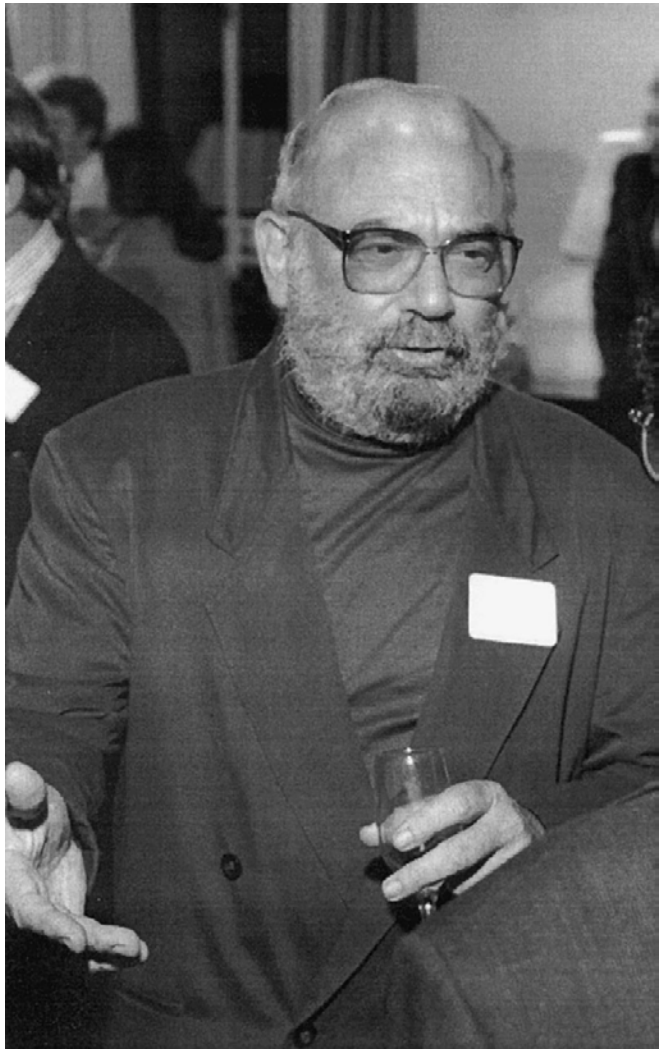
Random Forests for Regression and Classification

Adele Cutler

Utah State University

Leo Breiman, 1928 - 2005



1954: PhD Berkeley (mathematics)

1960 -1967: UCLA (mathematics)

1969 -1982: Consultant

1982 - 1993 Berkeley (statistics)

1984 “Classification & Regression Trees”
(with Friedman, Olshen, Stone)

1996 “Bagging”

2001 “Random Forests”

Impact

CART (1984)
27,926 citations

Bagging (1996)
13,090

Random Forests (2001)
17,492 citations

Total 72,796

Random Forests for Regression and Classification

Outline

- Background.
- Trees.
- Bagging.
- Random Forests.
- Variable importance.
- Partial dependence plots and interpretation of effects.
- Proximity.
- Visualization.
- New developments.

What is Regression?

Given data on predictor variables (inputs, X) and a **continuous response variable** (output, Y) build a model for:

- Predicting the value of the response from the predictors.
- Understanding the relationship between the predictors and the response.

e.g. predict a person's **systolic blood pressure** based on their age, height, weight, etc.

Regression Examples

- Y: **income**

  X: age, education, sex, occupation, …
- Y: **crop yield**

  X: rainfall, temperature, humidity, …
- Y: **test scores**

  X: teaching method, age, sex, ability, …
- Y: **selling price of homes**

  X: size, age, location, quality, …

Regression Background

- Linear regression $Y = \beta_0 + \beta_1 X + \varepsilon$
- Multiple linear regression e.g.

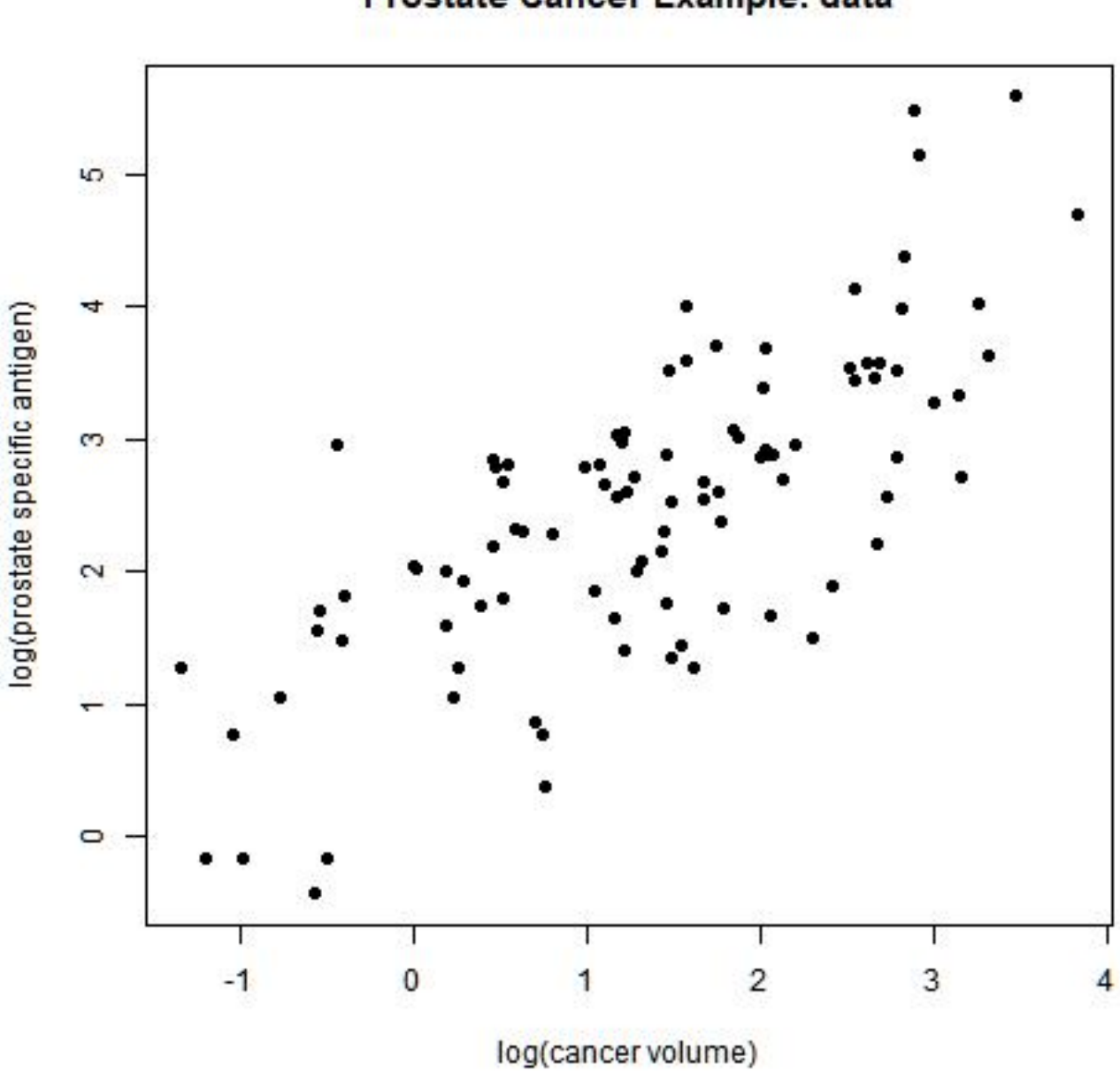
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

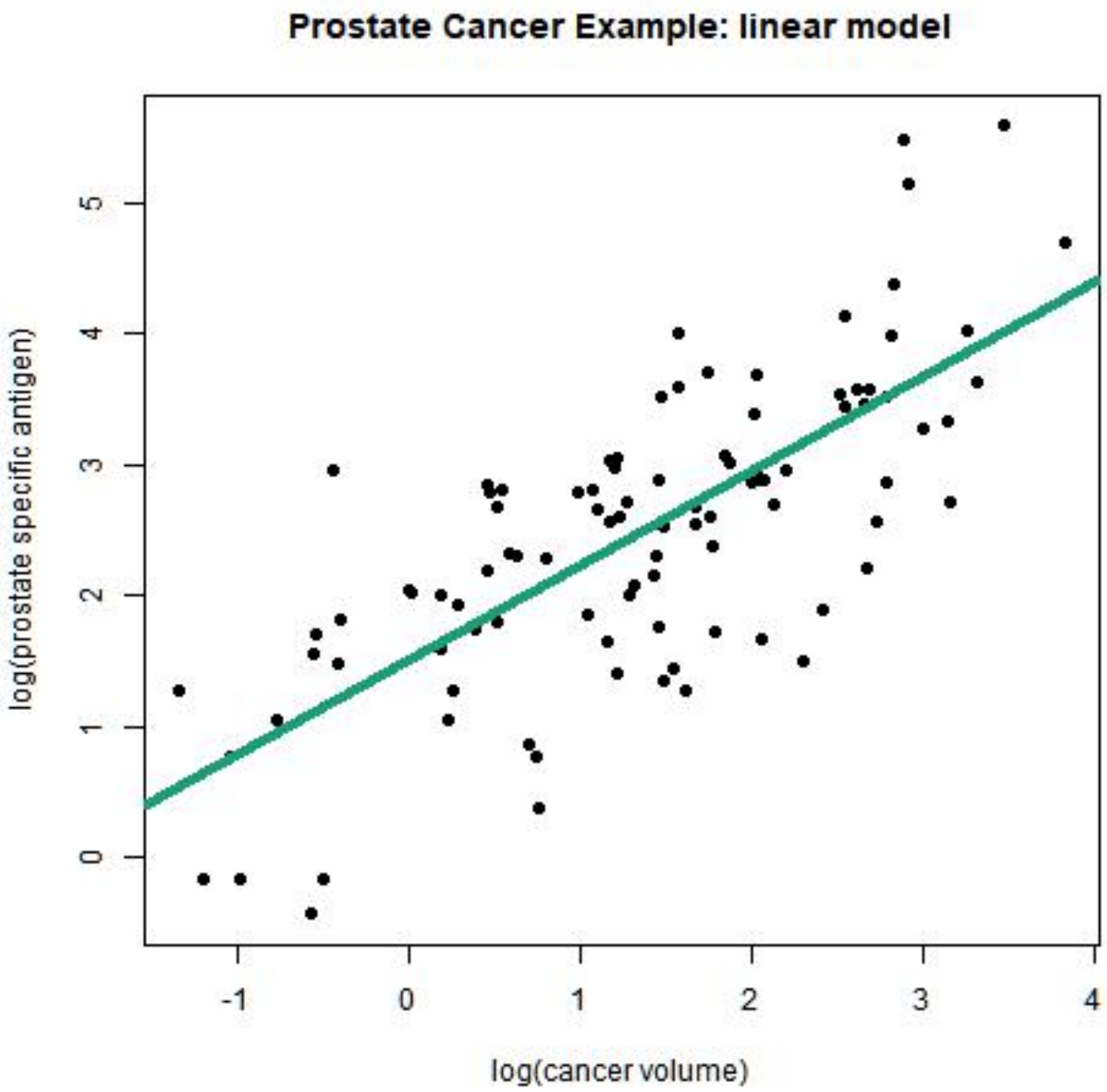
- Nonlinear regression (parametric) e.g.

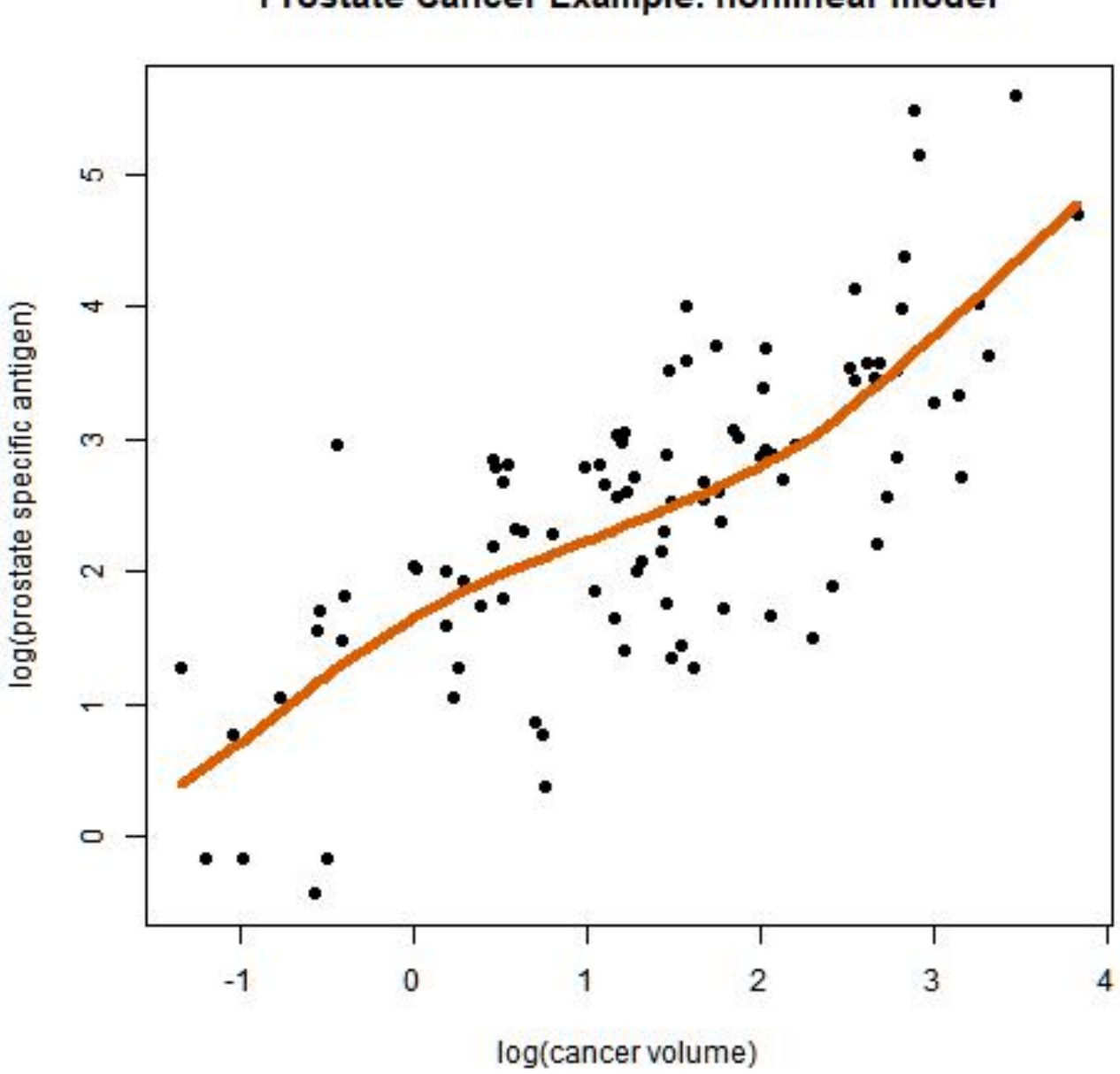
$$Y = \beta_0 + \beta_1 X + \beta_1 X^2 + \varepsilon$$

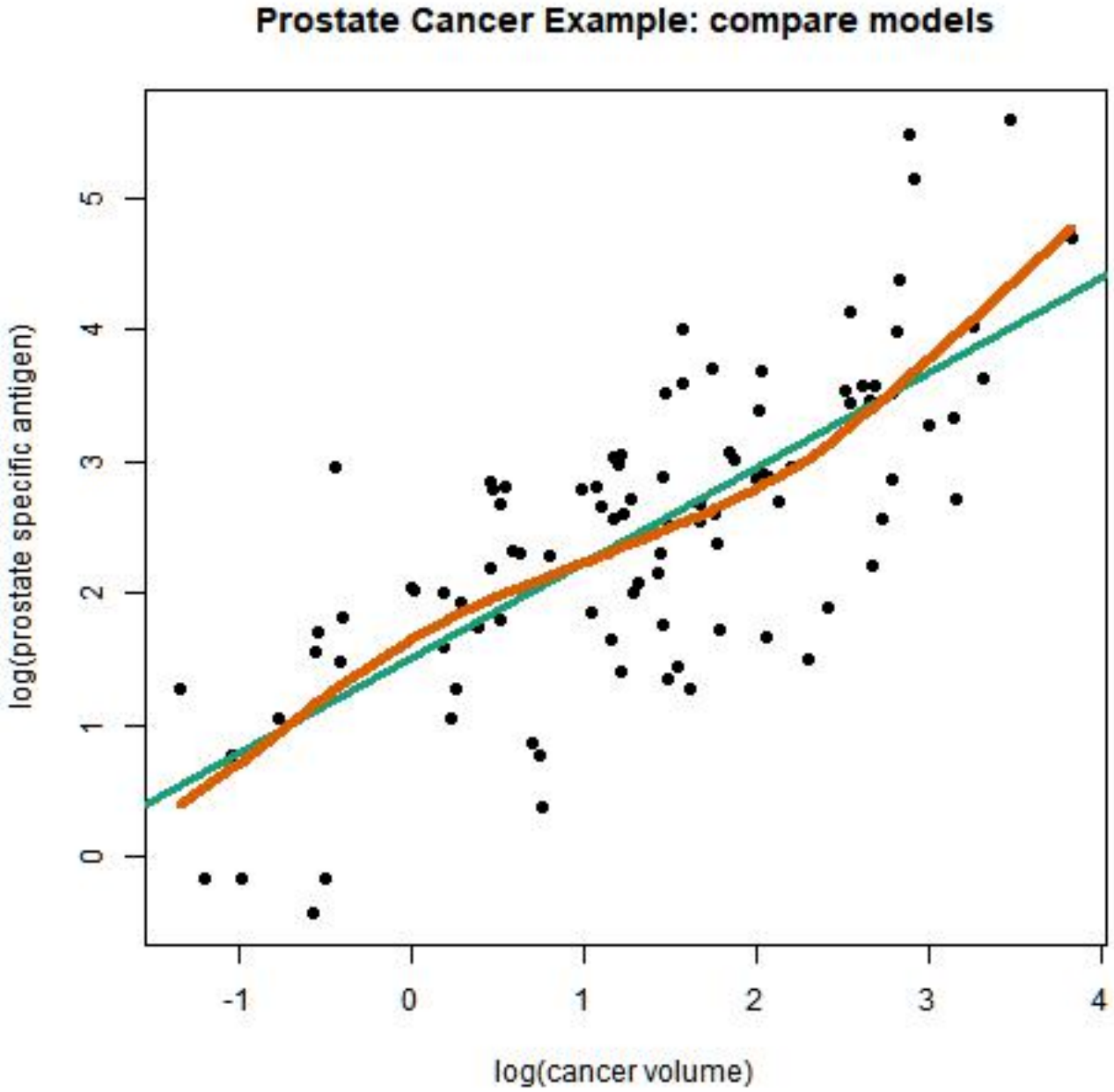
- Nonparametric regression (smoothing)
  - Kernel smoothing
  - B-splines
  - Smoothing splines
  - Wavelets

