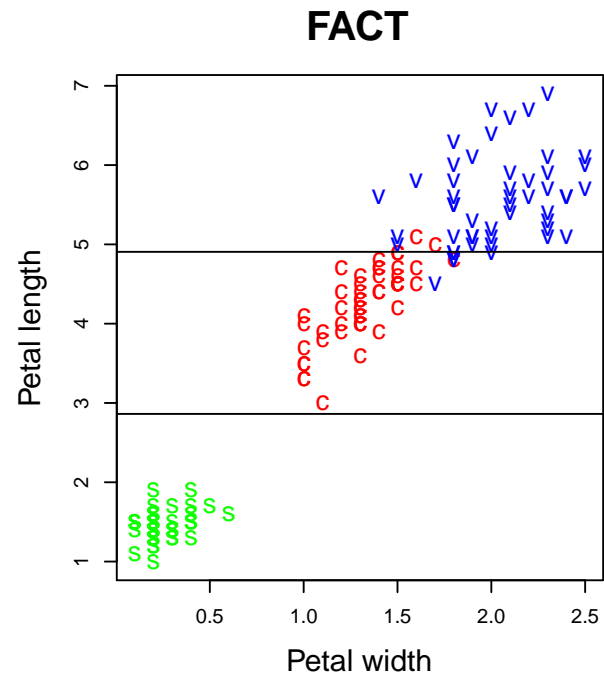
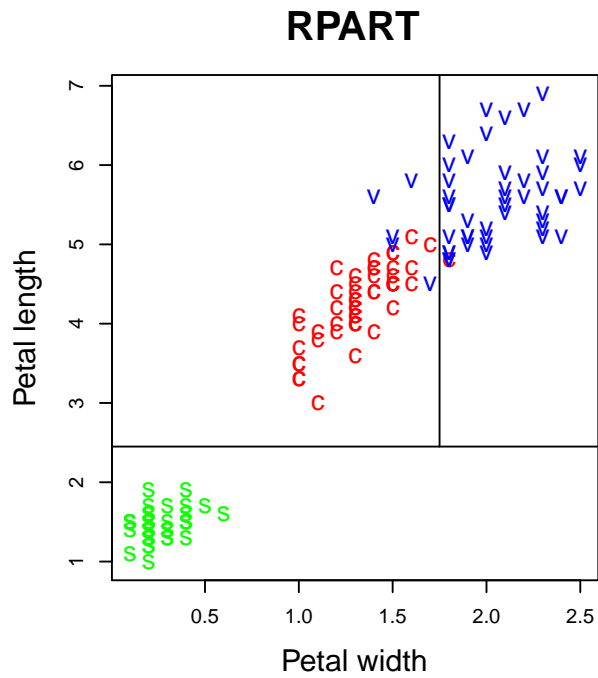
1. ANOVA F -tests to choose split variable
2. Univariate LDA to find split points (midpoints between class means)

Other features:

- Mean and mode missing value imputation at each node
- Each categorical variable transformed to dummy variables and then to 1st linear discriminant coord
- Stopping rule based on significance of F -tests

FACT univariate split tree for iris data





QUEST (Loh and Shih, 1997)