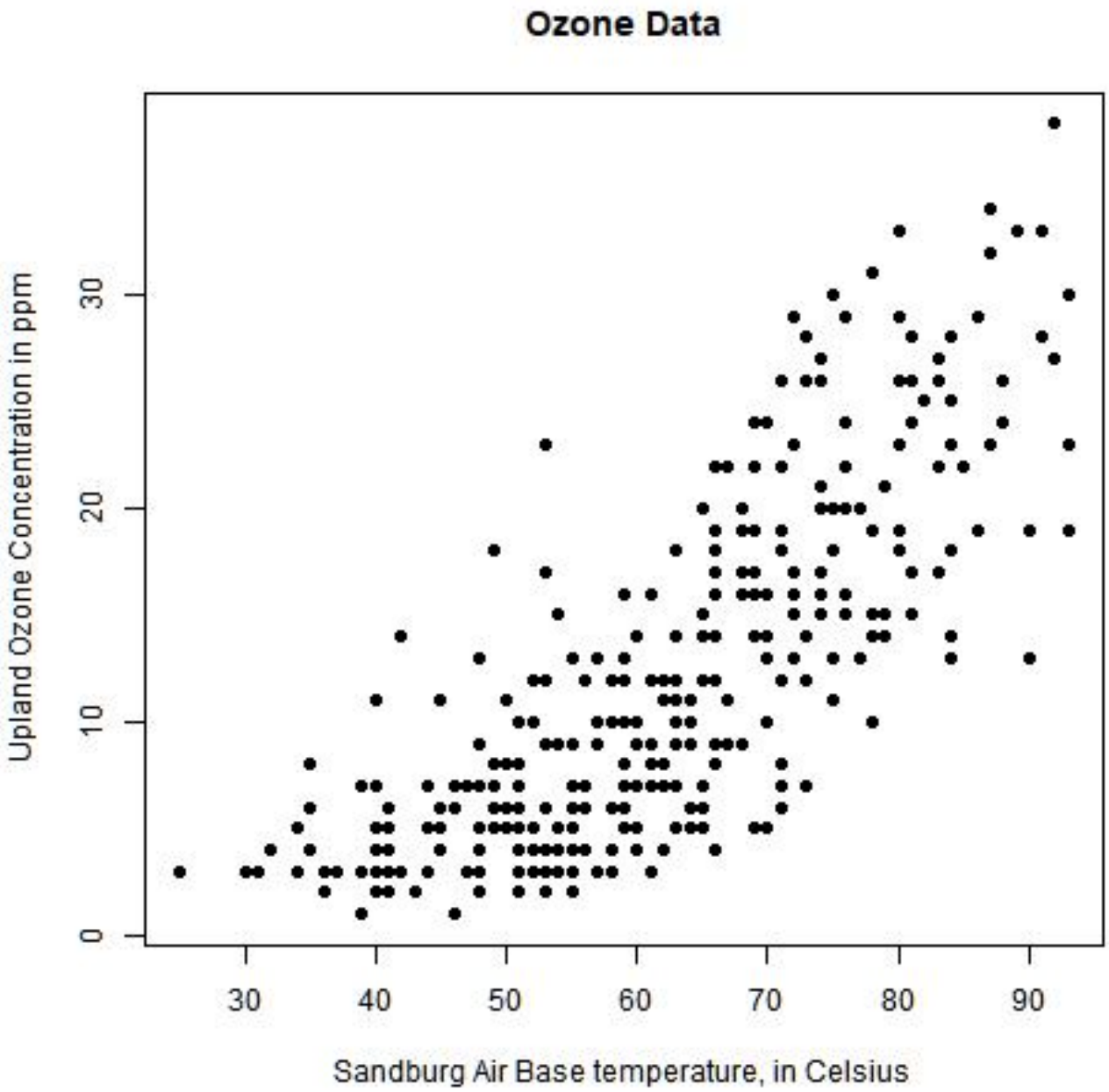
} Data-driven

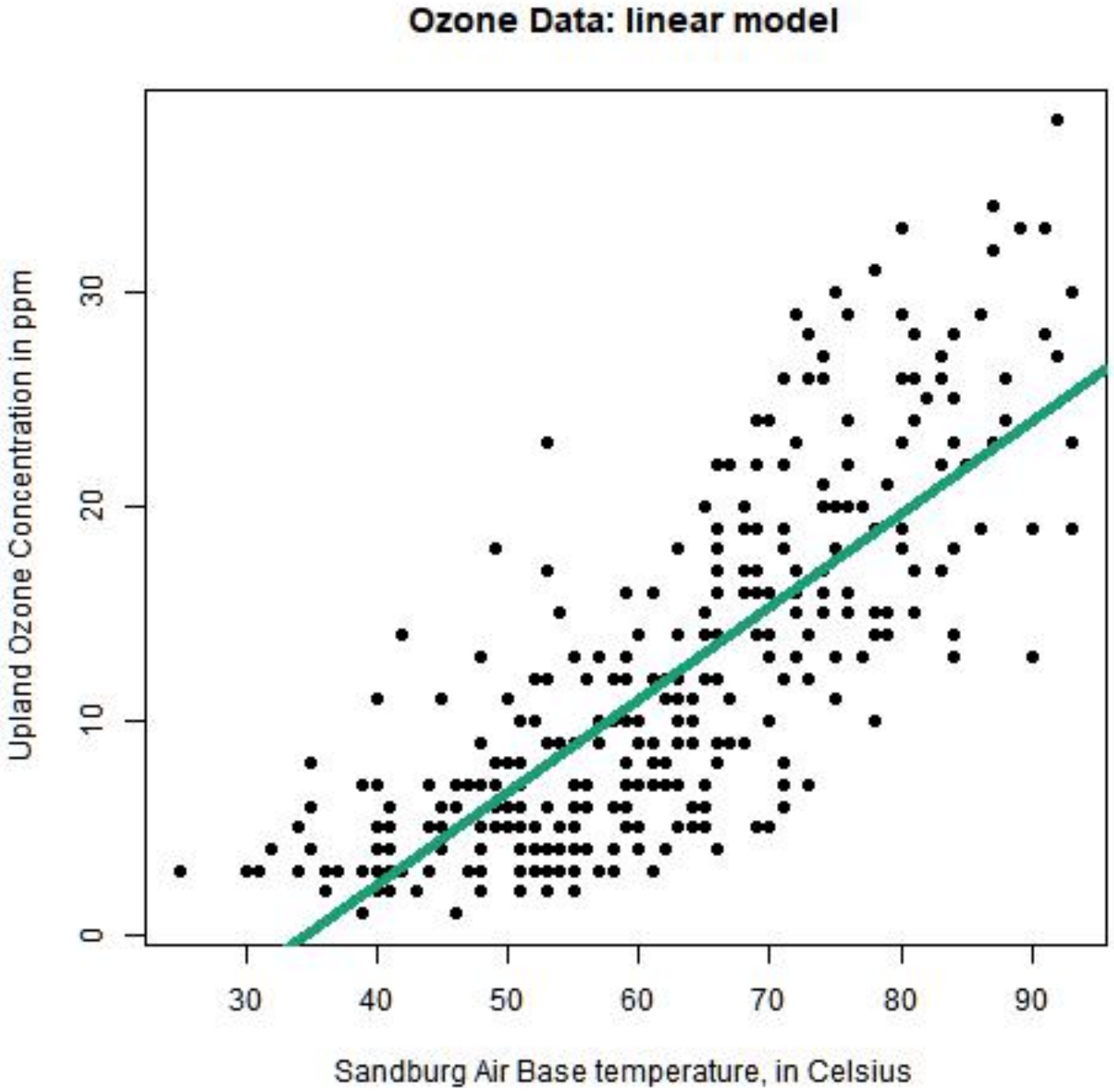

Prostate Cancer Example: data
log(prostate specific antigen)
log(cancer volume)
0
1
2
3
4
5
-1
0
1
2
3
4

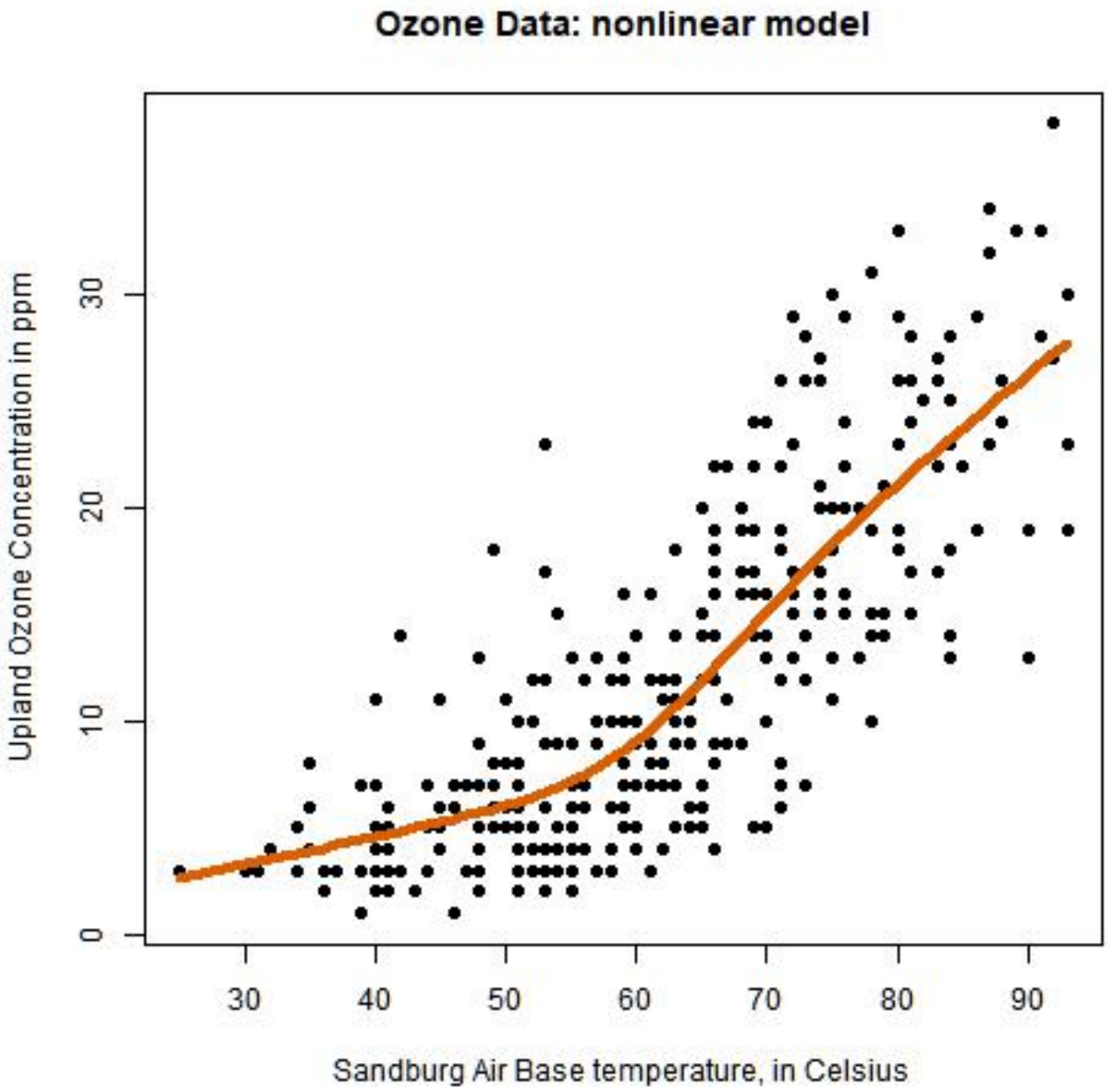

Prostate Cancer Example: linear model
log(prostate specific antigen)
log(cancer volume)
0
1
2
3
4
5
-1
0
1
2
3
4

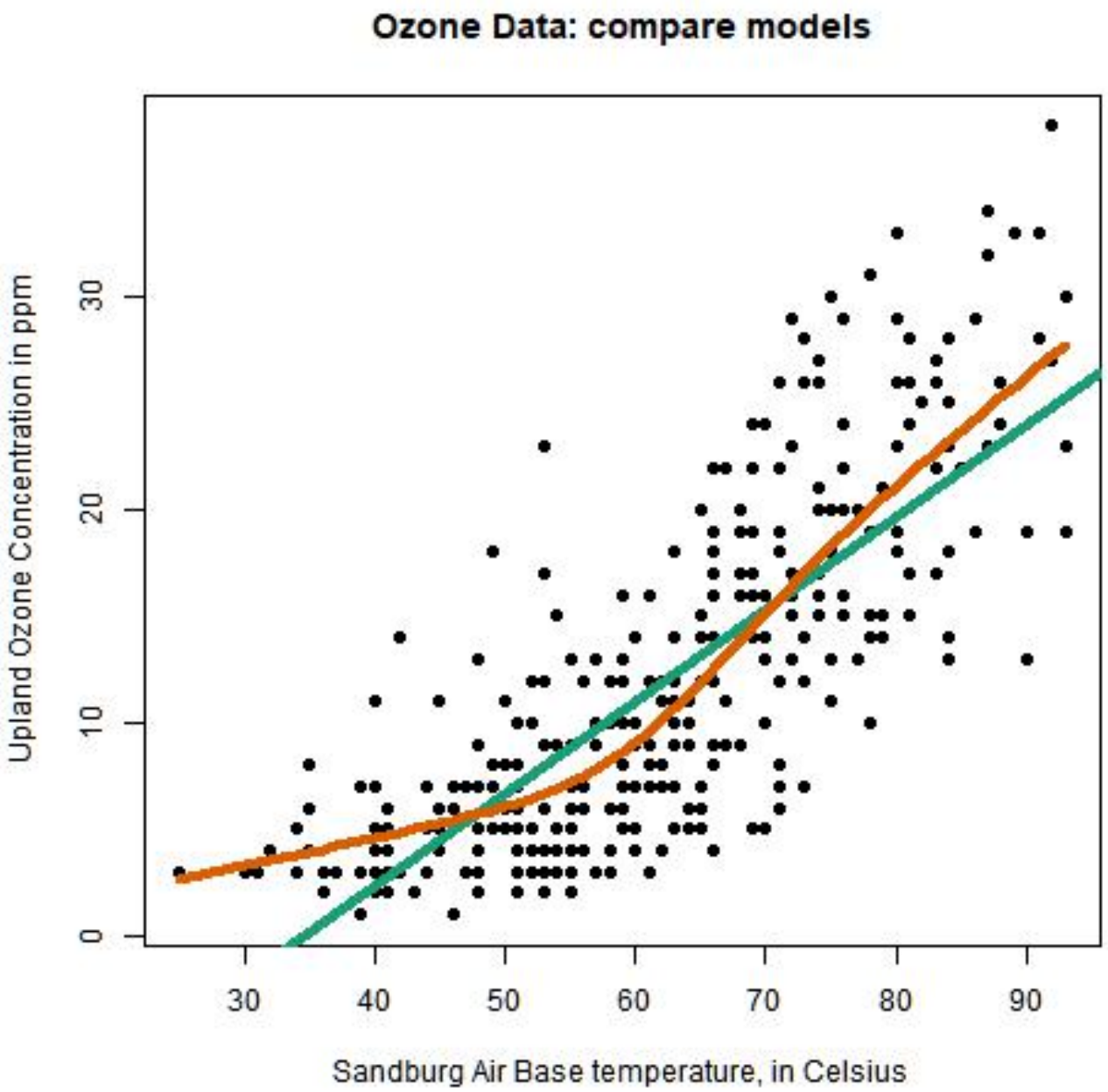

Prostate Cancer Example: nonlinear model
log(prostate specific antigen)
log(cancer volume)
0
1
2
3
4
5
-1
0
1
2
3
4


Prostate Cancer Example: compare models
log(prostate specific antigen)
log(cancer volume)
0
1
2
3
4
5
-1
0
1
2
3
4


Ozone Data
Upland Ozone Concentration in ppm
0
10
20
30
30
40
50
60
70
80
90
Sandburg Air Base temperature, in Celsius


Ozone Data: linear model
Upland Ozone Concentration in ppm
0
10
20
30
30
40
50
60
70
80
90
Sandburg Air Base temperature, in Celsius


Ozone Data: nonlinear model
Upland Ozone Concentration in ppm
0
10
20
30
30
40
50
60
70
80
90
Sandburg Air Base temperature, in Celsius


Ozone Data: compare models
Upland Ozone Concentration in ppm
0
10
20
30
30
40
50
60
70
80
90
Sandburg Air Base temperature, in Celsius

What is Classification?

Given data on predictor variables (inputs, X) and a **categorical response variable** (output, Y) build a model for:

- Predicting the value of the response from the predictors.
- Understanding the relationship between the predictors and the response.

e.g. predict a person's **5-year-survival (yes/no)** based on their age, height, weight, etc.

Classification Examples

- Y: **presence/absence of disease**

  X: diagnostic measurements
- Y: **land cover (grass, trees, water, roads…)**

  X: satellite image data (frequency bands)
- Y: **loan defaults (yes/no)**

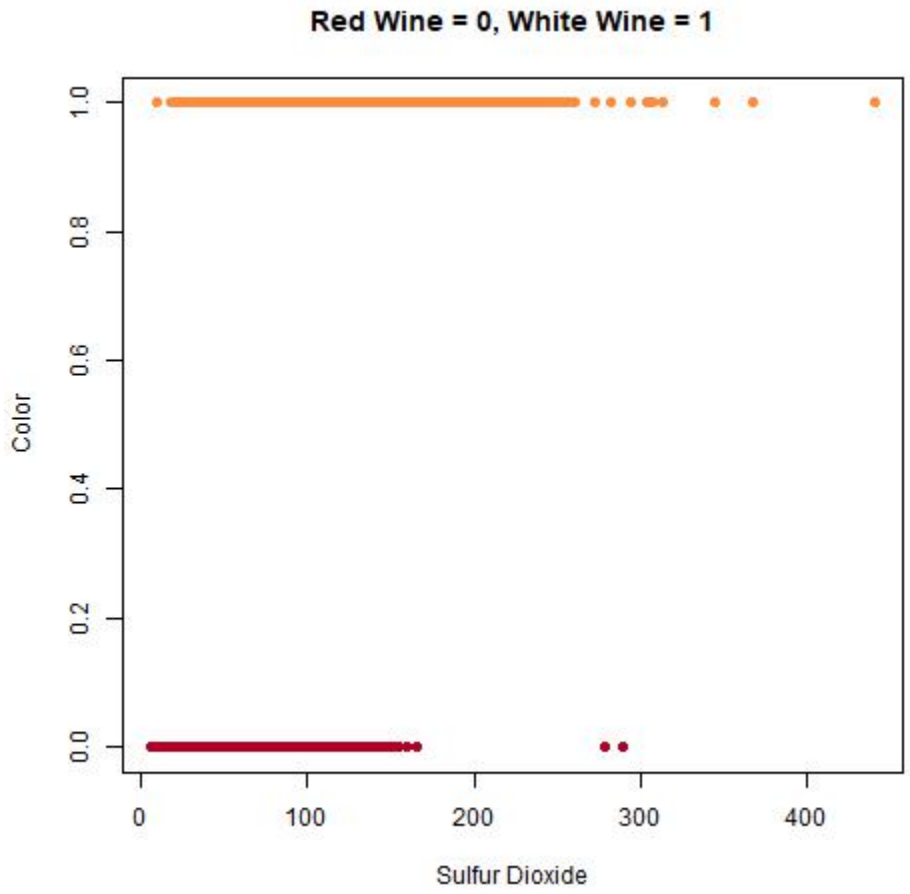
  X: credit score, own or rent, age, marital status, …
- Y: **dementia status**

  X: scores on a battery of psychological tests

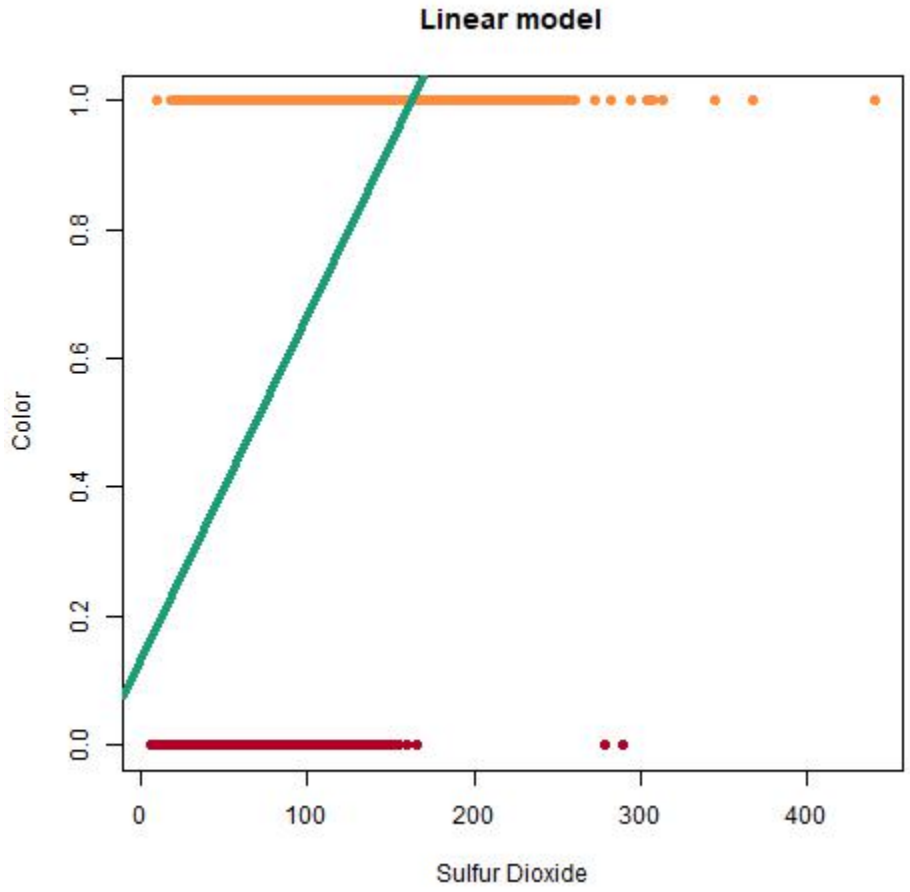
Classification Background

- Linear discriminant analysis (1930's)
- Logistic regression (1944)
- Nearest neighbors classifiers (1951)