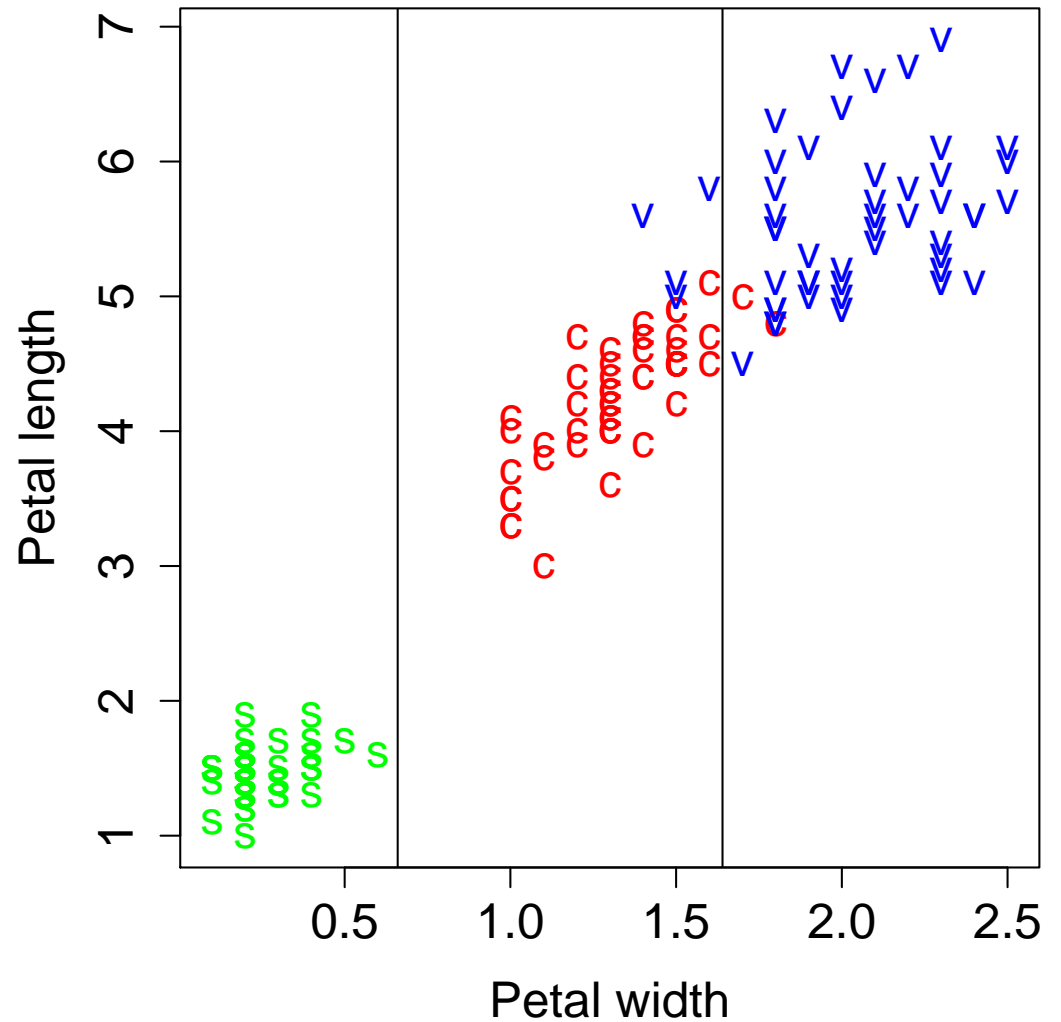
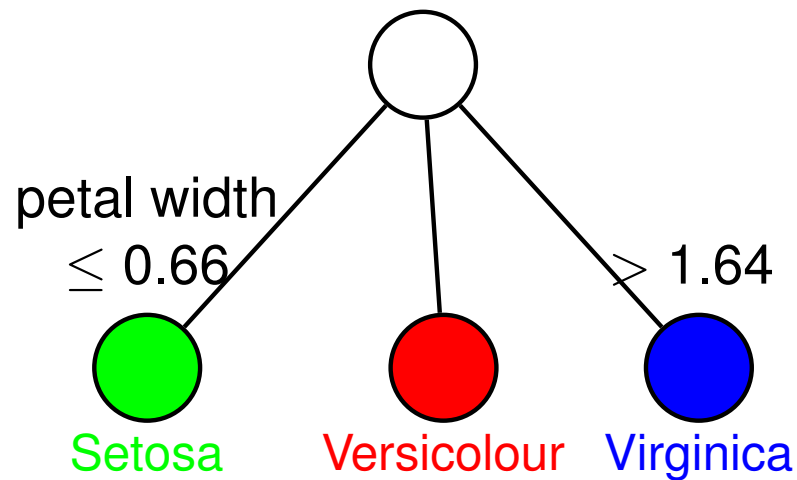
(first algorithm without selection bias)

1. Use ANOVA F and contingency table χ^2 -tests to select variables
2. Merge classes into two superclasses to get binary splits
3. Use QDA on the superclasses to find split point
4. If X is categorical, transform to largest discriminant coord
5. Use nodewise mean/mode imputation for with missing values
6. Prune with CART method

CRUISE (Kim and Loh, 2001, 2003)

- Split each node into two or more subnodes (à la FACT)
- Select split variables by contingency table χ^2 -tests
- Include tests for pairwise interactions
- Find split points by LDA after Box-Cox transformations
- Find linear splits by LDA on larger principal components
- Convert categorical X variables to discriminant coords
- Optionally fit LDA node models

CRUISE tree for iris data

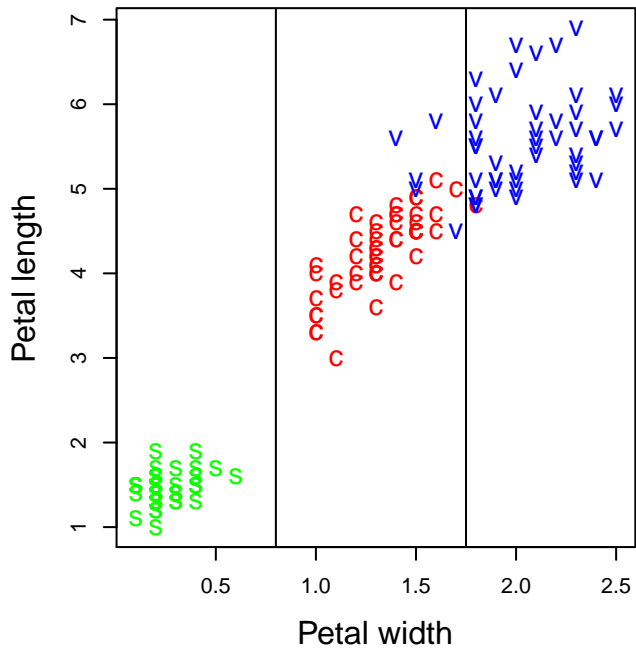


GUIDE classification (Loh, 2009)