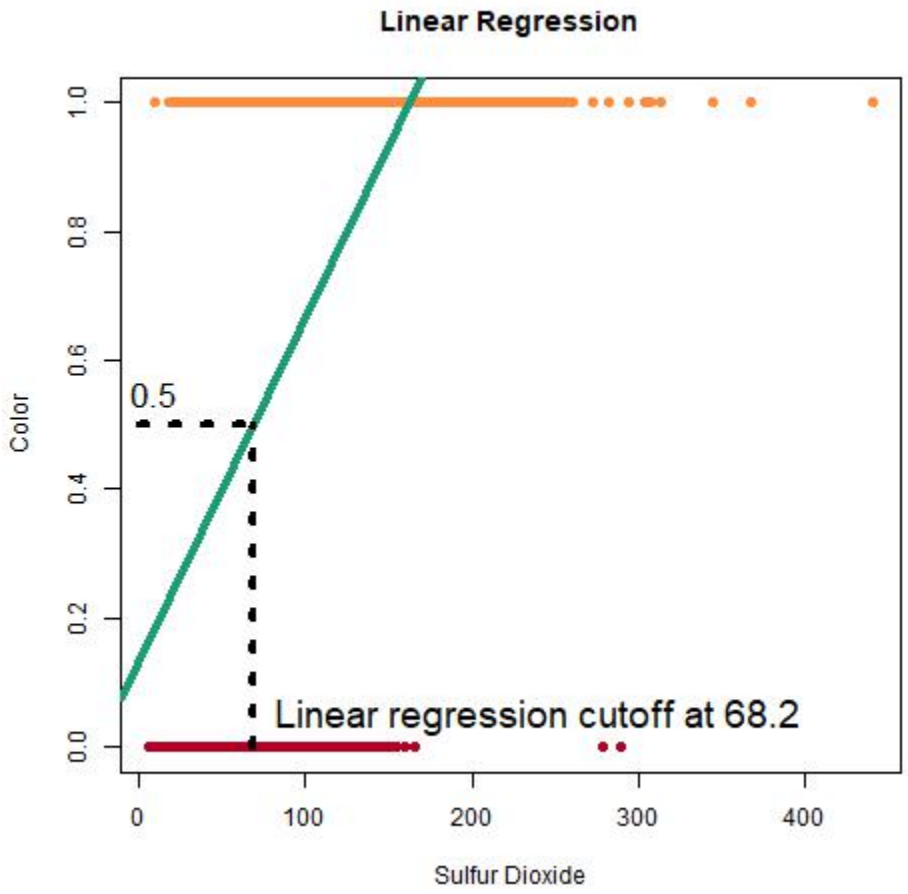
Classification Picture



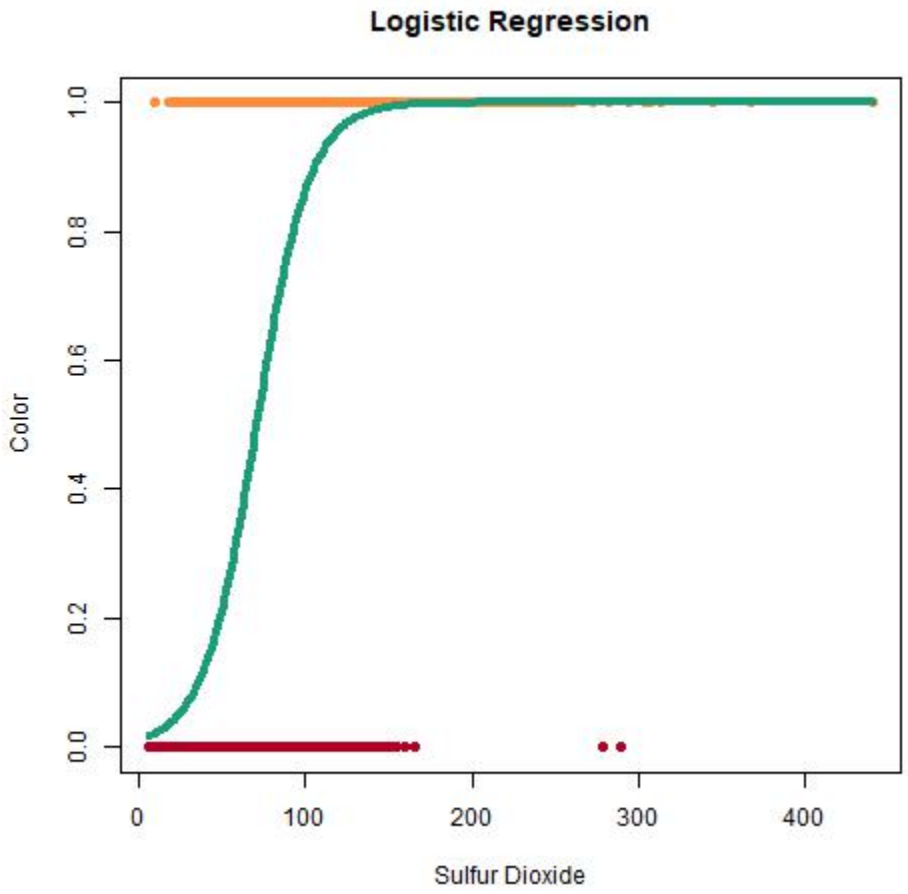
Classification Picture



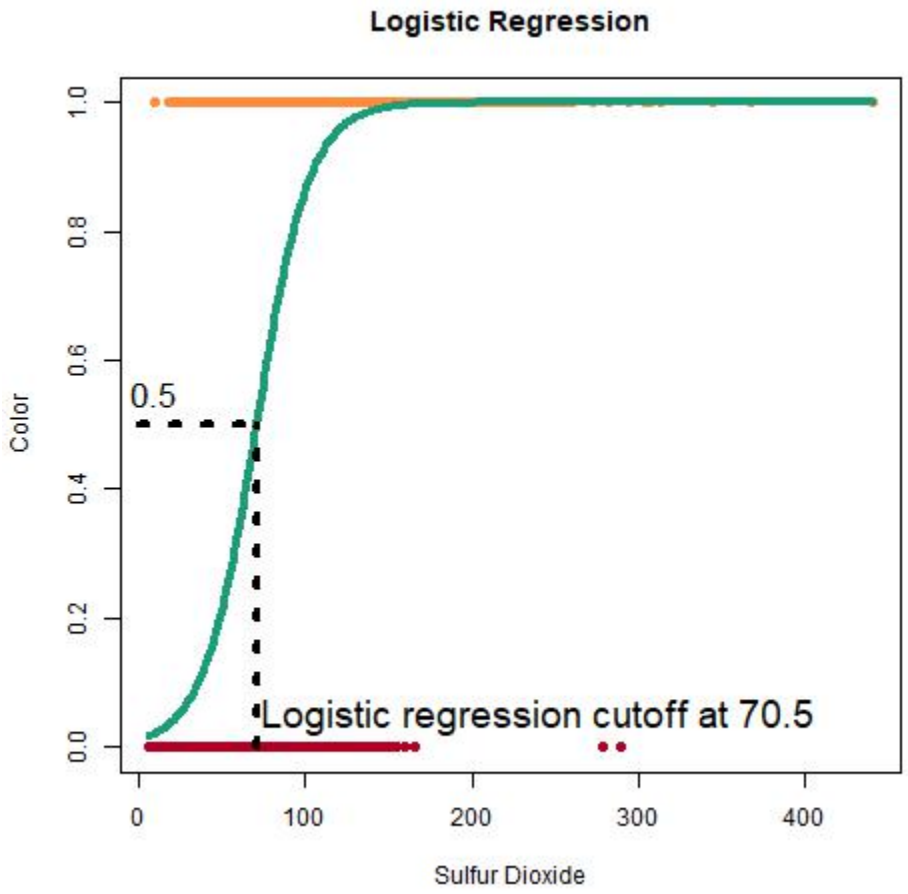
Classification Picture



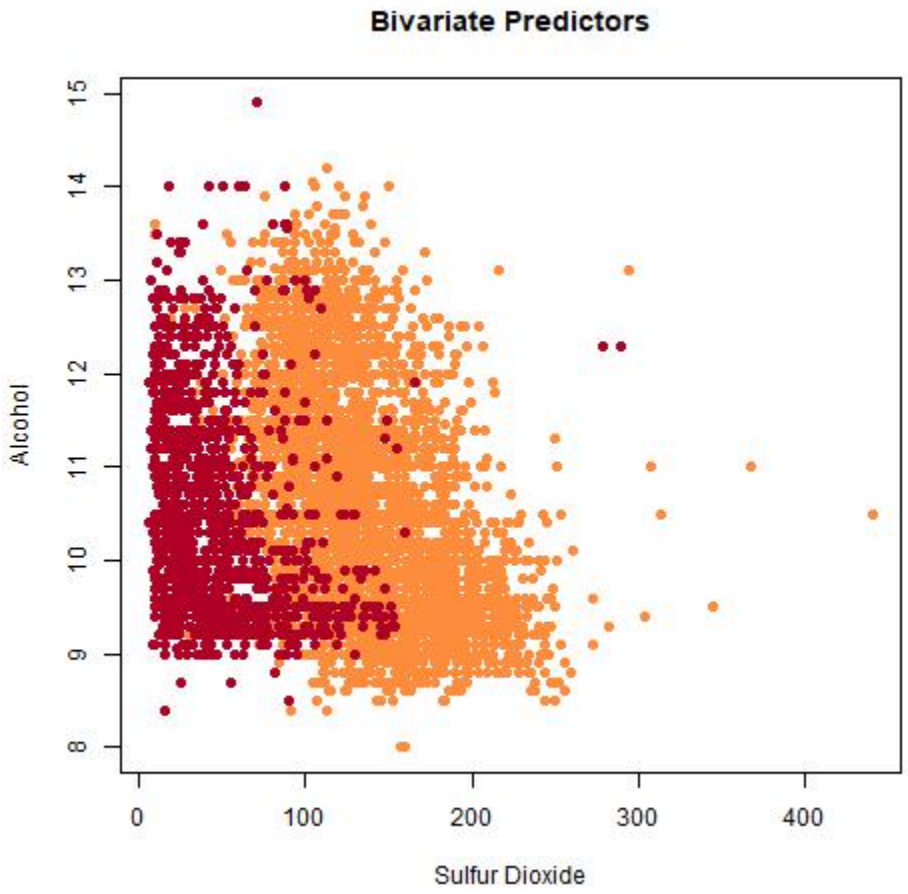
Classification Picture



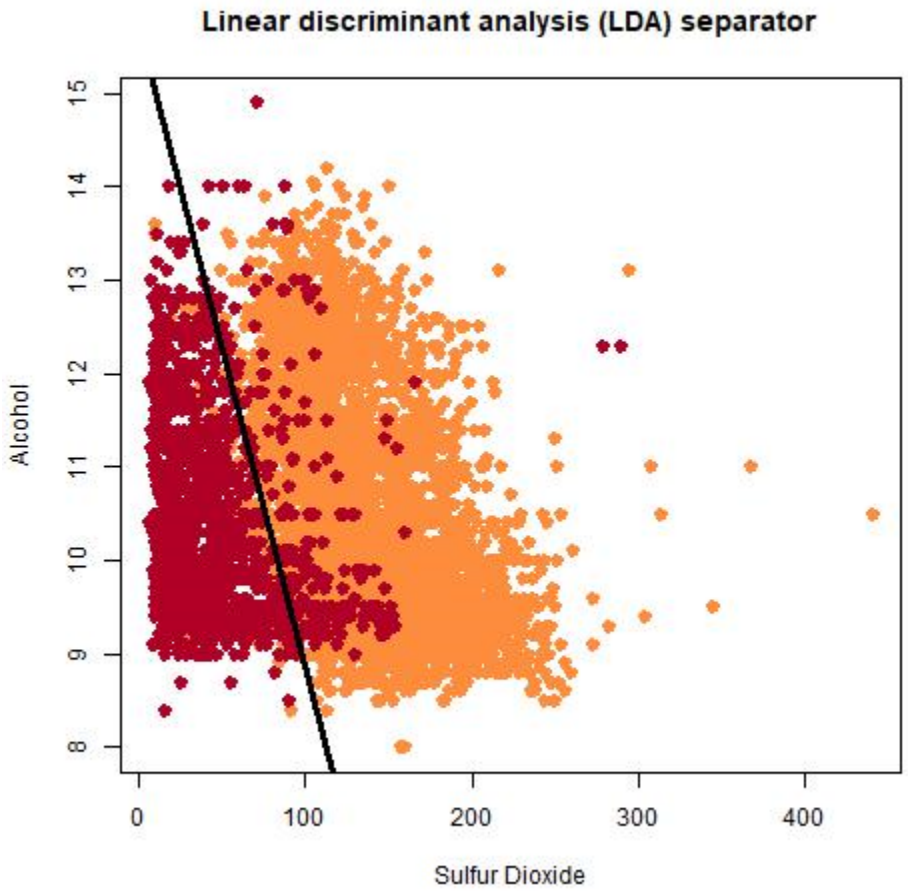
Classification Picture



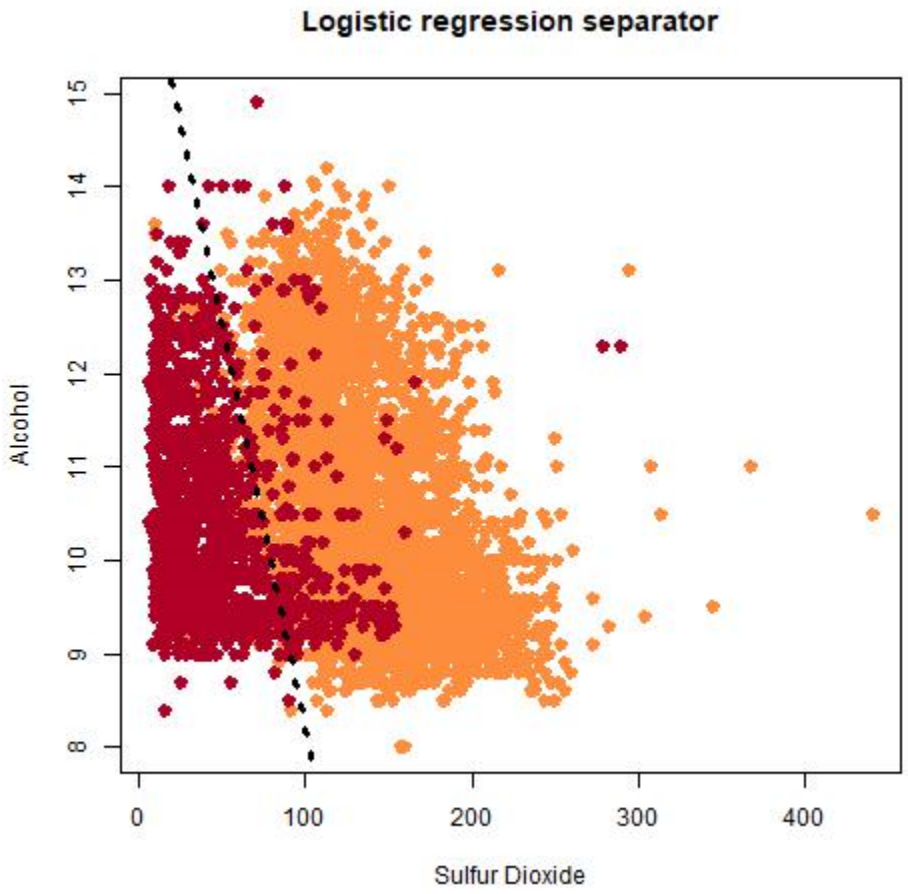
Classification Picture



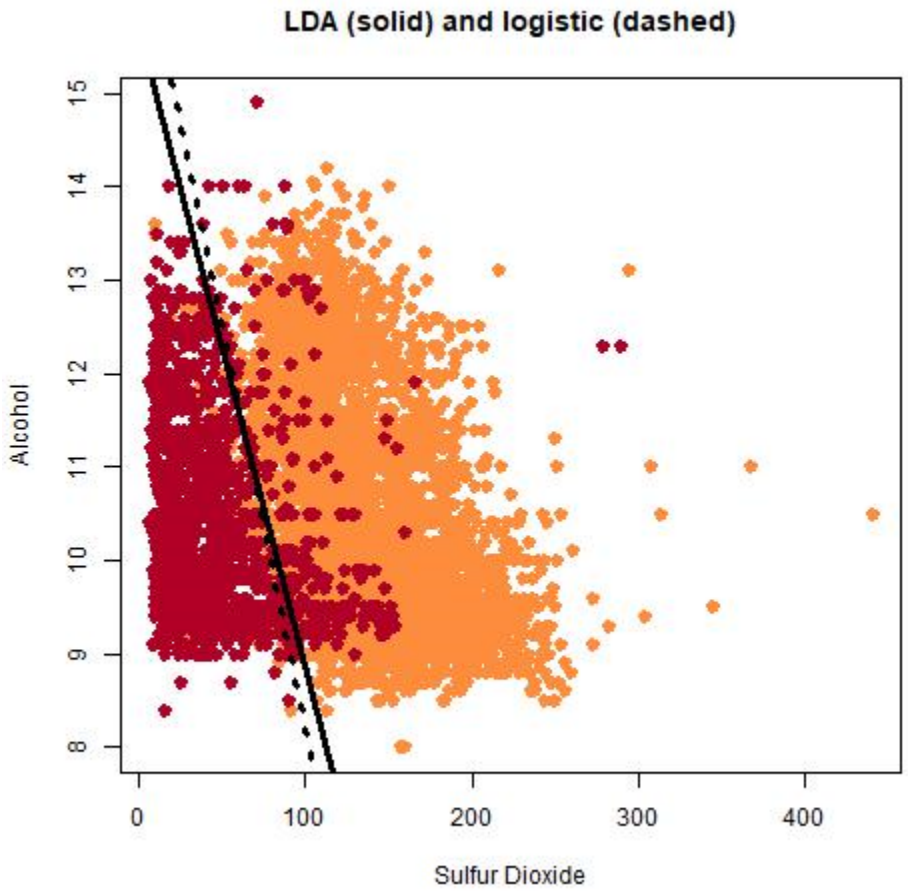
Classification Picture



Classification Picture



Classification Picture



Regression and Classification

Given data $\mathcal{D}$ = { ($\mathbf{x}_i$,y_i), i=1,…,n} where $\mathbf{x}_i$ =(x_{i1},…,x_{ip}), build a model $\hat{f}$ so that

$$\hat{Y} = \hat{f}(\mathbf{X})$$

for random variables $\mathbf{X}$ = (X_1,…,X_p) and Y.

Then $\hat{f}$ will be used for:

- Predicting the value of the response from the predictors: $\hat{y_0} = \hat{f}(\mathbf{x}_0)$ where $\mathbf{x}_0$ = (x_{01},…,x_{0p}).
- Understanding the relationship between the predictors and the response.

Assumptions

- Independent observations
  - Not autocorrelated over time or space
  - Not usually from a designed experiment
  - Not matched case-control
- Goal is prediction and (sometimes) understanding
  - Which predictors are useful? How? Where?
  - Is there any “interesting” structure?

Predictive Accuracy

- Regression
  - Expected mean squared error
- Classification
  - Expected (classwise) error rate

Estimates of Predictive Accuracy

- Resubstitution
  - Use the accuracy on the training set as an estimate of generalization error.
- AIC etc
  - Use assumptions about model.
- Crossvalidation
  - Randomly select a training set, use the rest as the test set.
  - 10-fold crossvalidation.

10-Fold Crossvalidation

Divide the data at random into 10 pieces, $D_1,\ldots,D_{10}$.

- Fit the predictor to $D_2,\ldots,D_{10}$; predict D_1.
- Fit the predictor to $D_1,D_3,\ldots,D_{10}$; predict D_2.
- Fit the predictor to $D_1,D_2,D_4,\ldots,D_{10}$; predict D_3.
- …
- Fit the predictor to $D_1,D_2,\ldots,D_9$; predict D_{10}.

Compute the estimate of predictive accuracy using the assembled predictions and their observed values.

Estimates of Predictive Accuracy

Typically, resubstitution estimates are optimistic compared to crossvalidation estimates.

Crossvalidation estimates tend to be pessimistic because they are based on smaller samples.

Random Forests has its own way of estimating predictive accuracy ("out-of-bag" estimates).

Case Study: Cavity Nesting birds in the Uintah Mountains, Utah

- Red-naped sapsucker (*Sphyrapicus nuchalis*) (*n* = 42 nest sites)

- Mountain chickadee
- (*Parus gambeli*) (*n* = 42 nest sites)

- Northern flicker (*Colaptes auratus*) (*n* = 23 nest sites)

- *n* = 106 non-nest sites

Case Study: Cavity Nesting birds in the Uintah Mountains, Utah

- Response variable is the presence (coded 1) or absence (coded 0) of a nest.
- Predictor variables (measured on 0.04 ha plots around the sites) are:
  - Numbers of trees in various size classes from less than 1 inch in diameter at breast height to greater than 15 inches in diameter.
  - Number of snags and number of downed snags.
  - Percent shrub cover.
  - Number of conifers.
  - Stand Type, coded as 0 for pure aspen and 1 for mixed aspen and conifer.

Assessing Accuracy in Classification

Actual Class	Predicted Class		Total
	Absence	Presence	
	0	1	
Absence, 0	a	**b**	a+b
Presence, 1	**c**	d	c+d
Total	a+c	b+d	n

$$Specificity = 100\% \times \frac{a}{a+b} \qquad Sensitivity = 100\% \times \frac{d}{c+d} \qquad PCC = 100\% \times \frac{a+d}{n}$$

$$\kappa = \frac{(Observed\ agreement) - (Chance\ agreement)}{1 - (Chance\ agreement)}$$

$$Chance\ agreement = \frac{a+b}{n} \times \frac{a+c}{n} + \frac{c+d}{n} \times \frac{b+d}{n} \qquad Observed\ agreement = \frac{a+d}{n}$$

Assessing Accuracy in Classification

Actual Class	**Predicted Class**		**Total**
	Absence	Presence	
	0	1	
Absence, 0	a	**b**	a+b
Presence, 1	**c**	d	c+d
Total	a+c	b+d	n

Error rate = (c + b) / n

For class 1: b/(a+b)

For class 2: c/(c+d)

Resubstitution Accuracy (fully grown tree)

Actual Class	Predicted Class		Total
	Absence	Presence	
	0	1	
Absence, 0	105	**1**	106
Presence, 1	**0**	107	107
Total	105	108	213

Error rate = (0 + 1)/213 = (approx) 0.005 or **0.5%**

Crossvalidation Accuracy (fully grown tree)

Actual Class	Predicted Class		Total
	Absence	Presence	
	0	1	
Absence, 0	83	**23**	106
Presence, 1	**22**	85	107
Total	105	108	213

Error rate = (22 + 23)/213 = (approx) .21 or **21%**

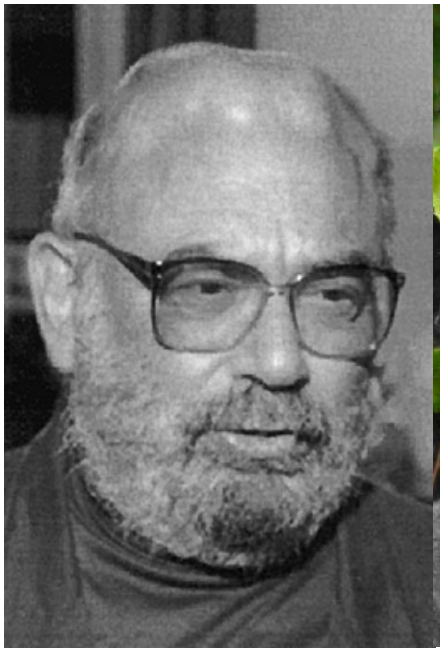
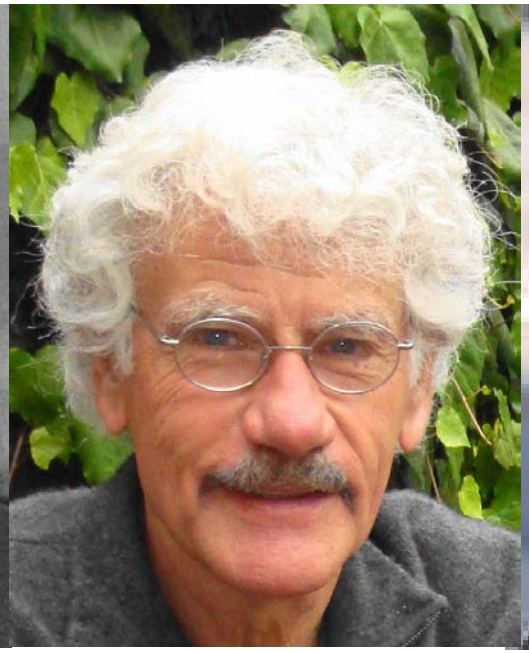
Outline

- Background.
- Trees.
- Bagging predictors.
- Random Forests algorithm.
- Variable importance.
- Proximity measures.
- Visualization.
- Partial plots and interpretation of effects.

Classification and Regression Trees

Pioneers:

- Morgan and Sonquist (1963).
- **Breiman, Friedman, Olshen, Stone (1984). *CART***
- Quinlan (1993). *C4.5*

Classification and Regression Trees

- Grow a binary tree.
- At each node, “split” the data into two “daughter” nodes.
- Splits are chosen using a splitting criterion.
- Bottom nodes are “terminal” nodes.