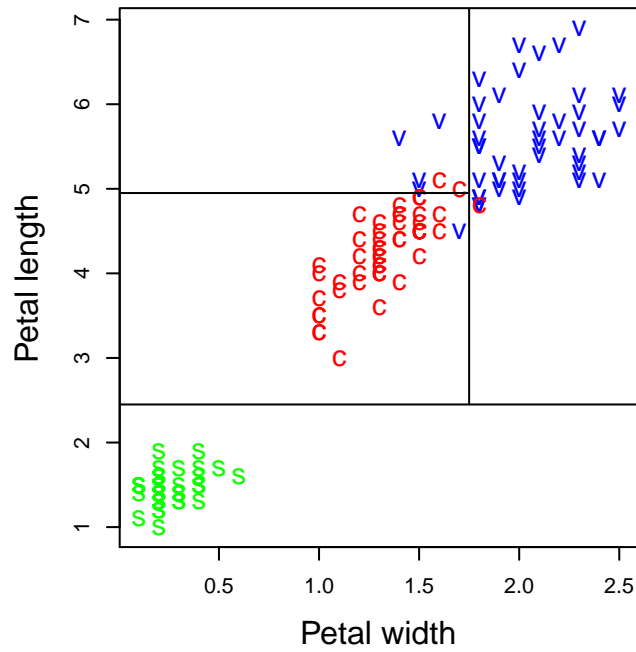
- Three-level tests for split selection: univariate, pairwise interaction, and pairwise linear
- Two-deep search if split is due to interactions
- Nearest-neighbor and kernel node models
- Missing values treated as separate category for split selection
- Bagging and random forest models
- Importance scores and thresholds

Iris data

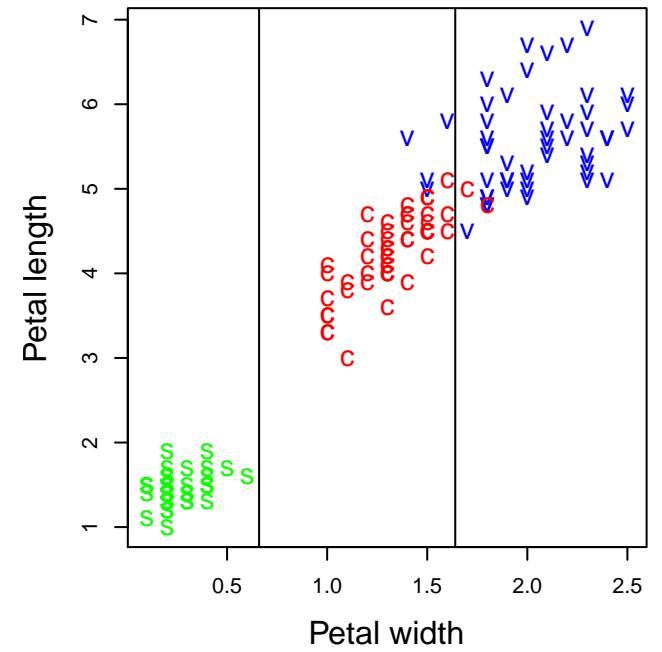
(d) GUIDE



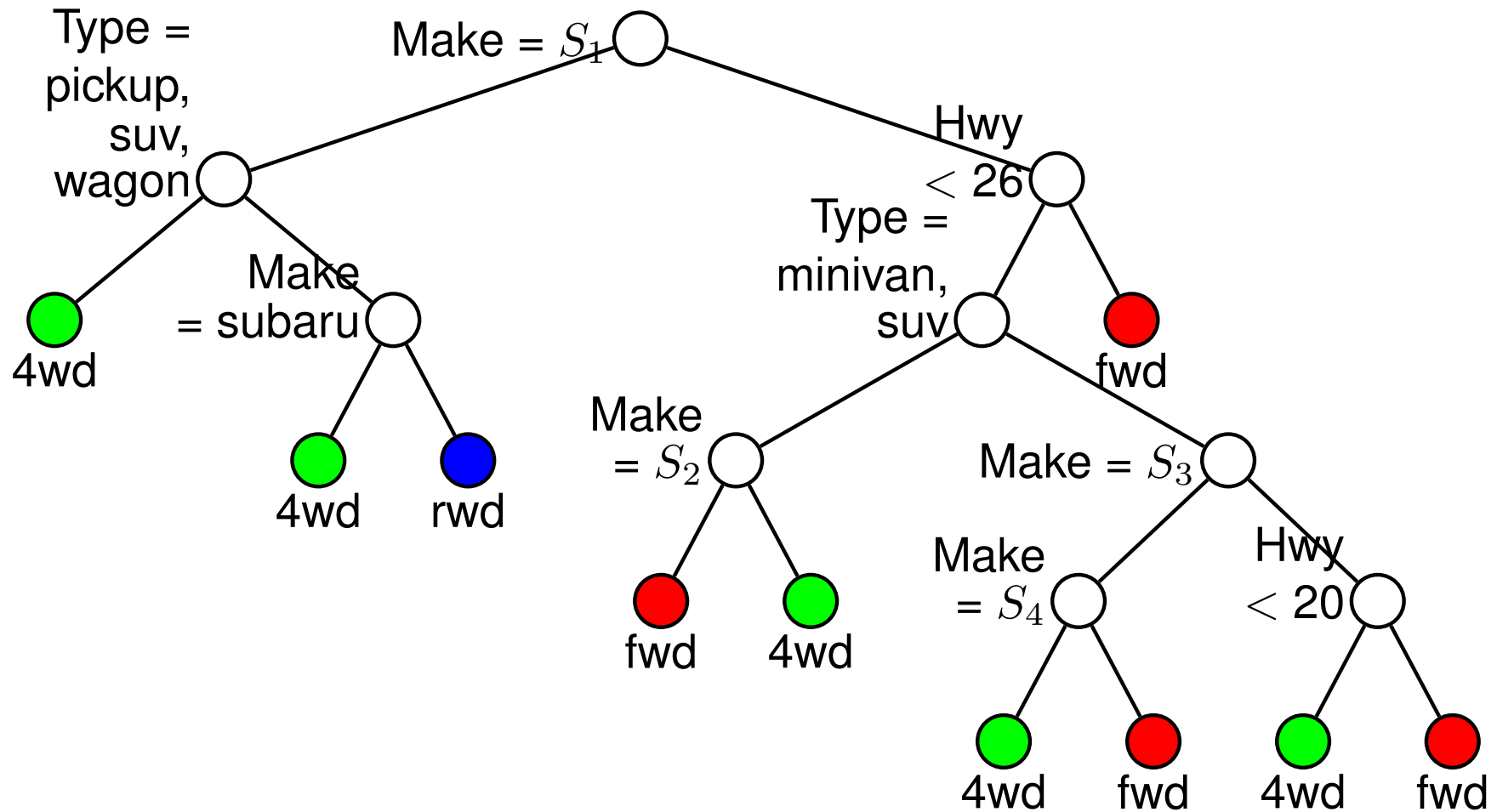
(e) QUEST



(f) CRUISE

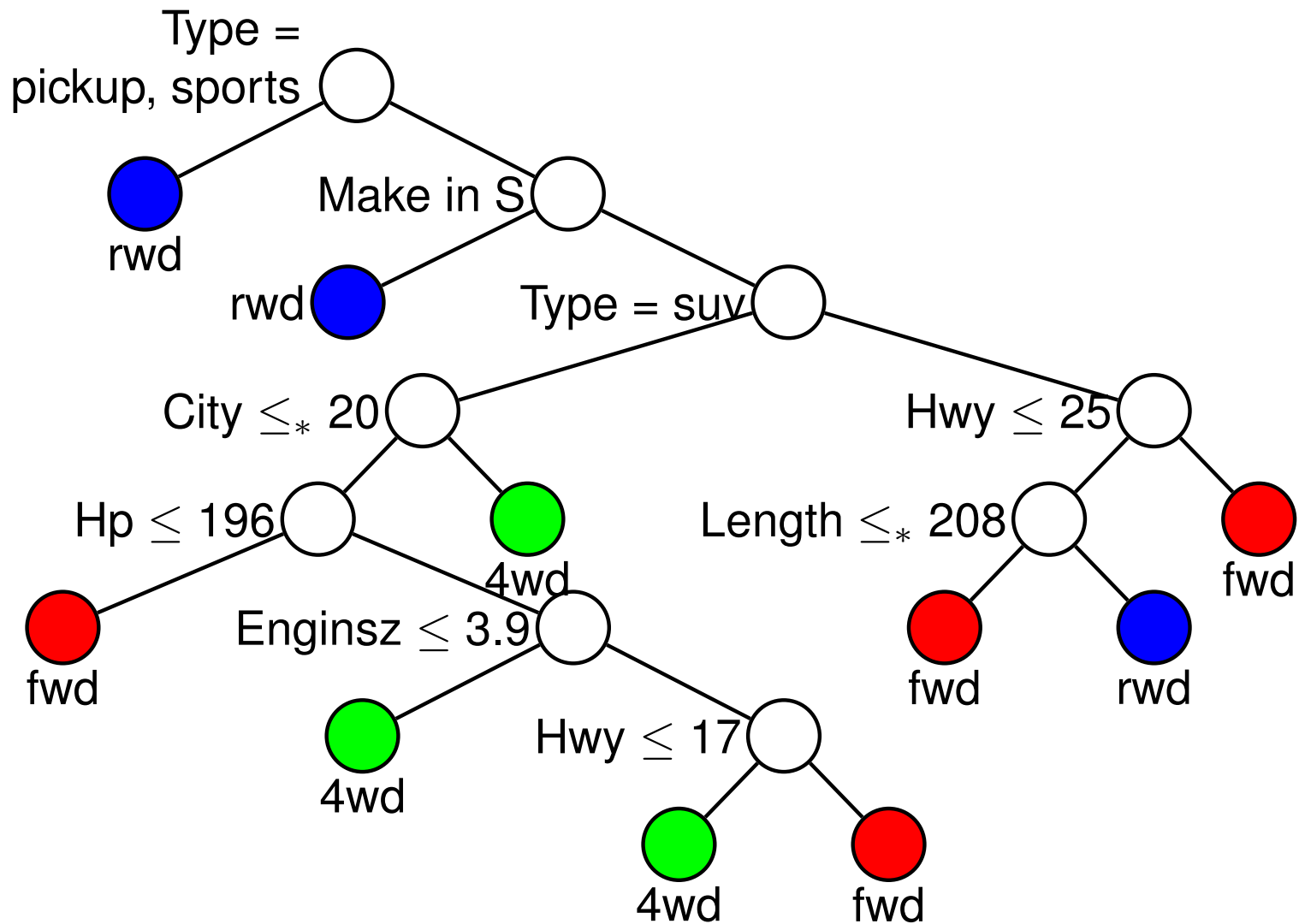


RPART tree for car data (took 22 cpu hrs!)



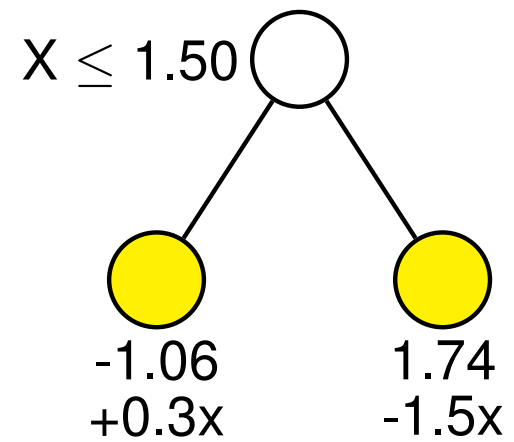