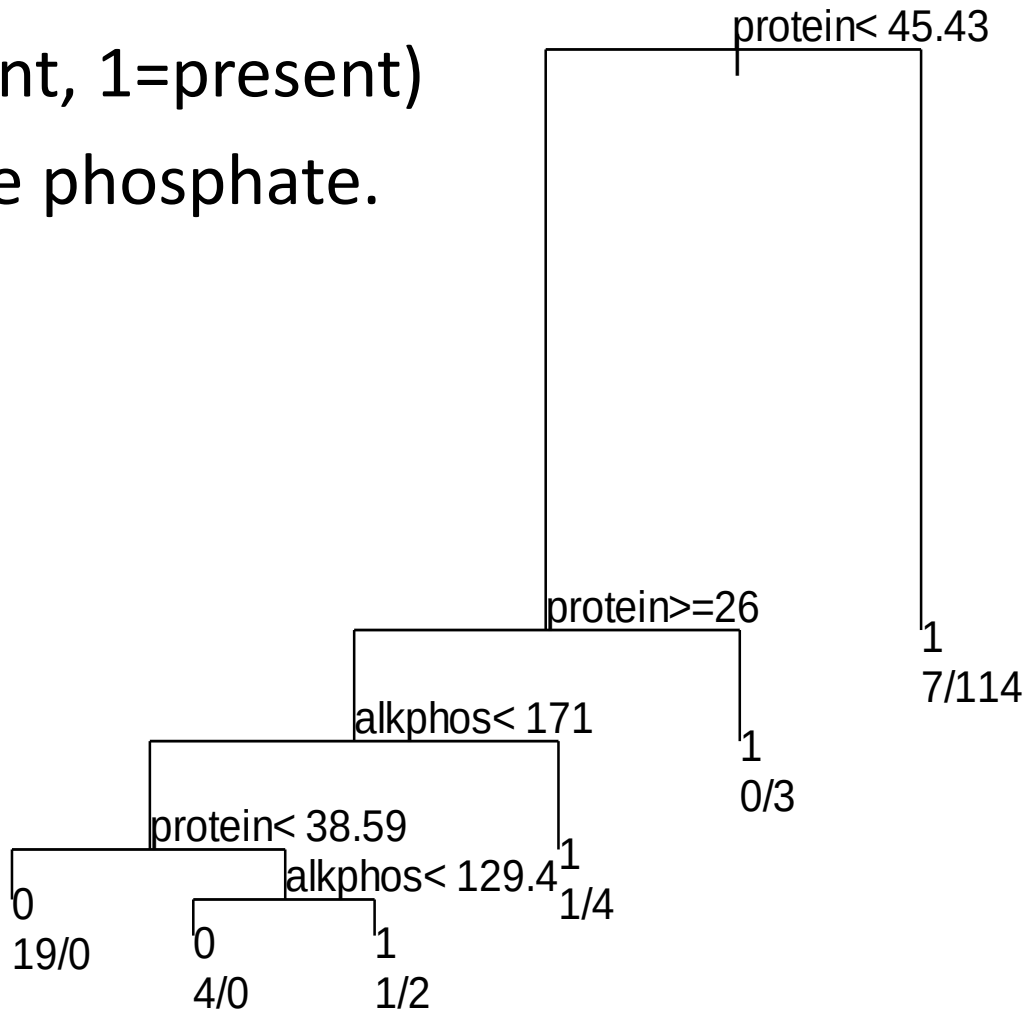
Classification and Regression Trees

- For regression the predicted value at a node is the *average* response variable for all observations in the node.
- For classification the predicted class is the *most common class* in the node (majority vote).
- For classification trees, can also get estimated probability of membership in each of the classes

A Classification Tree

Predict hepatitis (0=absent, 1=present) using protein and alkaline phosphate.

"Yes" goes left.



Splits are chosen to minimize a splitting criterion:

- **Regression**: residual sum of squares

$$RSS = \sum_{left} (y_i - y_L^*)^2 + \sum_{right} (y_i - y_R^*)^2$$

where y_L^* = mean y-value for left node

y_R^* = mean y-value for right node

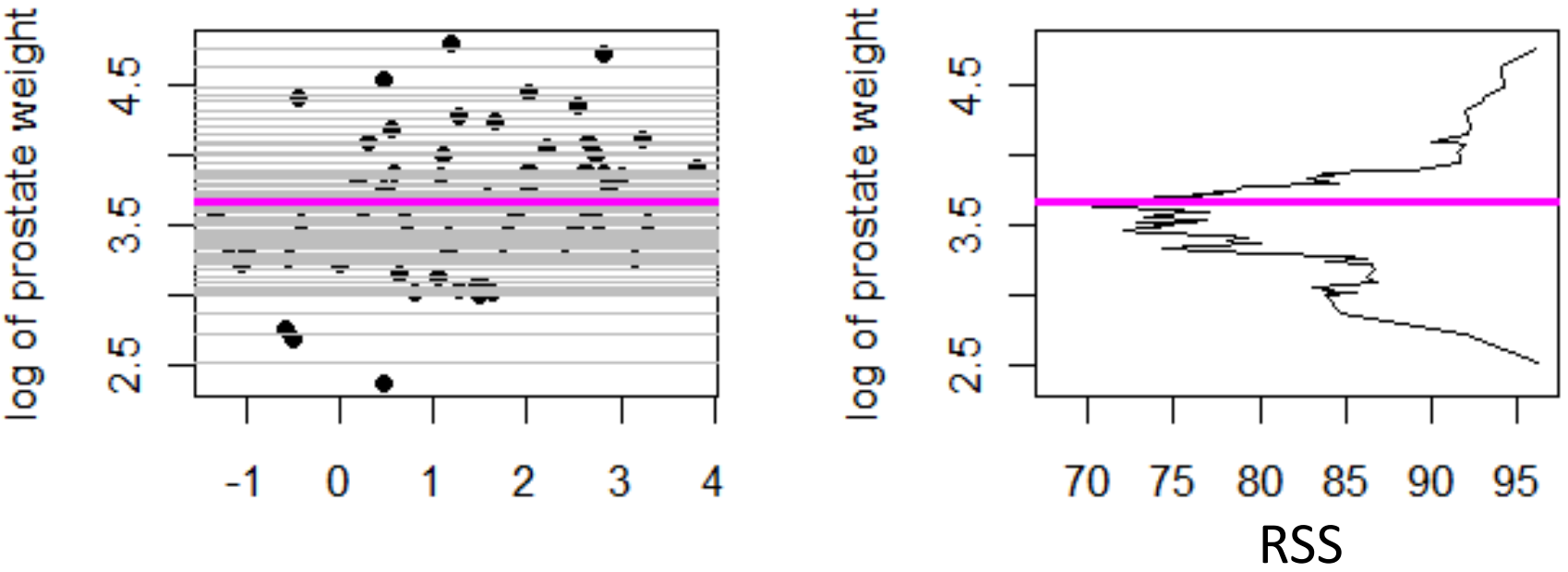
- **Classification**: Gini criterion

$$Gini = N_L \sum_{k=1,\ldots,K} p_{kL} (1- p_{kL}) + N_R \sum_{k=1,\ldots,K} p_{kR} (1- p_{kR})$$

where p_{kL} = proportion of class k in left node

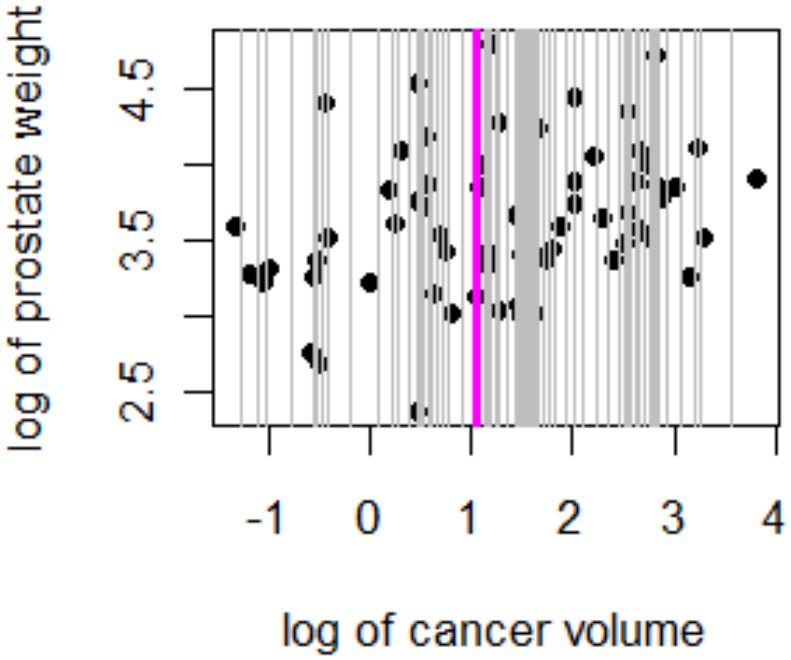
p_{kR} = proportion of class k in right node

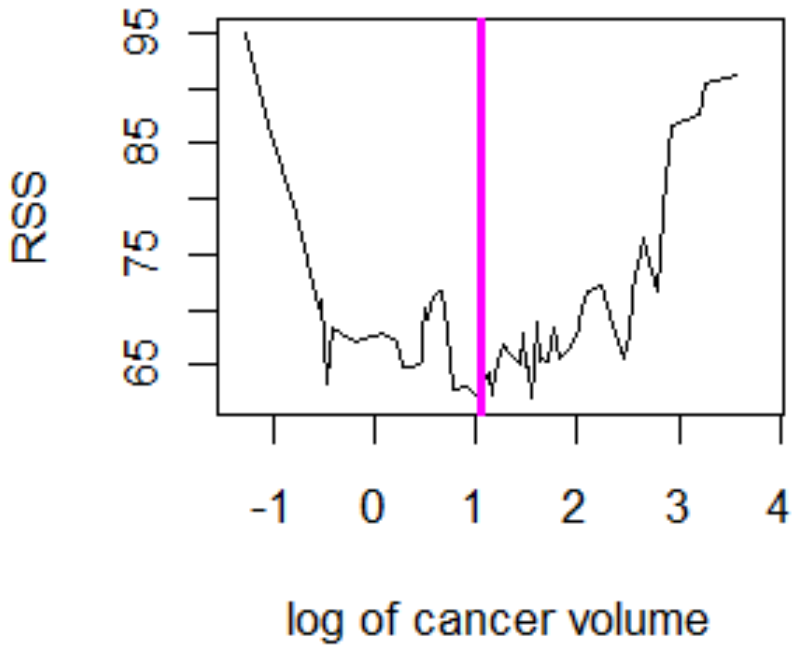
Choosing the best horizontal split



Best horizontal split is at 3.67 with RSS = 68.09.

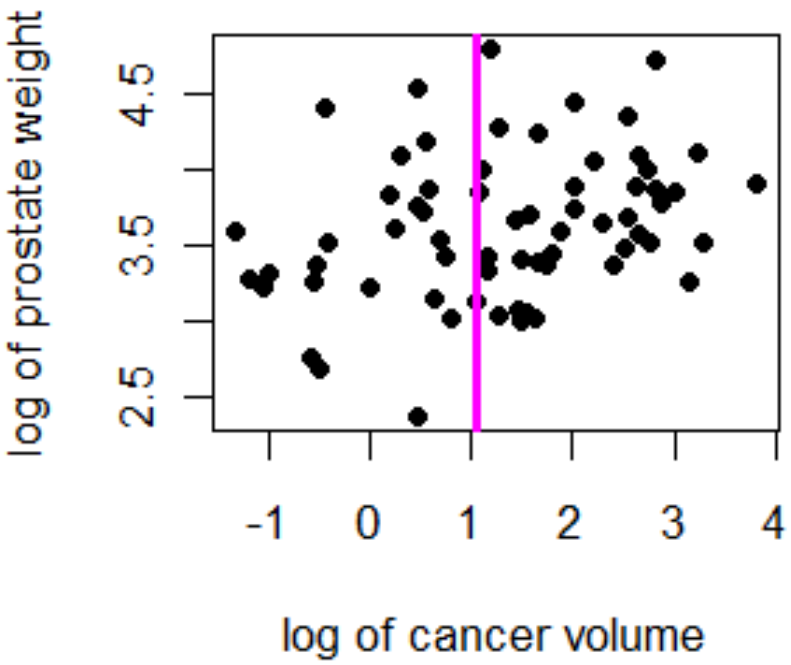
Choosing the best vertical split

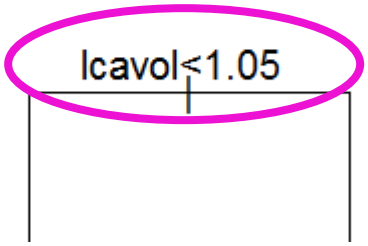




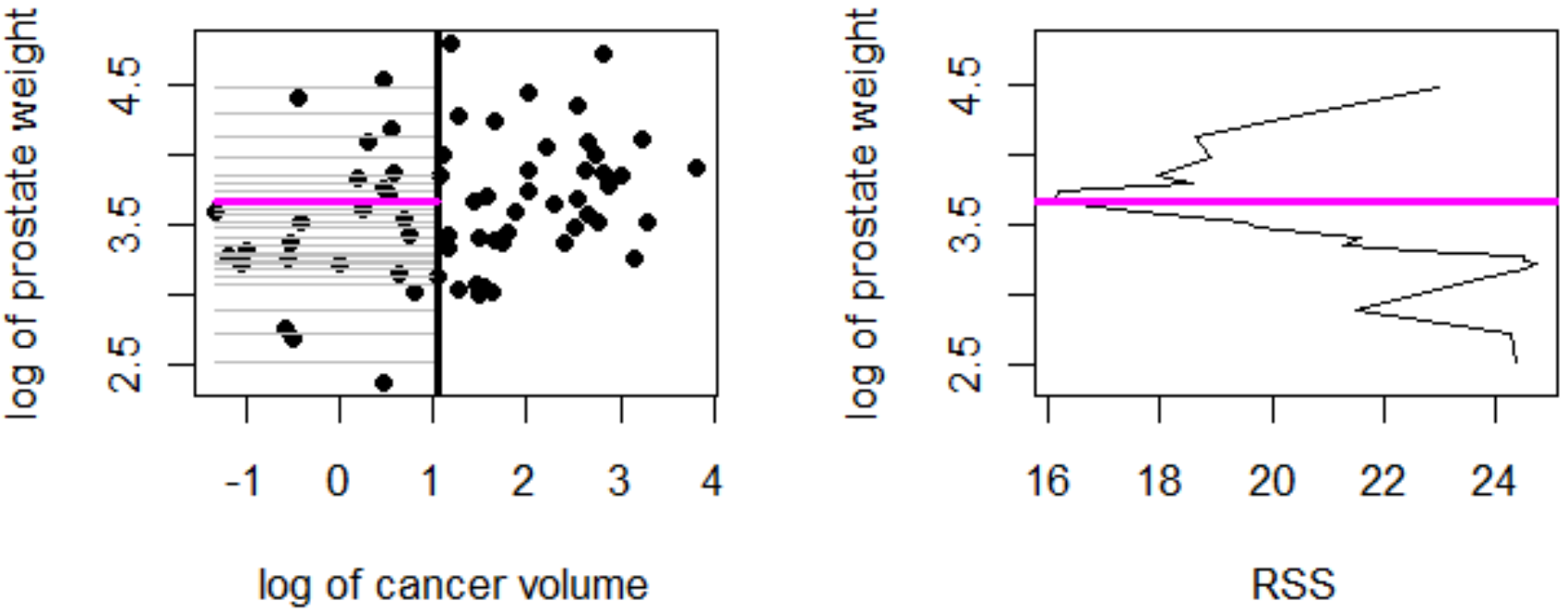
Best vertical split is at 1.05 with RSS = 61.76.

Regression tree (prostate cancer)



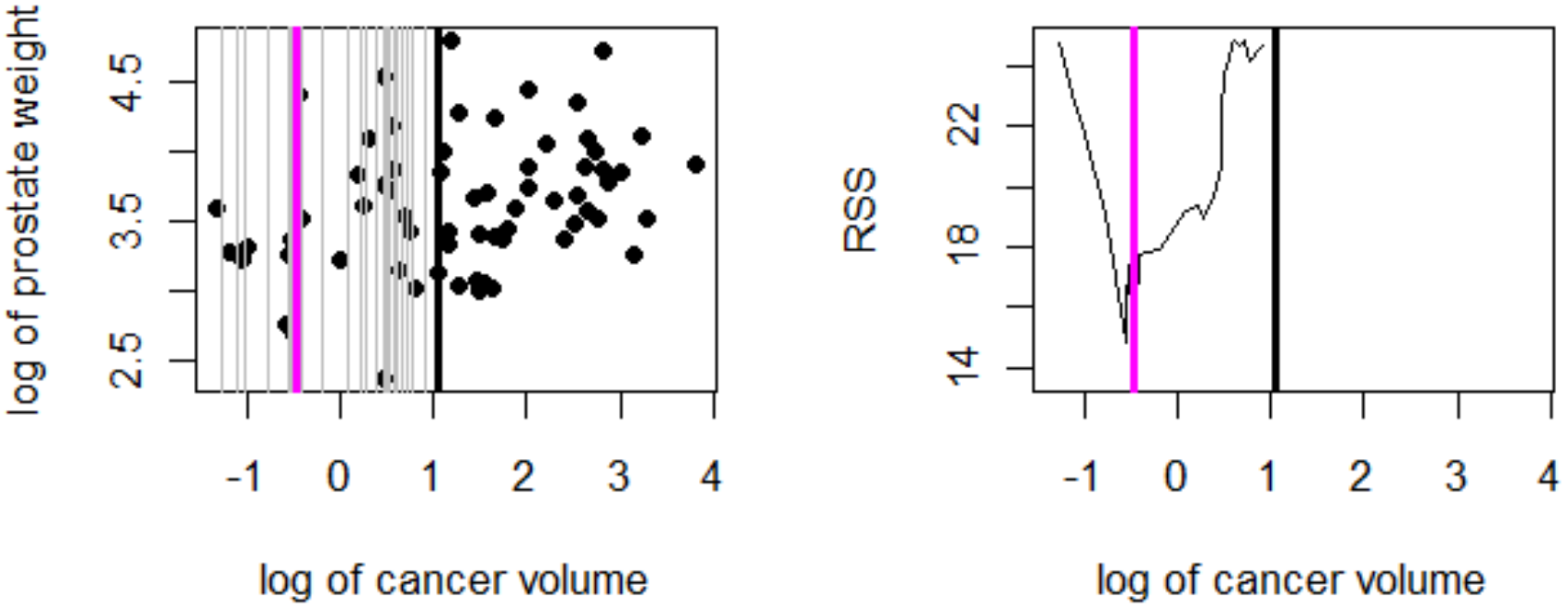


Choosing the best split in the left node



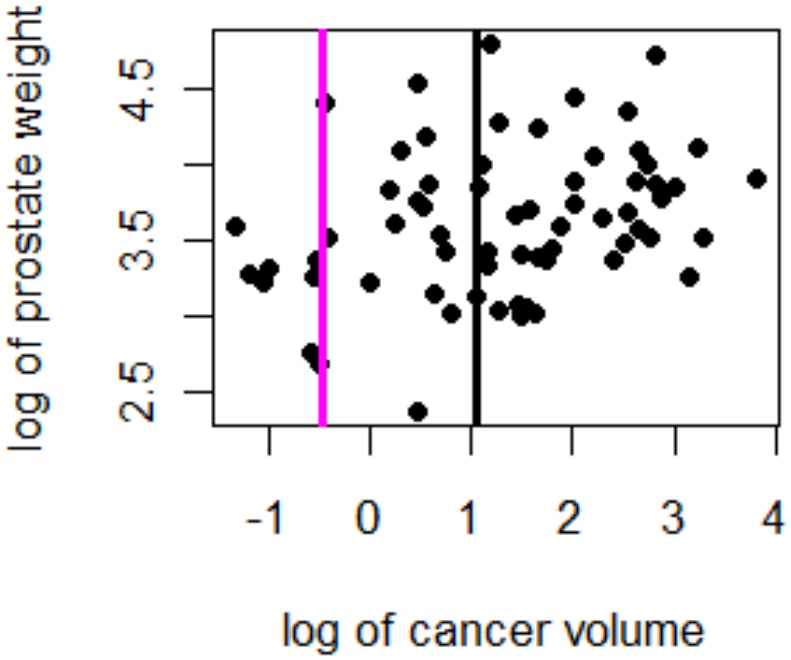
Best horizontal split is at 3.66 with RSS = 16.11.

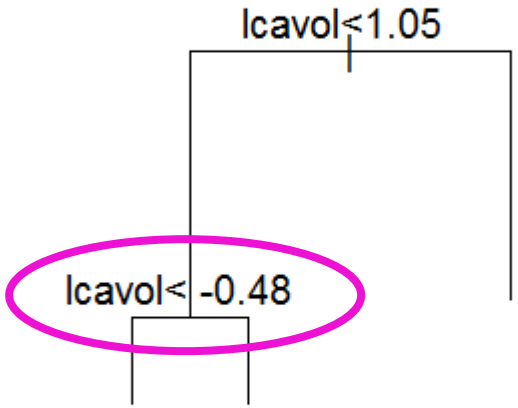
Choosing the best split in the left node



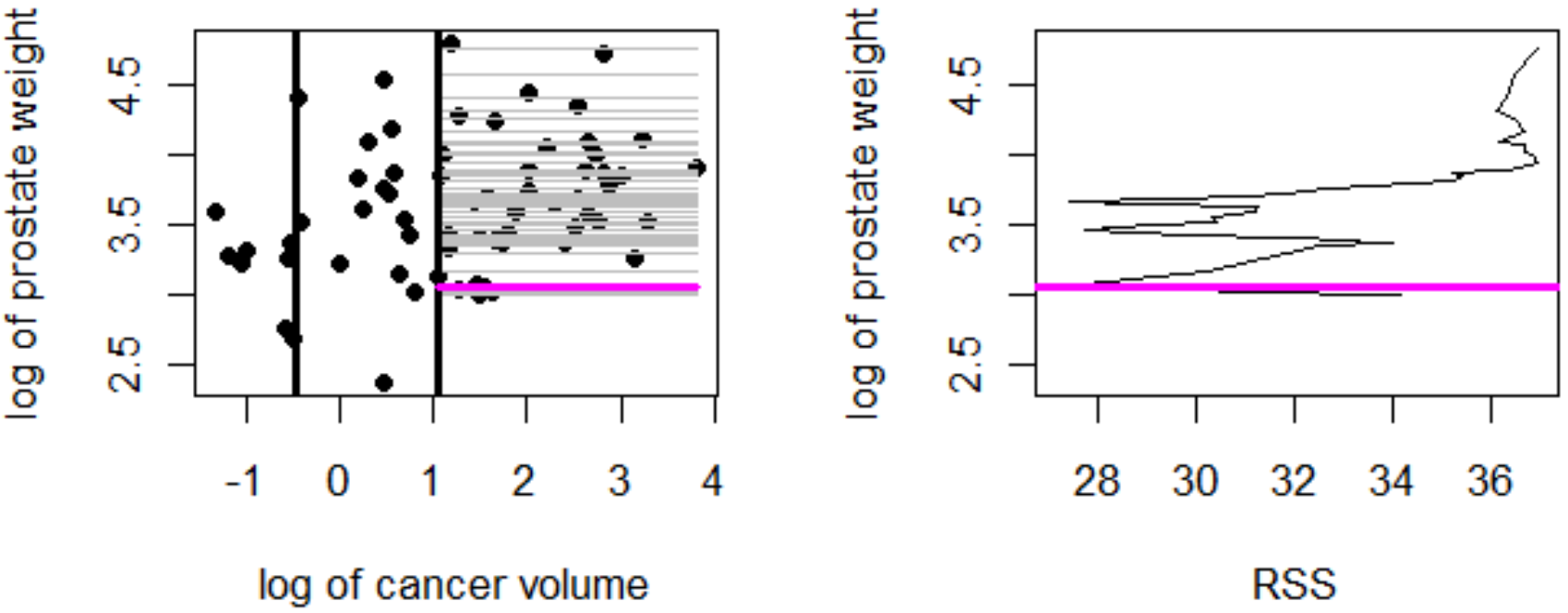
Best vertical split is at -.48 with RSS = 13.61.

Regression tree (prostate cancer)



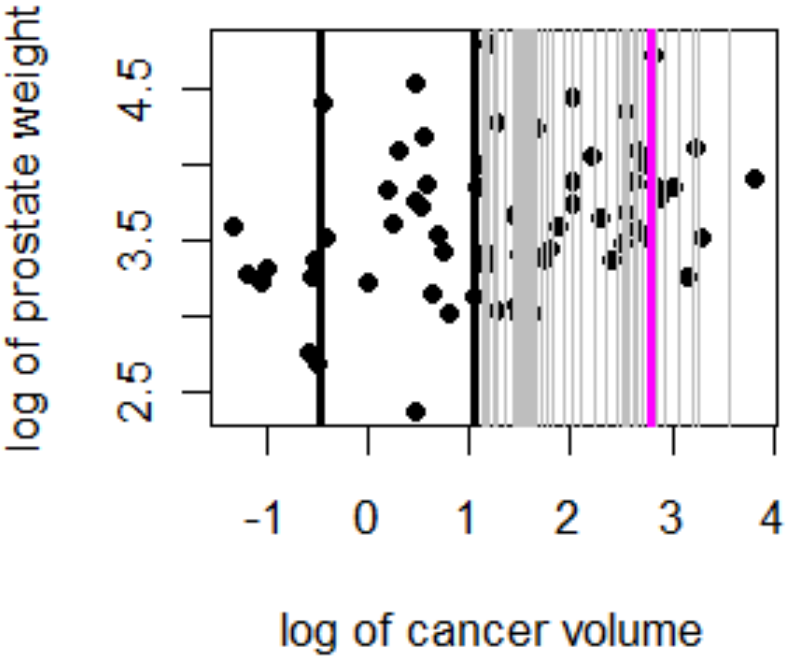


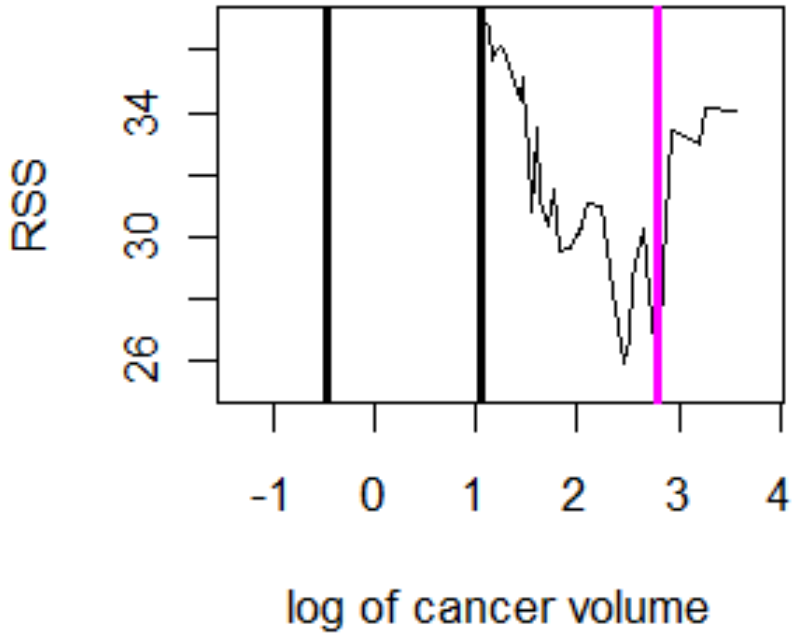
Choosing the best split in the right node



Best horizontal split is at 3.07 with RSS = 27.15.

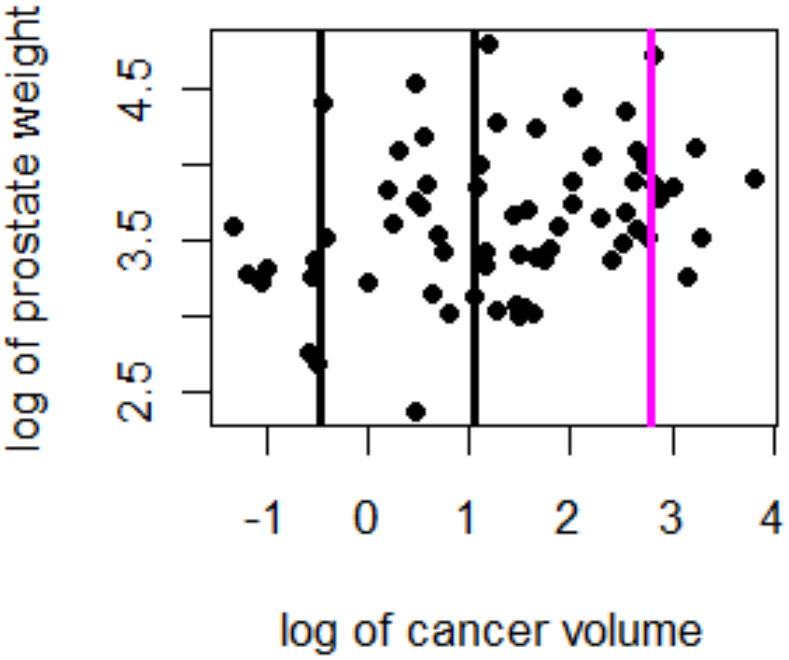
Choosing the best split in the right node

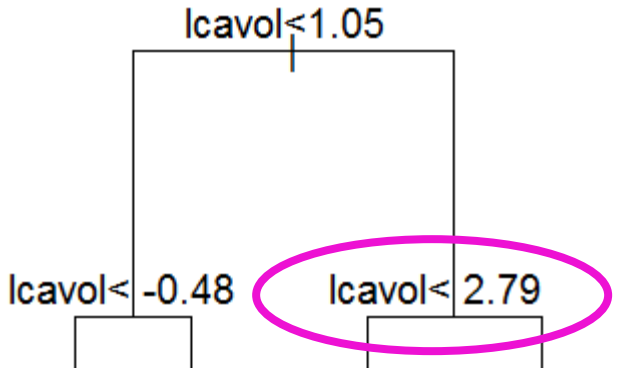




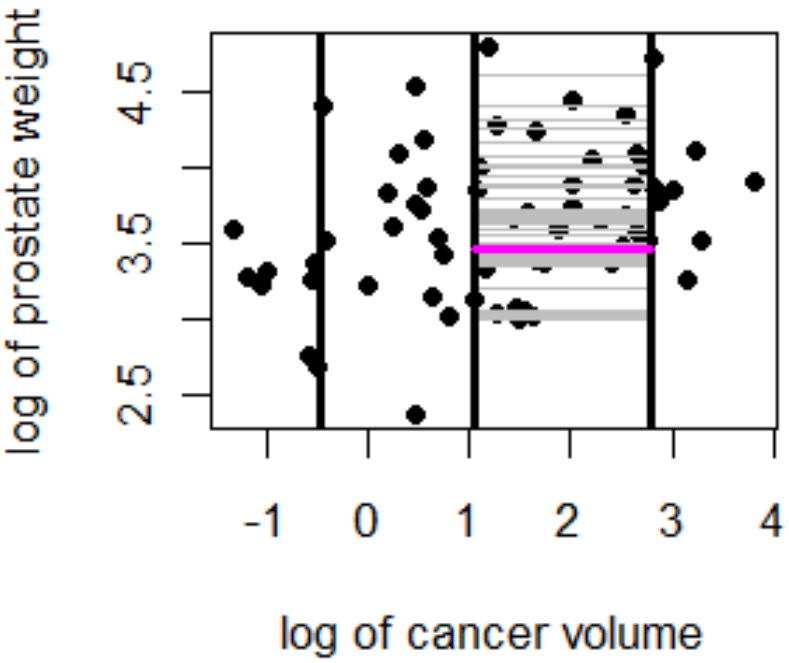
Best vertical split is at 2.79 with RSS = 25.11.

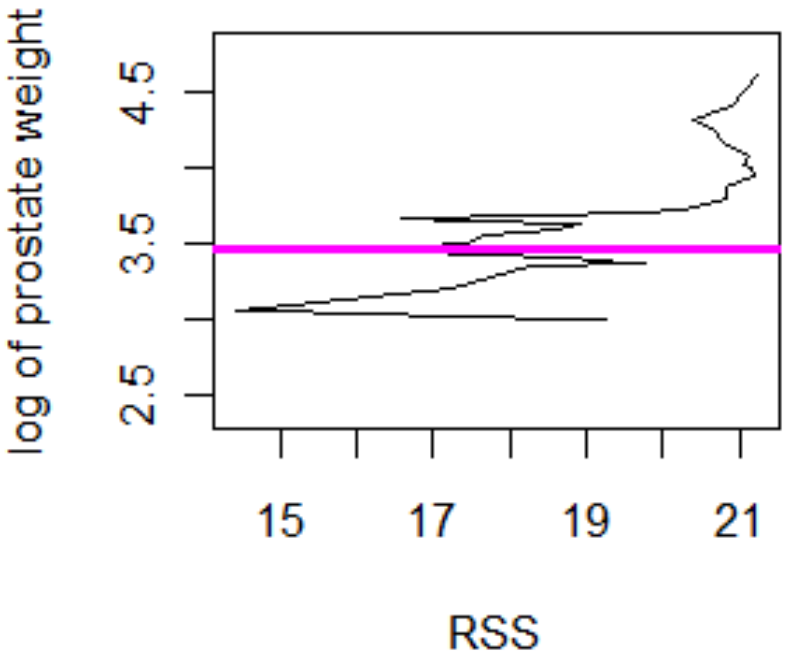
Regression tree (prostate cancer)





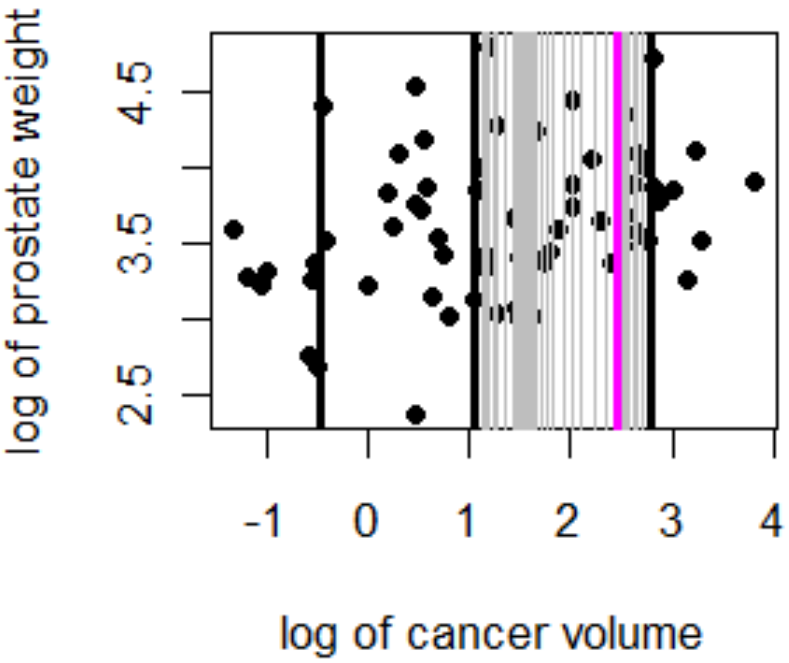
Choosing the best split in the third node

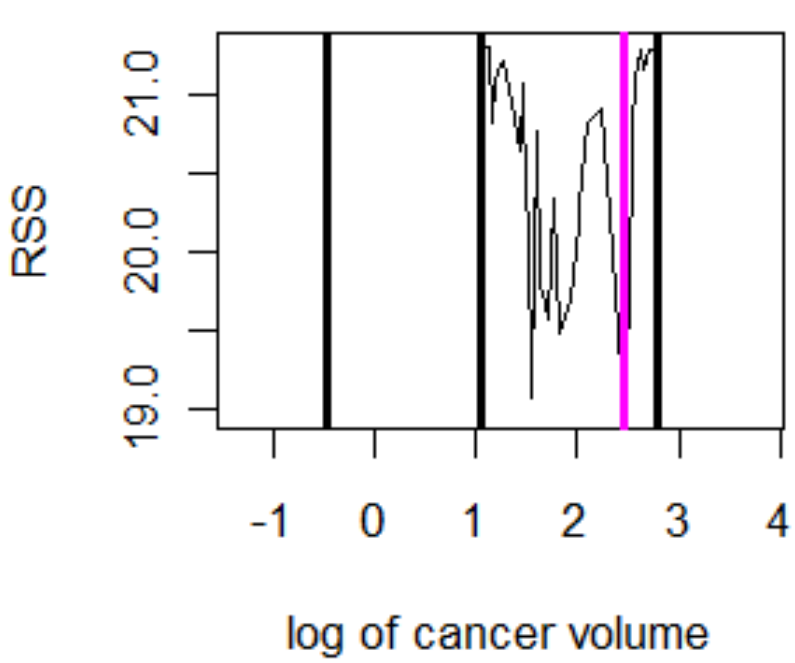




Best horizontal split is at 3.07 with RSS = 14.42, but this is too close to the edge. Use 3.46 with RSS = 16.14.

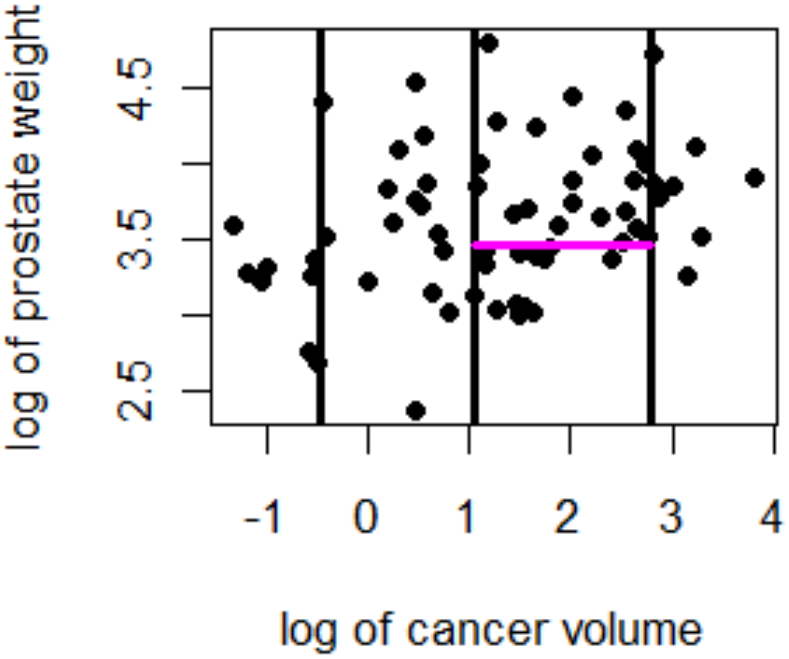
Choosing the best split in the third node

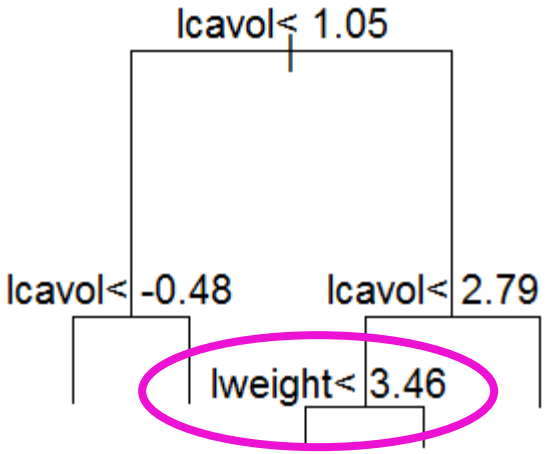




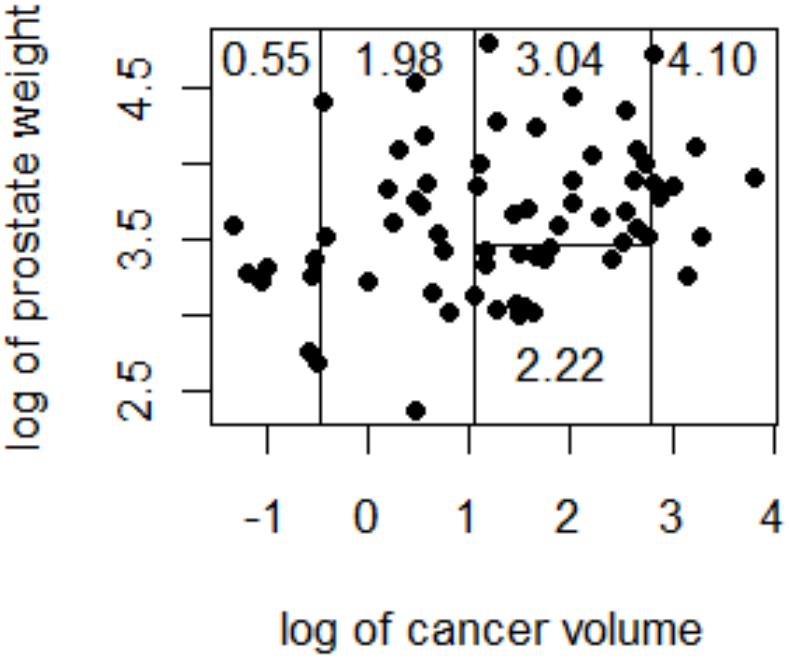
Best vertical split is at 2.46 with RSS = 18.97.

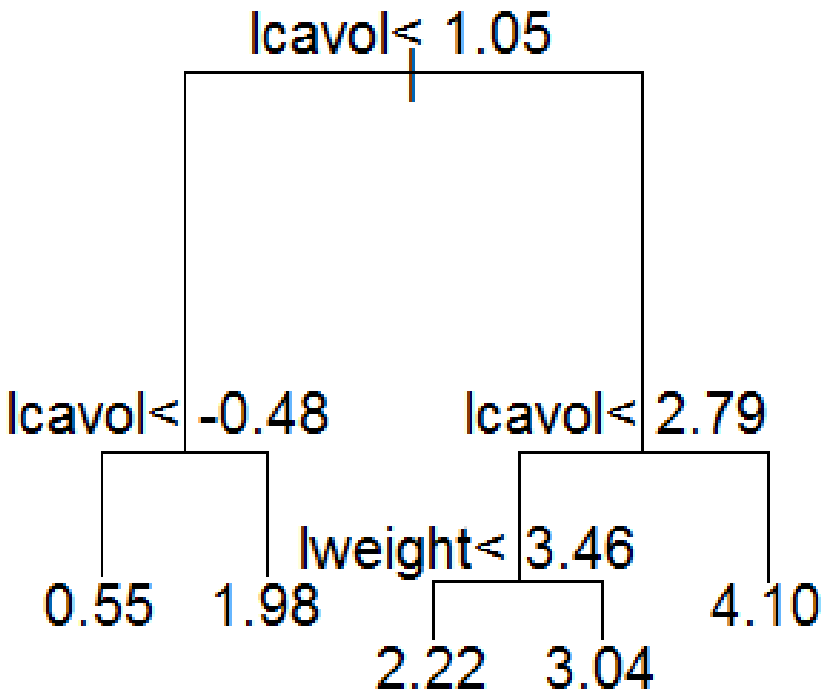
Regression tree (prostate cancer)