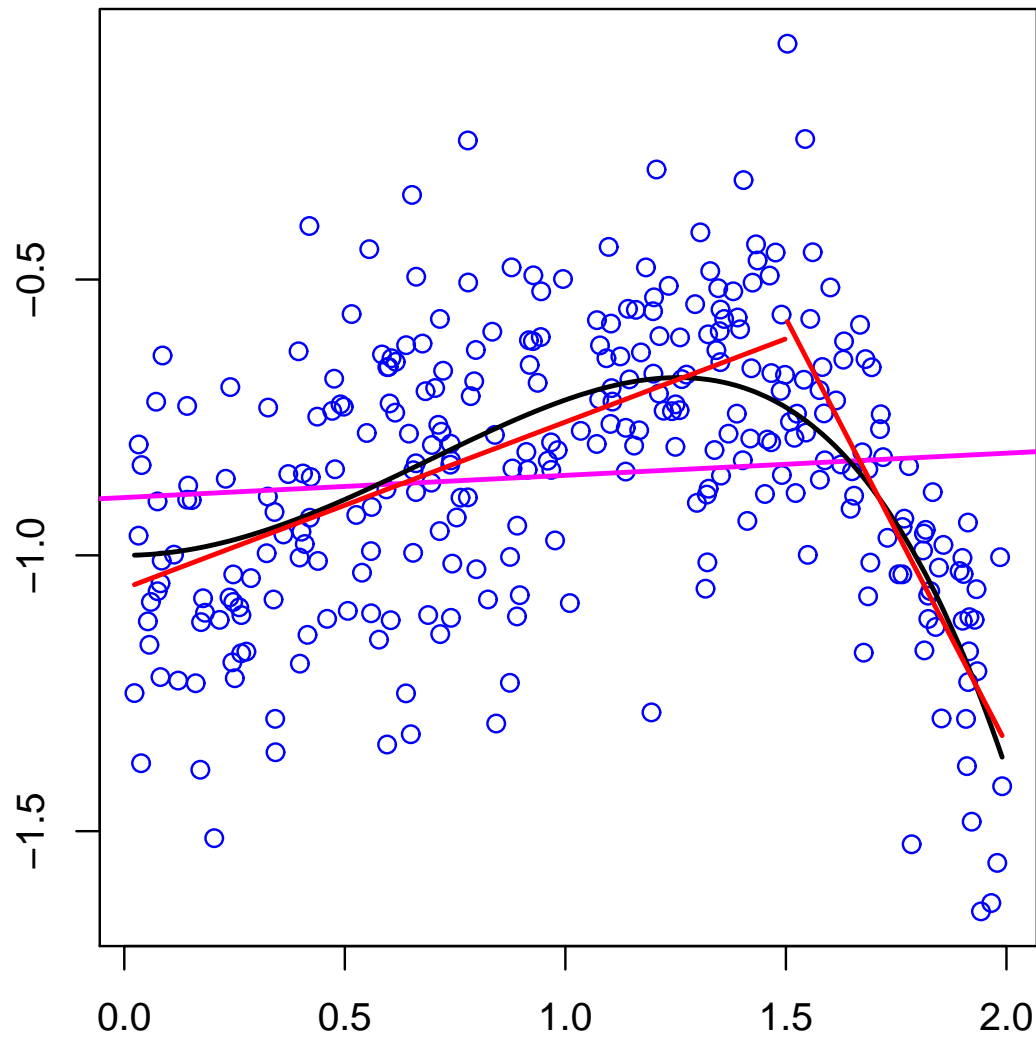
S_1 = BMW, GMC, Hummer, Infiniti, Jaguar, Land-Rover, Lexus, Lincoln, Mazda, Mercedes, Porsche, Subaru; S_2 = Cadillac, Chrysler, Kia, Mercury, Nissan; S_3 = Audi, Kia, Mitsubishi, Nissan, Pontiac, Volkswagen, Volvo; S_4 = Audi, Nissan, Volvo