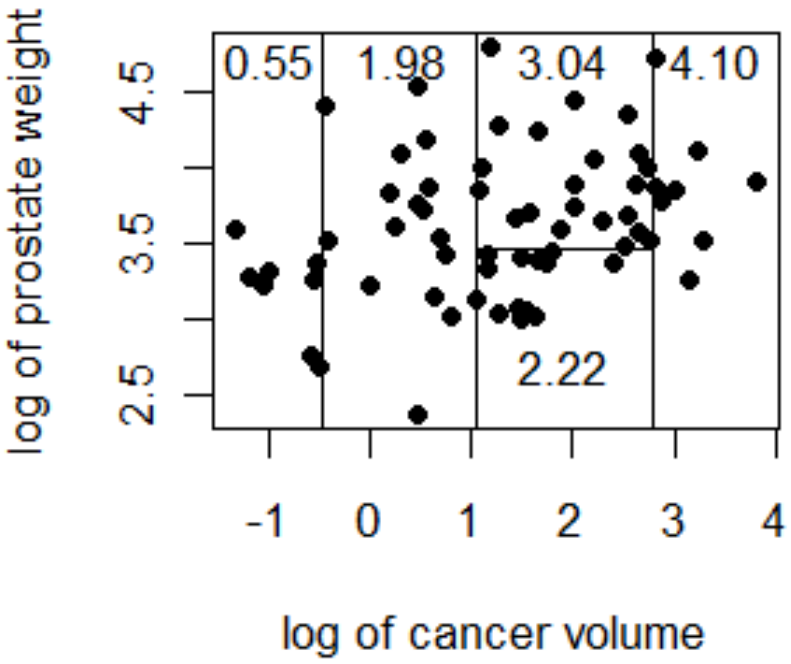


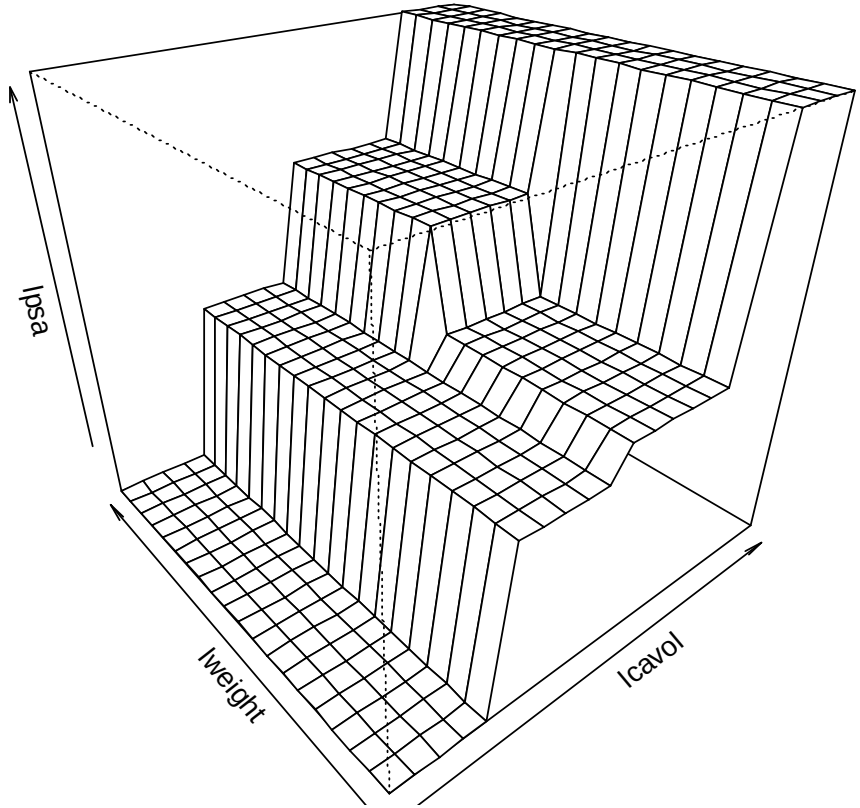
Regression tree (prostate cancer)



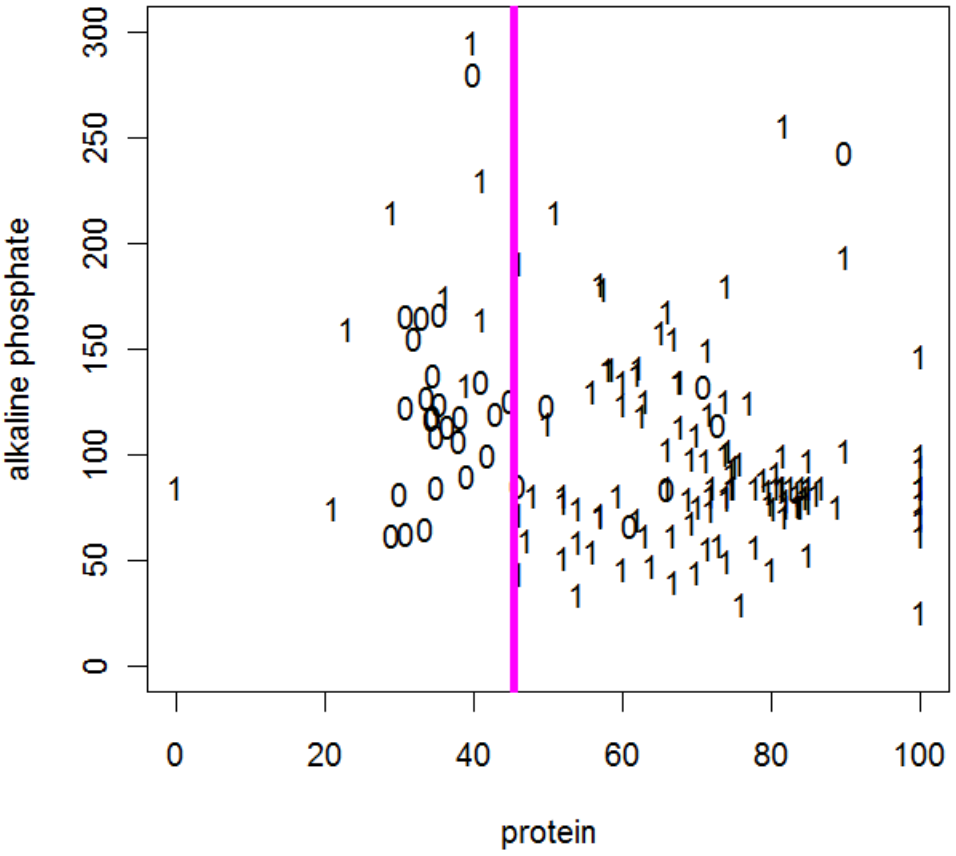


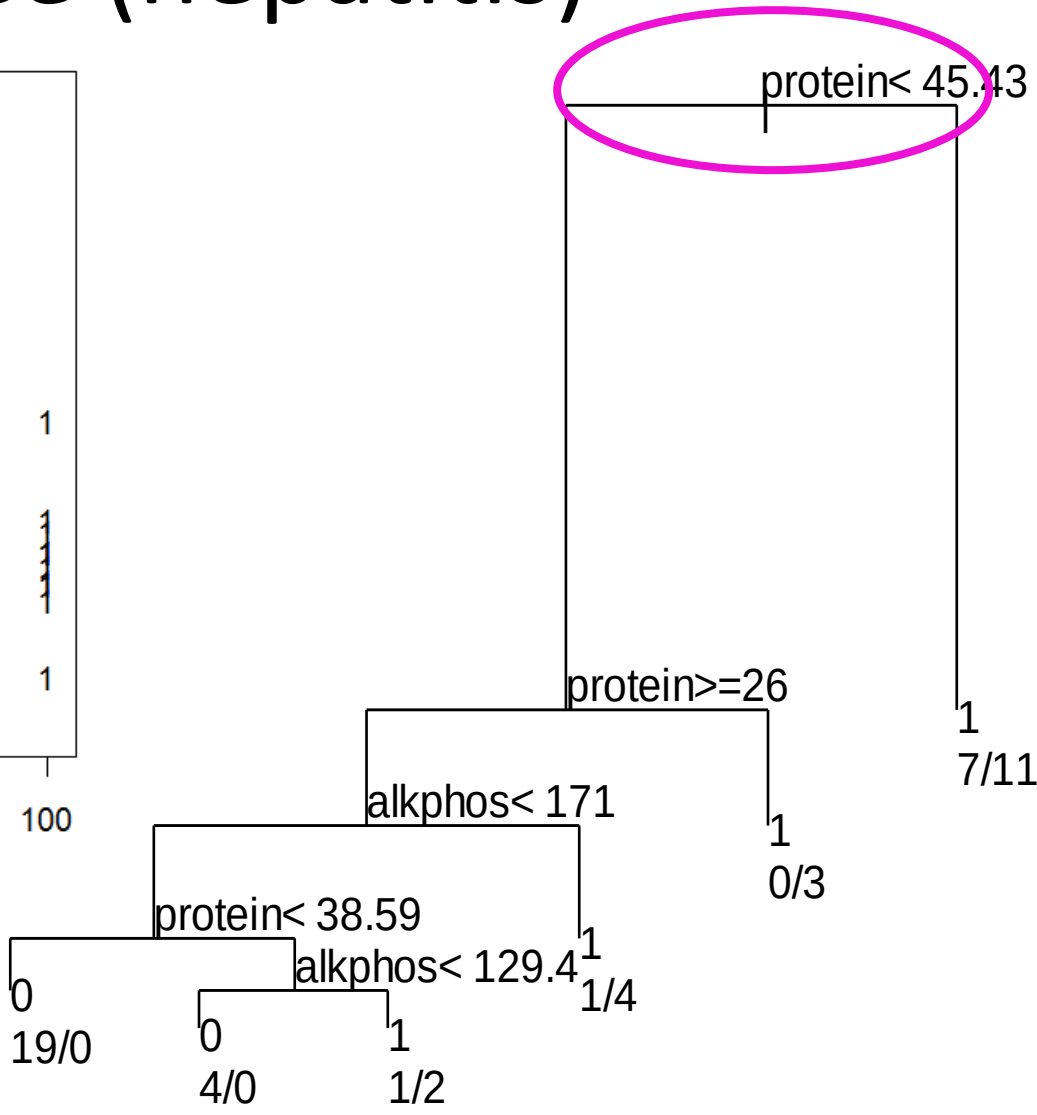
Regression tree (prostate cancer)



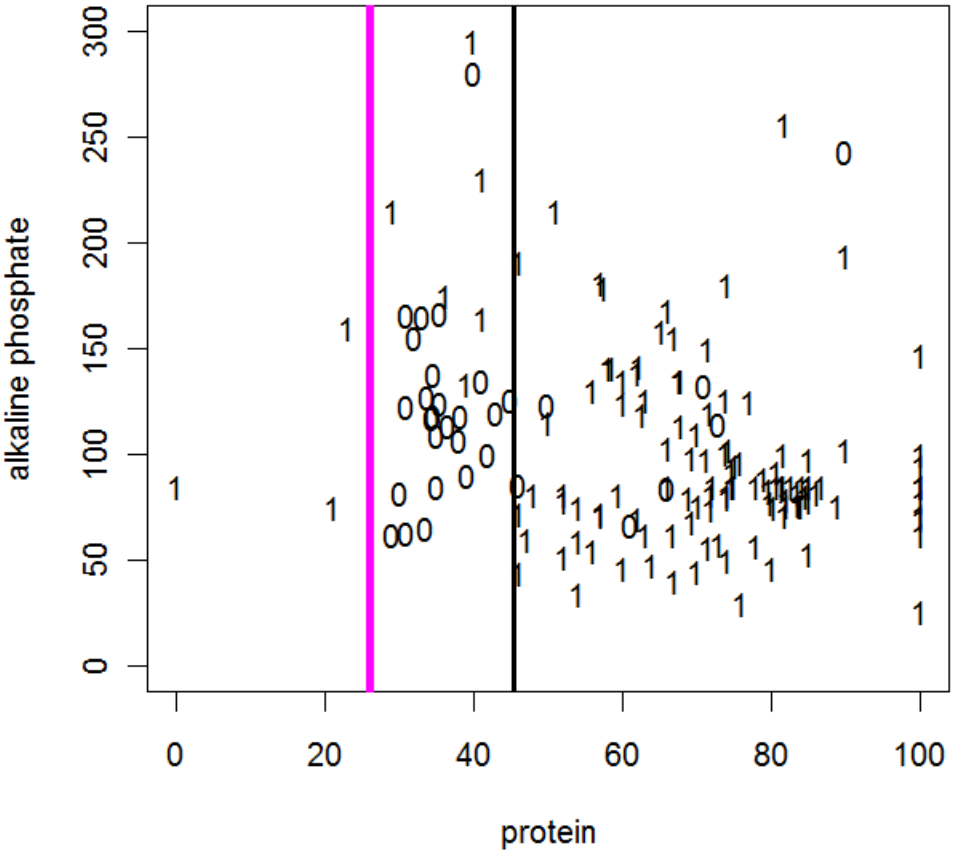


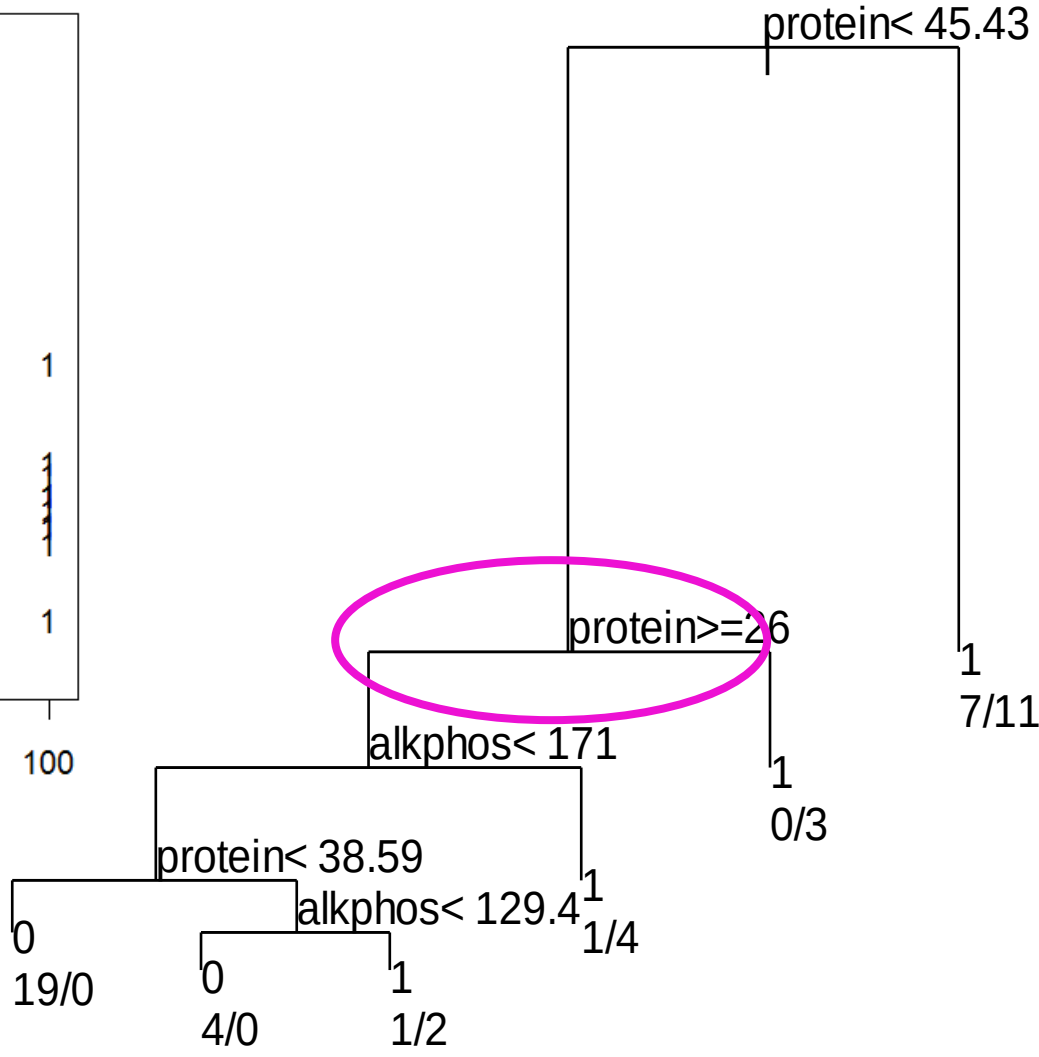
Classification tree (hepatitis)



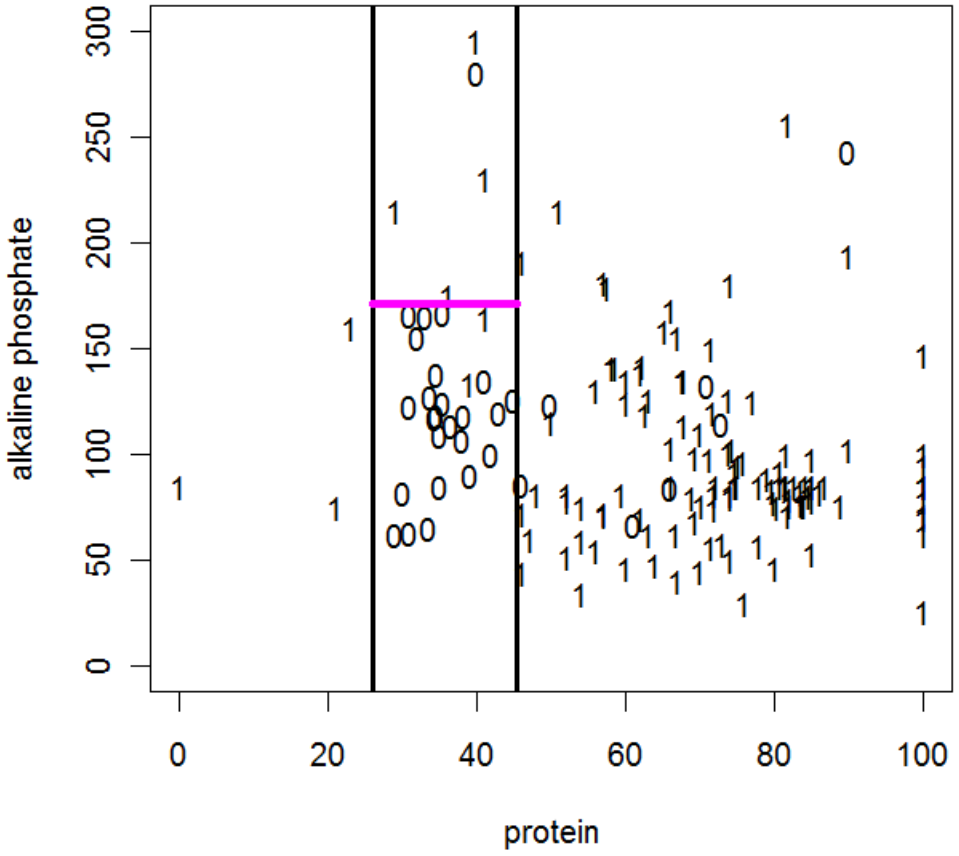


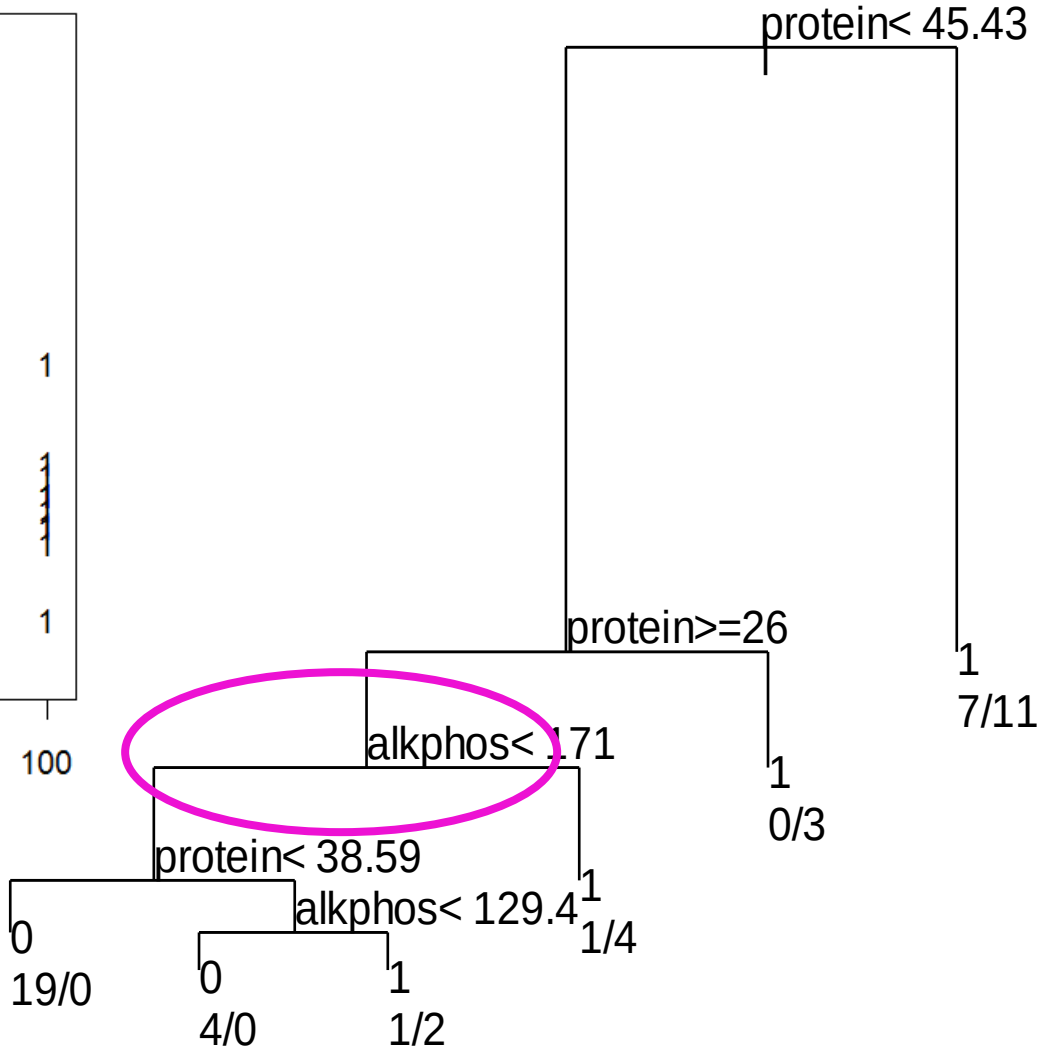
Classification tree (hepatitis)



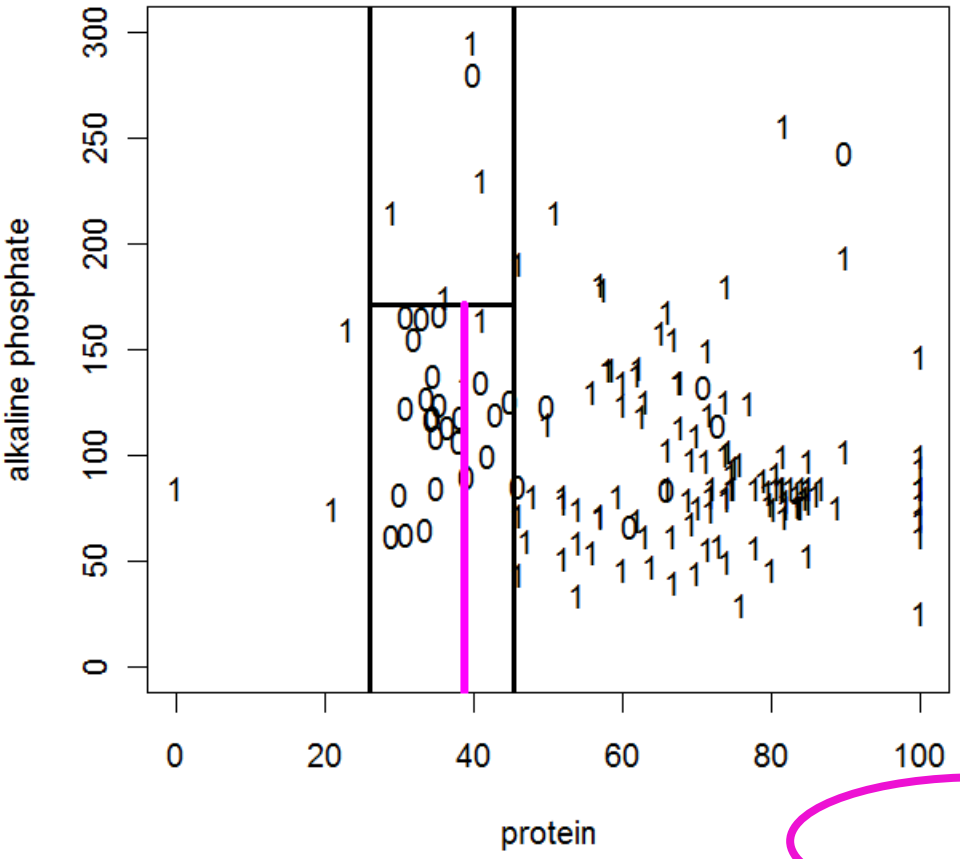


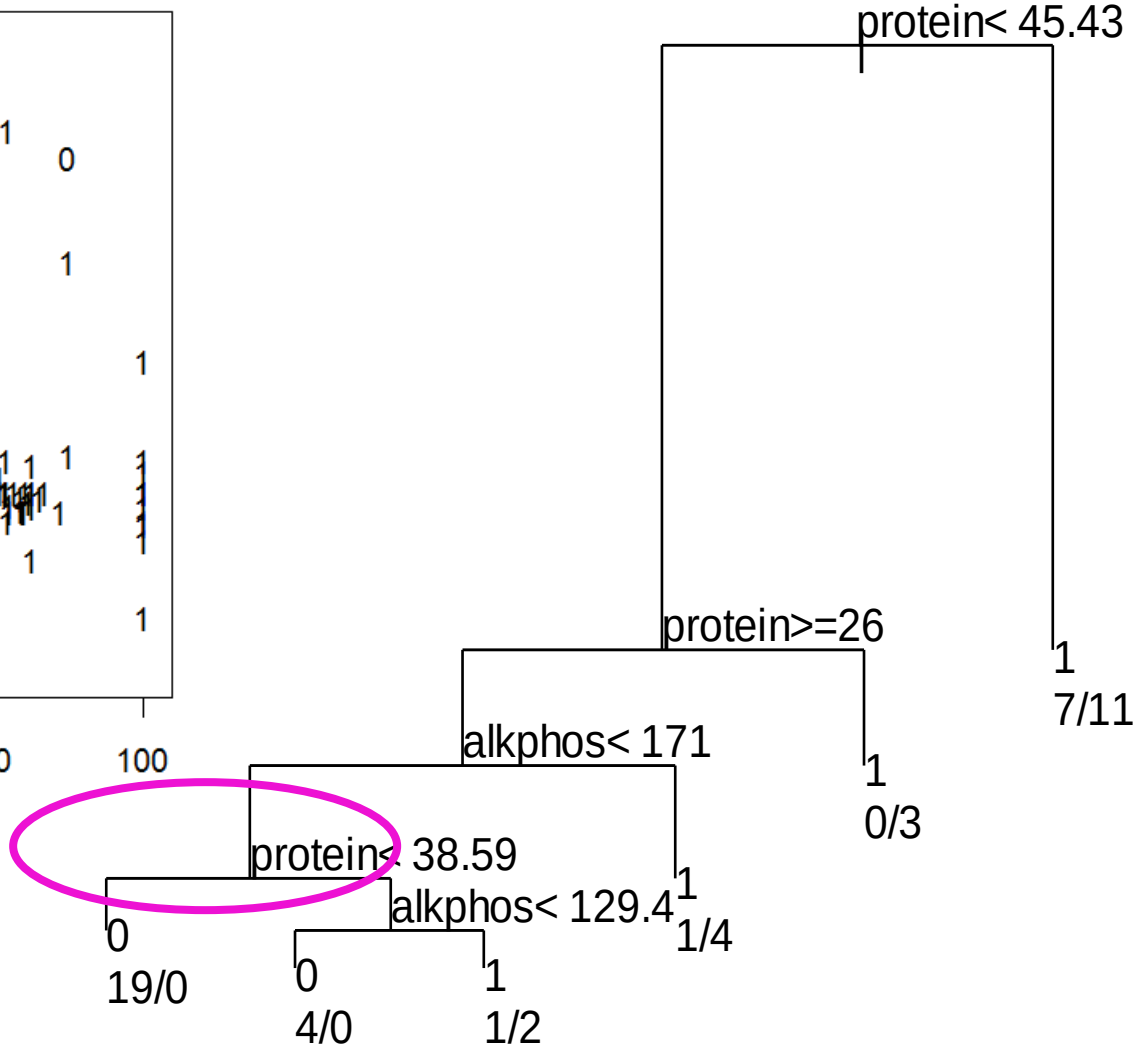
Classification tree (hepatitis)



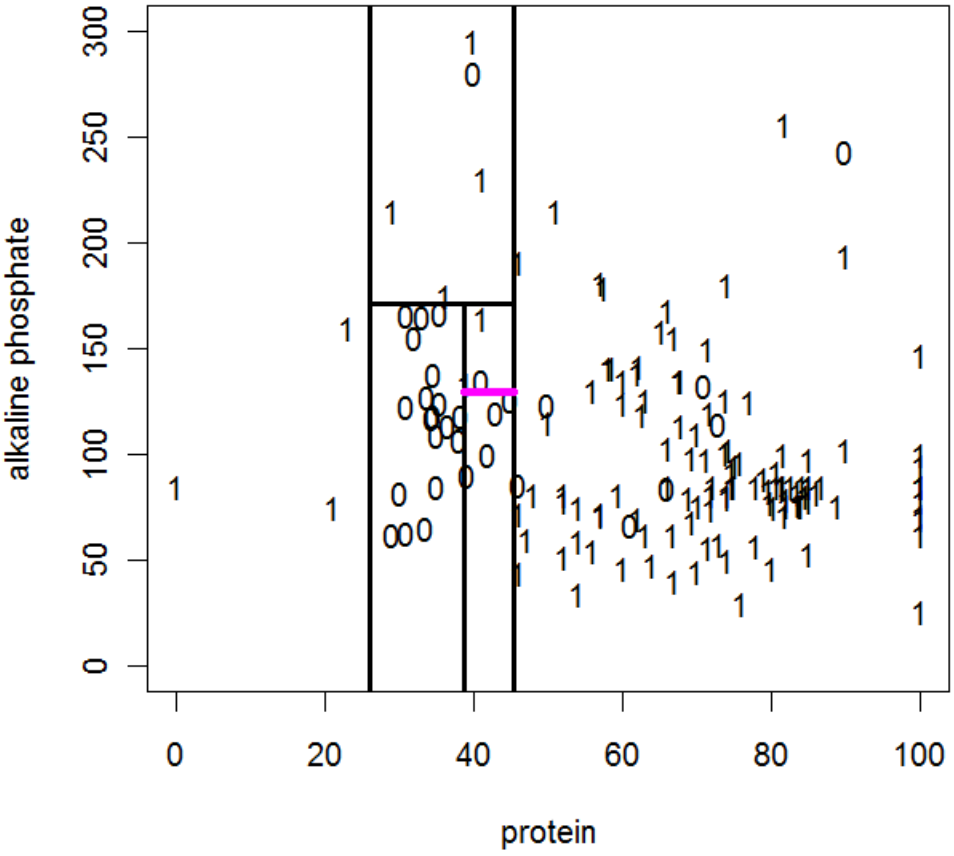


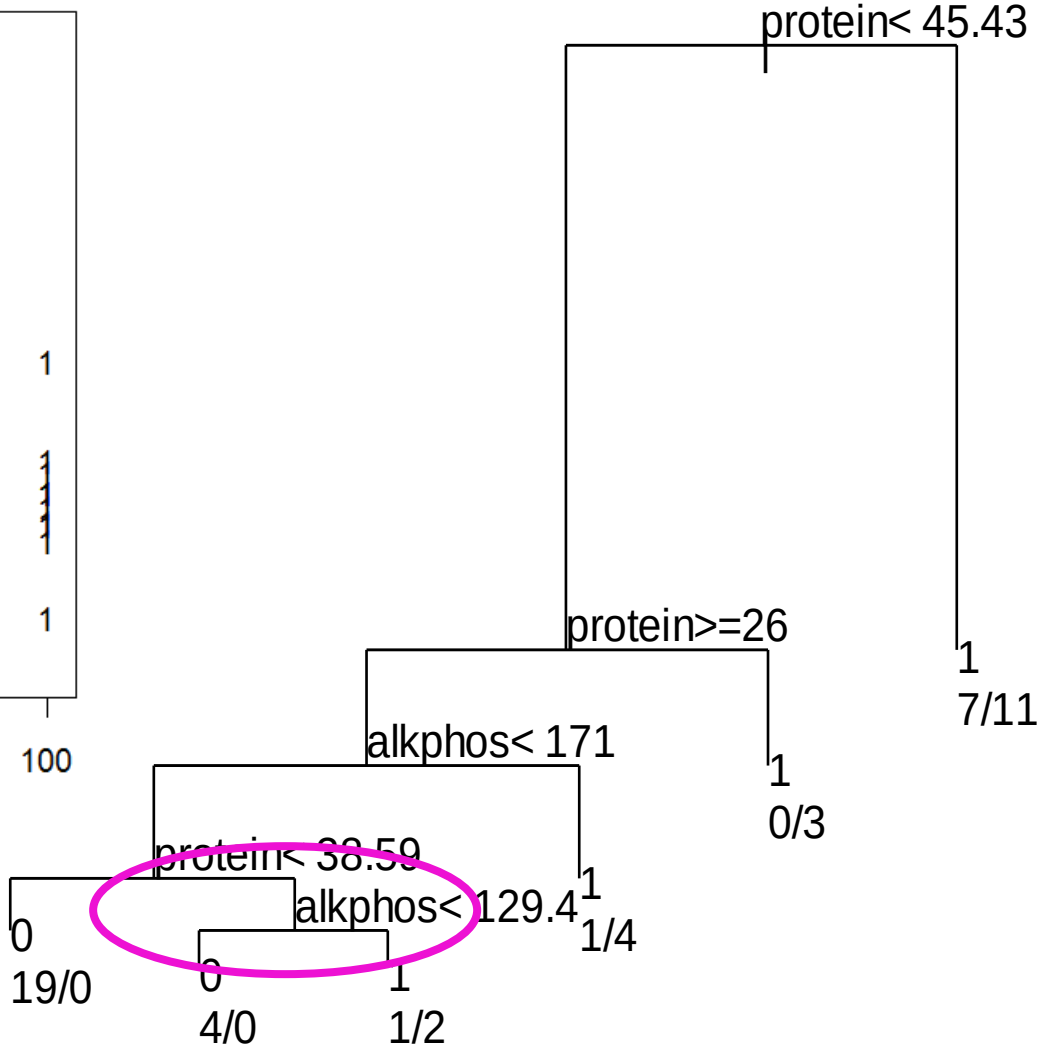
Classification tree (hepatitis)





Classification tree (hepatitis)

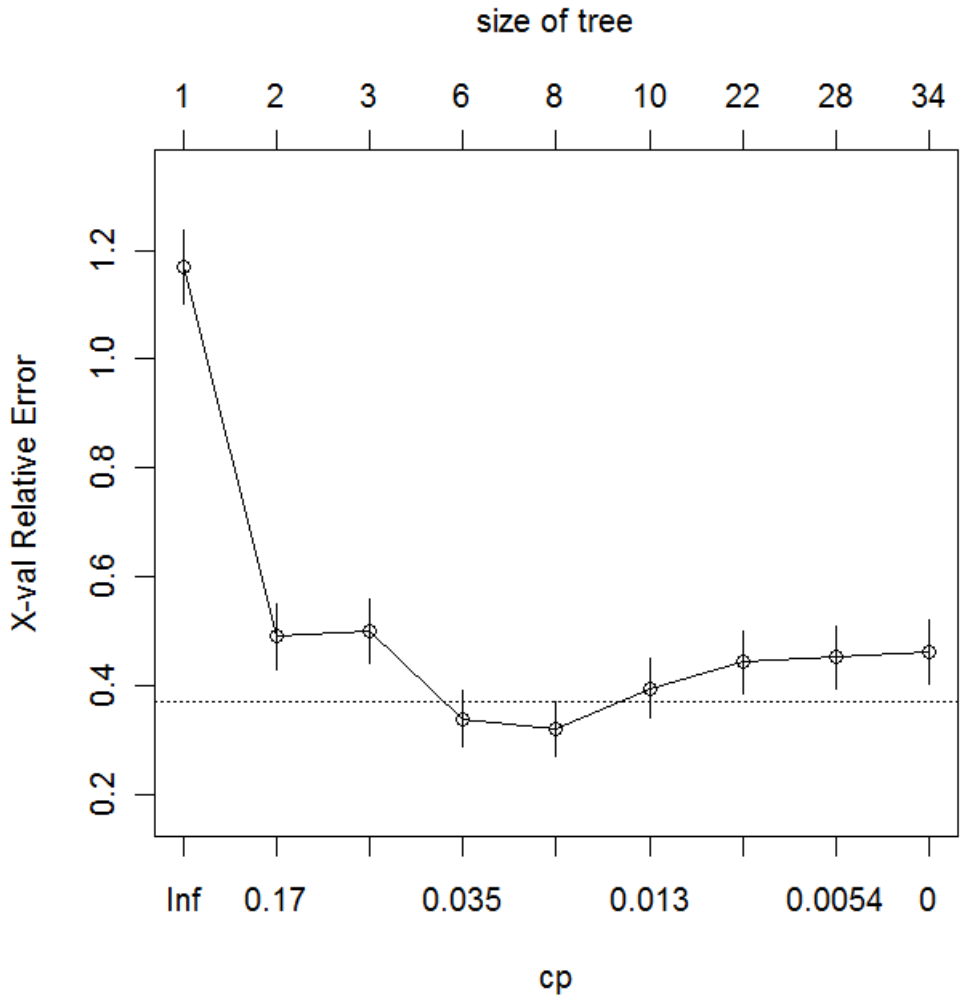




Pruning

- If the tree is too big, the lower “branches” are modeling noise in the data (“overfitting”).
- The usual paradigm is to grow the trees large and “**prune**” back unnecessary splits.
- Methods for **pruning** trees have been developed. Most use some form of crossvalidation. Tuning may be necessary.

Case Study: Cavity Nesting birds in the Uintah Mountains, Utah



Choose cp = .035

Crossvalidation Accuracy (cp = .035)

Actual Class	**Predicted Class**		**Total**
	Absence	Presence	
	0	1	
Absence, 0	85	21	106
Presence, 1	19	88	107
Total	104	109	213

Error rate = (19 + 21)/213 = (approx) .19 or 19%

Classification and Regression Trees

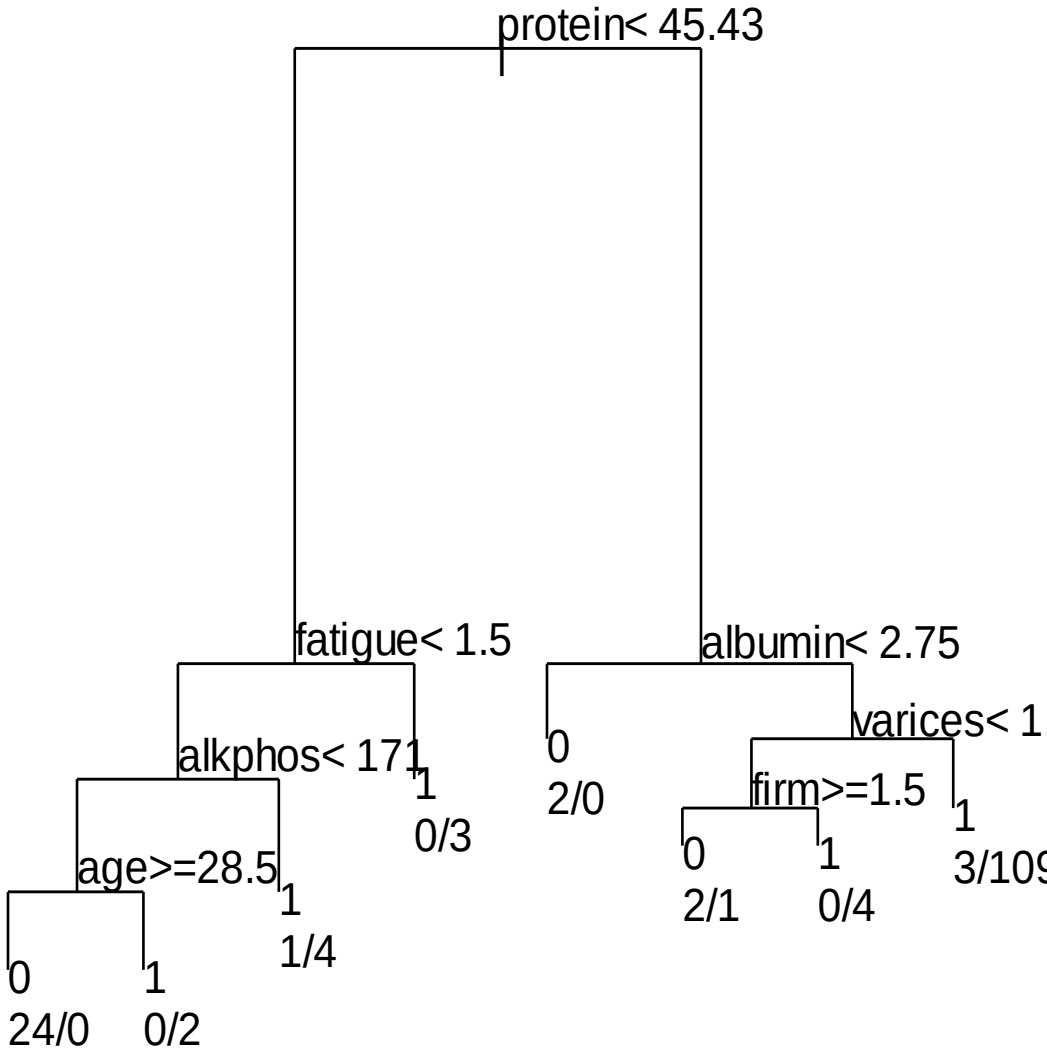
Advantages

- Applicable to both regression and classification problems.
- Handle categorical predictors naturally.
- Computationally simple and quick to fit, even for large problems.
- No formal distributional assumptions (non-parametric).
- Can handle highly non-linear interactions and classification boundaries.
- Automatic variable selection.
- Handle missing values through surrogate variables.
- Very easy to interpret if the tree is small.

Classification and Regression Trees

Advantages (ctnd)

- The picture of the tree can give valuable insights into which variables are important and where.
- The terminal nodes suggest a natural clustering of data into homogeneous groups.



Classification and Regression Trees

Disadvantages

- *Inaccuracy* - current methods, such as support vector machines and ensemble classifiers often have 30% lower error rates than CART.
- *Instability* – if we change the data a little, the tree picture can change a lot. So the interpretation is not as straightforward as it appears.

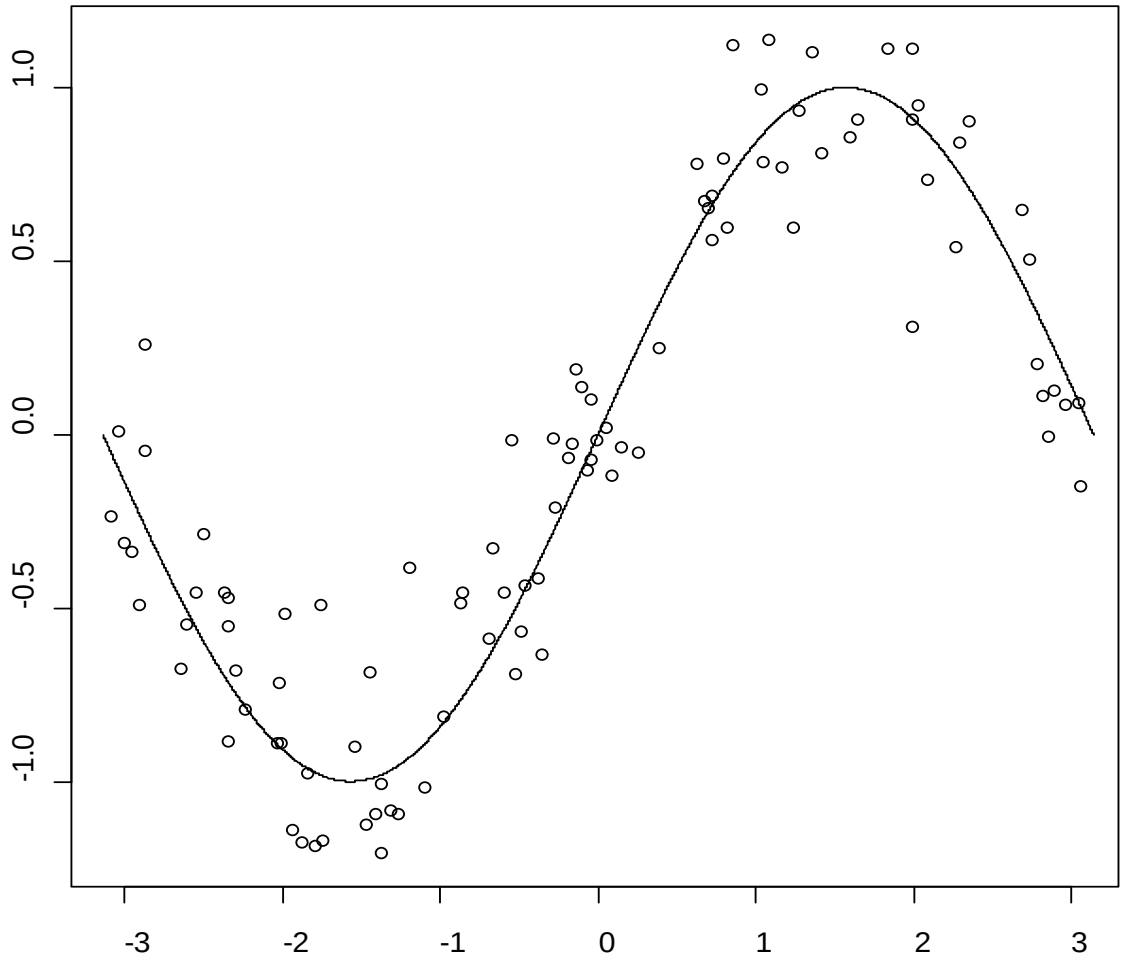
Today, we can do better!