GUIDE tree for car data (0.25 sec.)



$S = \{\text{Audi, BMW, Hummer, Infiniti, Isuzu, Jaguar, Jeep, Land-Rover, Lexus, Lincoln, Mercedes, Porsche, Subaru}\}$

Piecewise-linear regression tree



M5 (Quinlan, 1992)

1. First fit a piecewise-constant regression tree
2. Then fit a stepwise linear regression model to data in each node using split variables below the node

Torgo (1997) takes a similar approach, but allows kernel and nearest-neighbor models in terminal nodes.

GUIDE regression (Loh, 2002)

- Classification techniques applied to residuals to fit regression trees
- Constant, multiple, stepwise, polynomial and ANCOVA models
- Bootstrap bias correction for linear models
- Missing values treated as separate category for split selection
- For linear models, missing regressor values imputed with node means or fitted to separate constant models
- Bagging and random forest models
- Importance scores and thresholds

GUIDE regression models

- Quantile regression (Chaudhuri and Loh, 2002)
- Logistic regression (Chan and Loh, 2004)
- Poisson regression (Loh, 2006)
- Least-median of squares regression
- Longitudinal & multiresponse data (Loh and Zheng, 2013)
- Proportional hazards regression (Loh et al., 2015)
- Subgroup identification for differential treatment effects in randomized experiments (Loh et al., 2015)
- Detection of differential item functioning
- Propensity score matching and causal inference

Random forest (Breiman, 2001)

- Large set of CART trees constructed from bootstrap samples
- At each node, split selected from a random subset of variables
- Trees intentionally over-fitted (not pruned)
- Final predicted value is average of values from the trees
- Biased variable importance scores (Strobl et al., 2007)

Bayesian approaches (Denison et al., 1998; Chipman et al., 1998, 2002, 2010)

- Prior distributions on set of tree models
- Stochastic search to find good models
- Tree with largest posterior likelihood selected
- Resulting trees are random

Other randomized solutions

- TARGET (Fan and Gray, 2005; Gray and Fan, 2008) uses genetic algorithms for tree construction.
- *Extremely randomized trees* (Guerts et al., 2006) select splits from randomly picked subsets of split variables and split points.

PARTY: unbiased variable selection based on permutation tests

- CTREE (Hothorn et al., 2006)
 - Limited to classification and piecewise-constant regression
 - Surrogate splits for missing values
 - Response may be univariate, multivariate, ordinal, or censored
- MOB (Zeileis et al., 2008)
 - Least-squares, logistic, and parametric survival models
 - Based on tests of randomness of residual process along each X
 - Bonferroni adjustment of p-values to determine stopping
 - No method for missing values (na.action defaults to na.omit)
 - Not unbiased if X variables are used for fitting and splitting

Resurgent interest in trees: subgroup identification for differential treatment effects

- RTA (Dusseldorp and Meulman, 2004)
- STIMA (Dusseldorp et al., 2010)
- Interaction trees (Su et al., 2008, 2009)
- Virtual twins (Foster et al., 2011)
- SIDES (Lipkovich et al., 2011)
- QUINT (Dusseldorp and Van Mechelen, 2014)
- GUIDE (Loh et al., 2015)

Comment on Loh (2014) by Strobl (2014) regarding unbiased variable selection

*“One should think that the results shown here and in many previous studies . . . are so clear that **any statistically educated person should never want to use a biased recursive partitioning algorithm again.**”*

Yet I encounter so many cases where biased recursive partitioning algorithms are still employed both in applied and methodological publications.

*I really wonder why this is the case. Does it mean that the authors of those publications don't consider variable selection bias an issue of concern, or **willingly ignore decades of research?**”*

References

- Alexander, W. P. and Grimshaw, S. D. (1996). Treed regression. *Journal of Computational and Graphical Statistics*, 5:156–175.
- Breiman, L. (2001). Random forests. *Machine Learning*, 45:5–32.
- Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). *Classification and Regression Trees*. Chapman & Hall/CRC.
- Chan, K.-Y. and Loh, W.-Y. (2004). LOTUS: An algorithm for building accurate and comprehensible logistic regression trees. *Journal of Computational and Graphical Statistics*, 13:826–852.
- Chaudhuri, P. and Loh, W.-Y. (2002). Nonparametric estimation of conditional quantiles using quantile regression trees. *Bernoulli*, 8:561–576.

- Chipman, H. A., George, E. I., and McCulloch, R. E. (1998). Bayesian CART model search. *Journal of the American Statistical Association*, 93:935–948.
- Chipman, H. A., George, E. I., and McCulloch, R. E. (2002). Bayesian treed models. *Machine Learning*, 48:299–320.
- Chipman, H. A., George, E. I., and McCulloch, R. E. (2010). BART: Bayesian additive regression trees. *Annals of Applied Statistics*, 4:266–298.
- Ciampi, A. (1991). Generalized regression trees. *Computational Statistics and Data Analysis*, 12:57–78.
- Ciampi, A., Hogg, S. A., McKinney, S., and Thiffault, J. (1988). RECPAM: a computer program for recursive partition and amalgamation for censored survival data and other situations

frequently occurring in biostatistics. *Computer Methods and Programs in Biomedicine*, 26:239–256.

Davis, R. B. and Anderson, J. R. (1989). Exponential survival trees. *Statistics in Medicine*, 8:947–961.

De'ath, G. (2002). Multivariate regression trees: a new technique for modeling species-environment relationships. *Ecology*, 83:1105–1117.

Denison, D. G. T., Mallick, B. K., and Smith, A. F. M. (1998). A Bayesian CART algorithm. *Biometrika*, 85:363–377.

Dusseldorp, E., Conversano, C., and Van Os, B. J. (2010). Combining an additive and tree-based regression model simultaneously: STIMA. *Journal of Computational and Graphical Statistics*, 19:514–530.

Dusseldorp, E. and Meulman, J. J. (2004). The regression trunk approach to discover treatment covariate interaction. *Psychometrika*, 69:355–374.

Dusseldorp, E. and Van Mechelen, I. (2014). Qualitative interaction trees: a tool to identify qualitative treatment-subgroup interactions. *Statistics in Medicine*, 33:219–237.

Fan, G. and Gray, J. B. (2005). Regression tree analysis using TARGET. *Journal of Computational and Graphical Statistics*, 14:1–13.

Foster, J. C., Taylor, J. M. G., and Ruberg, S. J. (2011). Subgroup identification from randomized clinical trial data. *Statistics in Medicine*, 30:2867–2880.

Gordon, L. and Olshen, R. A. (1985). Tree-structured survival analysis. *Cancer Treatment Reports*, 69:1065–1069.

Gray, J. B. and Fan, G. (2008). Classification tree analysis using TARGET. *Computational Statistics and Data Analysis*, 52:1362–1372.

Guerts, P., Ernst, D., and Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning*, 63:3–42.

Hothorn, T., Hornik, K., and Zeileis, A. (2006). Unbiased recursive partitioning: a conditional inference framework. *Journal of Computational and Graphical Statistics*, 15:651–674.

Kass, G. V. (1980). An exploratory technique for investigating large quantities of categorical data. *Applied Statistics*, 29:119–127.

Kim, H. and Loh, W.-Y. (2001). Classification trees with unbiased multiway splits. *Journal of the American Statistical Association*, 96:589–604.

Kim, H. and Loh, W.-Y. (2003). Classification trees with bivariate linear discriminant node models. *Journal of Computational and Graphical Statistics*, 12:512–530.

LeBlanc, M. and Crowley, J. (1992). Relative risk trees for censored survival data. *Biometrics*, 48:411–425.

Lee, S. K. (2005). On generalized multivariate decision tree by using GEE. *Computational Statistics and Data Analysis*, 49:1105–1119.

Lipkovich, I., Dmitrienko, A., Denne, J., and Enas, G. (2011). Subgroup identification based on differential effect search — a recursive partitioning method for establishing response to treatment in patient subpopulations. *Statistics in Medicine*, 30:2601–2621.

Loh, W.-Y. (2002). Regression trees with unbiased variable

selection and interaction detection. *Statistica Sinica*, 12:361–386.

Loh, W.-Y. (2006). Regression tree models for designed experiments. In Rojo, J., editor, *Second E. L. Lehmann Symposium*, volume 49 of *IMS Lecture Notes-Monograph Series*, pages 210–228. Institute of Mathematical Statistics.

Loh, W.-Y. (2009). Improving the precision of classification trees. *Annals of Applied Statistics*, 3:1710–1737.

Loh, W.-Y. (2014). Fifty years of classification and regression trees (with discussion). *International Statistical Review*, 34:329–370.

Loh, W.-Y., He, X., and Man, M. (2015). A regression tree approach to identifying subgroups with differential treatment effects. *Statistics in Medicine*, 34:1818–1833.

- Loh, W.-Y. and Shih, Y.-S. (1997). Split selection methods for classification trees. *Statistica Sinica*, 7:815–840.
- Loh, W.-Y. and Vanichsetakul, N. (1988). Tree-structured classification via generalized discriminant analysis (with discussion). *Journal of the American Statistical Association*, 83:715–728.
- Loh, W.-Y. and Zheng, W. (2013). Regression trees for longitudinal and multiresponse data. *Annals of Applied Statistics*, 7:495–522.
- Messenger, R. and Mandell, L. (1972). A modal search technique for predictive nominal scale multivariate analysis. *Journal of the American Statistical Association*, 67:768–772.
- Morgan, J. N. and Sonquist, J. A. (1963). Problems in the analysis of survey data, and a proposal. *Journal of the American Statistical Association*, 58:415–434.

- Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, 1:81–106.
- Quinlan, J. R. (1992). Learning with continuous classes. In *5th Australian Joint Conference on Artificial Intelligence*, pages 343–348.
- Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann.
- Segal, M. R. (1988). Regression trees for censored data. *Biometrics*, 44:35–47.
- Segal, M. R. (1992). Tree structured methods for longitudinal data. *Journal of the American Statistical Association*, 87:407–418.
- Sela, R. J. and Simonoff, J. S. (2012). RE-EM trees: a data mining approach for longitudinal and clustered data. *Machine Learning*, 86:169–207.

- Strobl, C. (2014). Discussion of Loh (2014), “Fifty years of classification and regression trees”. *International Statistical Review*. In press.
- Strobl, C., Boulesteix, A., Zeileis, A., and Hothorn, T. (2007). Bias in random forest variable importance measures: illustrations, sources and a solution. *BMC Bioinformatics*, 8:25.
- Su, X., Tsai, C. L., Wang, H., Nickerson, D. M., and Bogong, L. (2009). Subgroup analysis via recursive partitioning. *Journal of Machine Learning Research*, 10:141–158.
- Su, X., Zhou, T., Yan, X., Fan, J., and Yang, S. (2008). Interaction trees with censored survival data. *International Journal of Biostatistics*, 4. Article 2.
- Su, X. G., Wang, M., and Fan, J. J. (2004). Maximum likelihood regression trees. *Journal of Computational and Graphical Statistics*, 13:586–598.

- Therneau, T., Atkinson, B., and Ripley, B. (2014). *rpart: Recursive Partitioning and Regression Trees*. R package version 4.1-8.
- Torgo, L. (1997). Functional models for regression tree leaves. In Fisher, D. H., editor, *Proceedings of the Fourteenth International Conference on Machine Learning*, pages 385–393, Burlington, MA. Morgan Kaufmann.
- Yu, Y. and Lambert, D. (1999). Fitting trees to functional data, with an application to time-of-day patterns. *Journal of Computational and Graphical Statistics*, 8:749–762.
- Zeileis, A., Hothorn, T., and Hornik, K. (2008). Model-based recursive partitioning. *Journal of Computational and Graphical Statistics*, 17:492–514.
- Zhang, H. (1998). Classification trees for multiple binary responses. *Journal of the American Statistical Association*, 93:180–193.