Random Forests

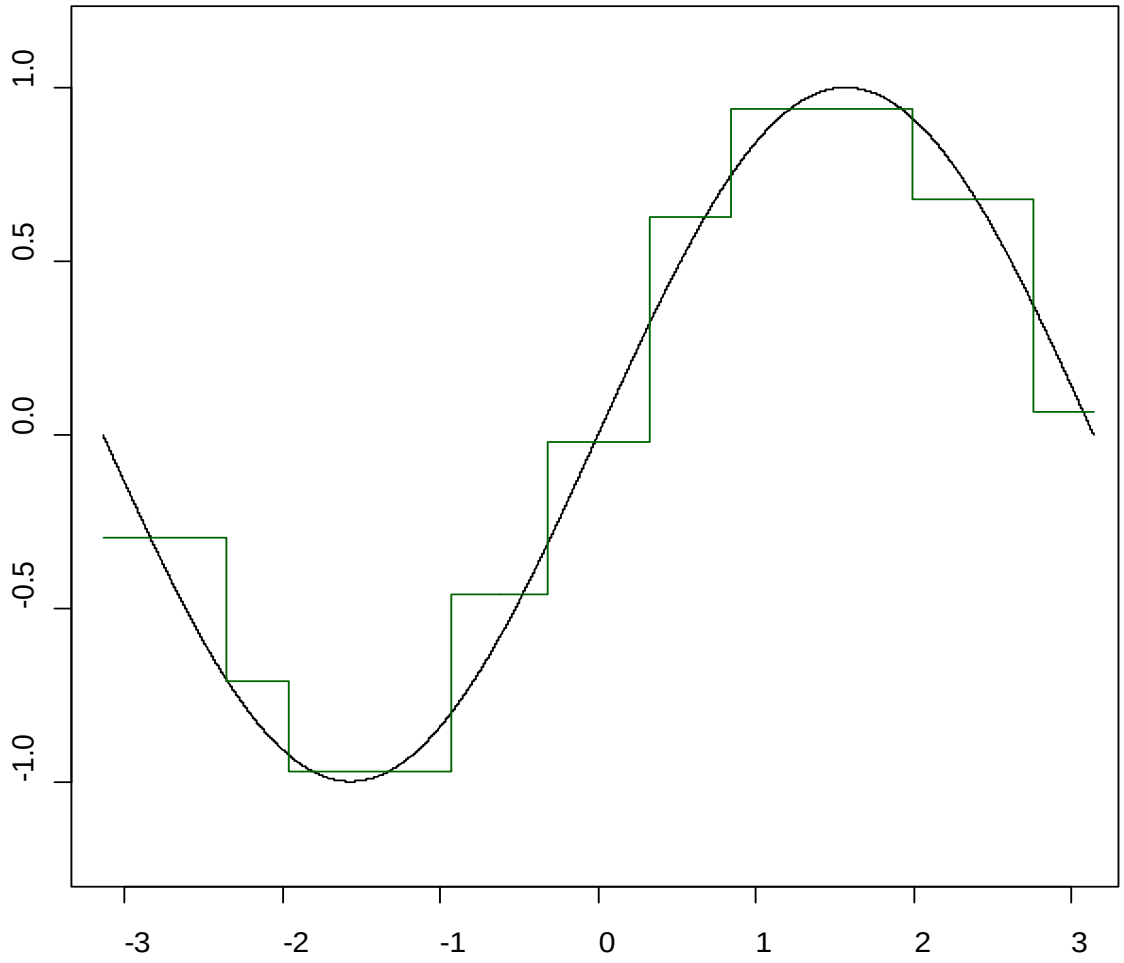
Outline

- Background.
- Trees.
- Bagging predictors.
- Random Forests algorithm.
- Variable importance.
- Proximity measures.
- Visualization.
- Partial plots and interpretation of effects.

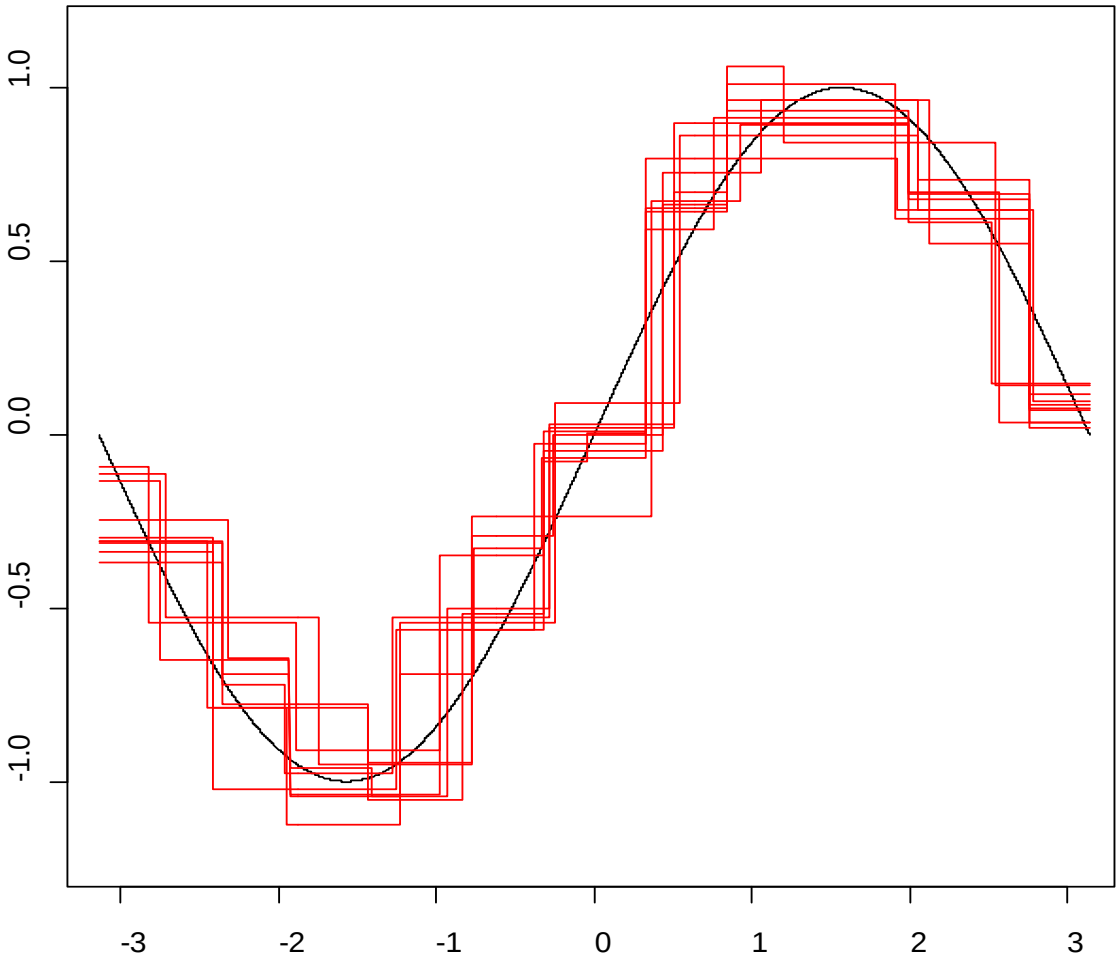
Data and Underlying Function



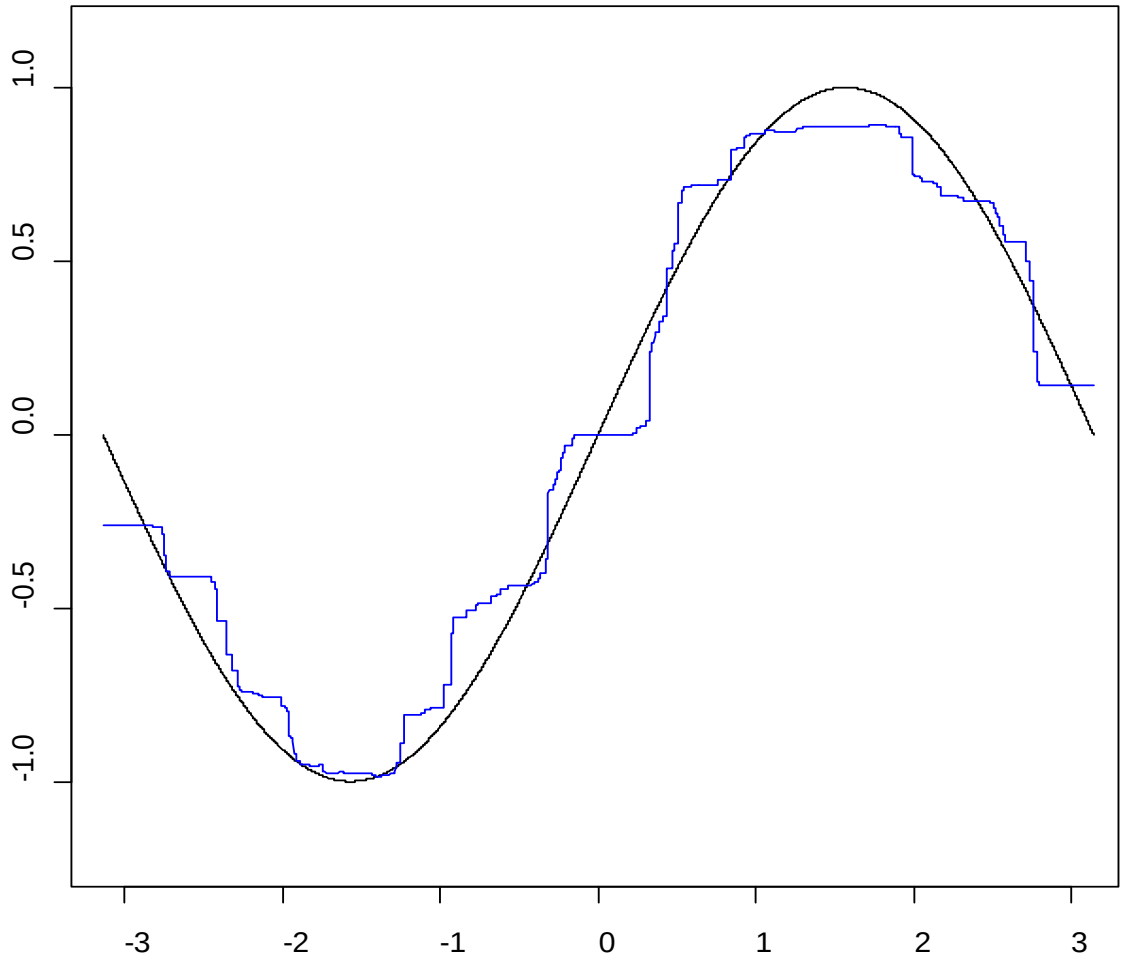
Single Regression Tree



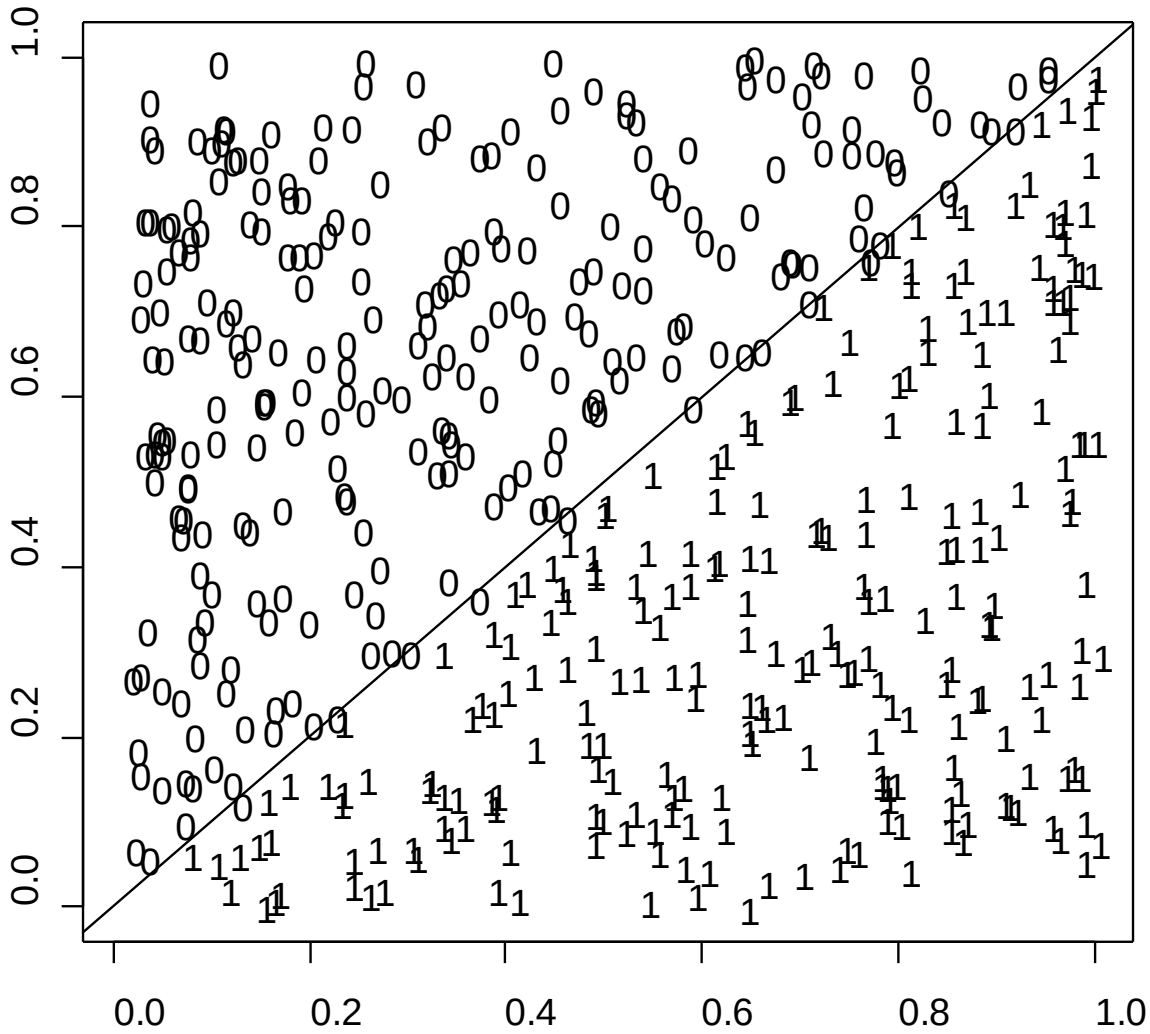
10 Regression Trees



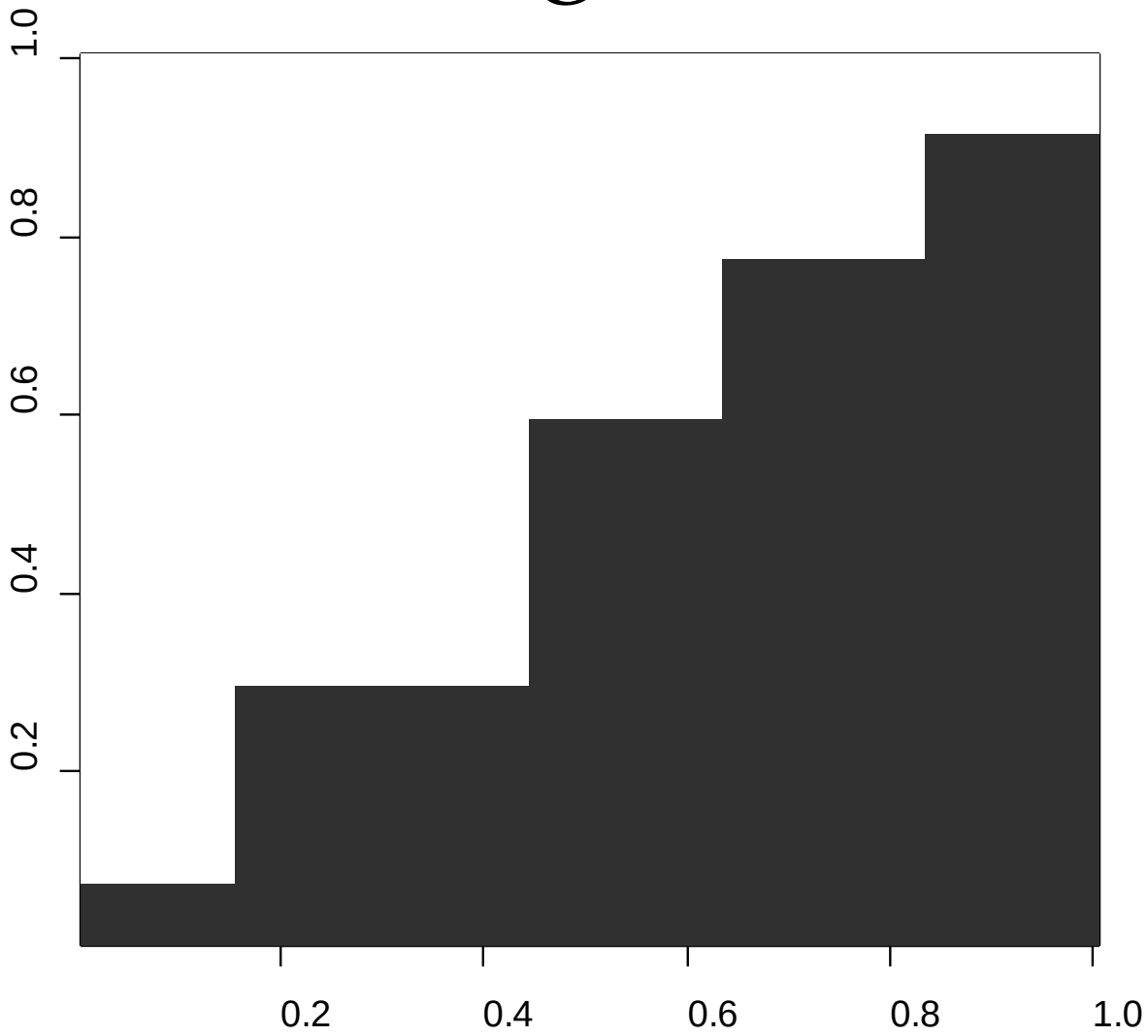
Average of 100 Regression Trees



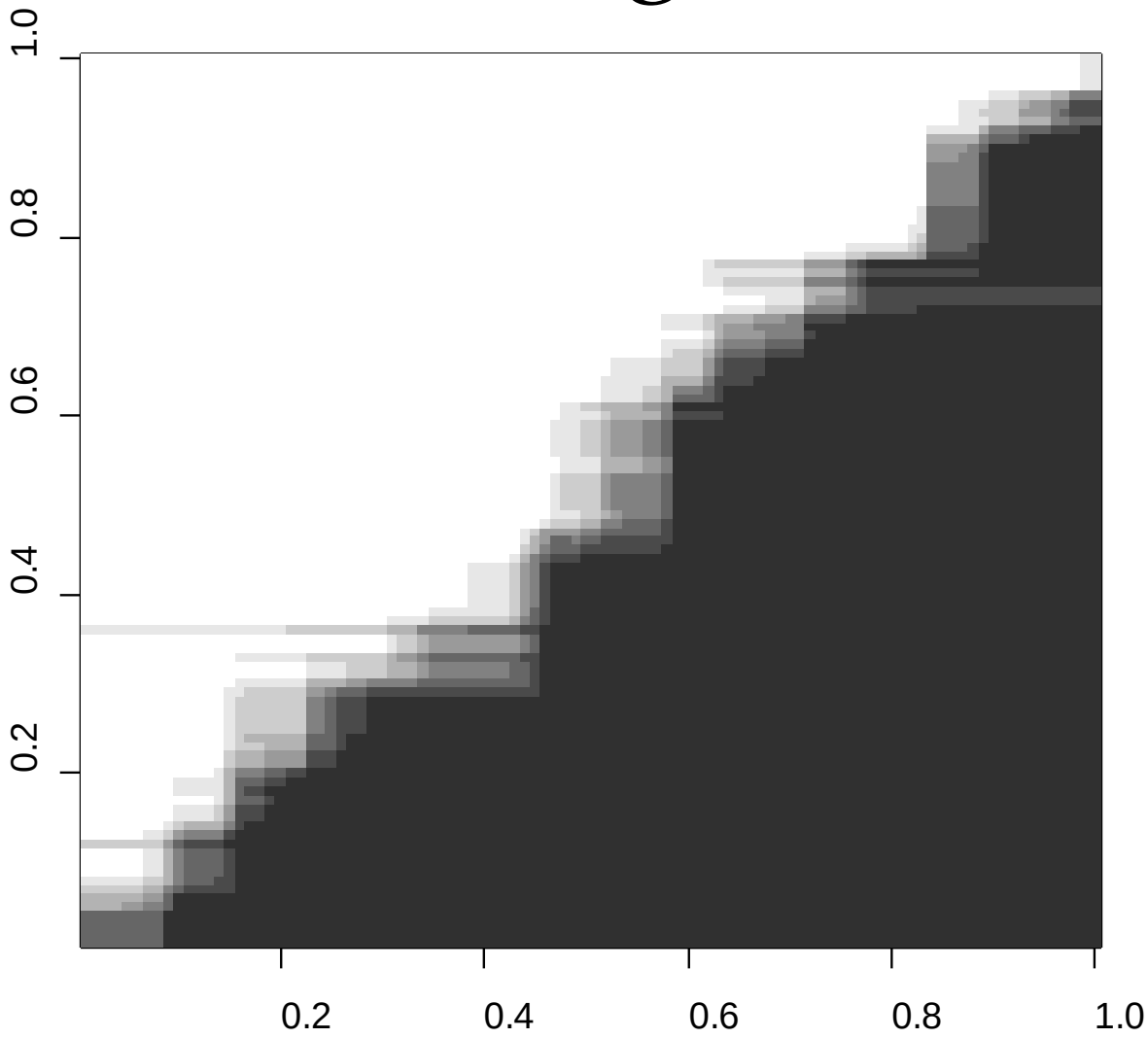
Hard problem for a single tree:



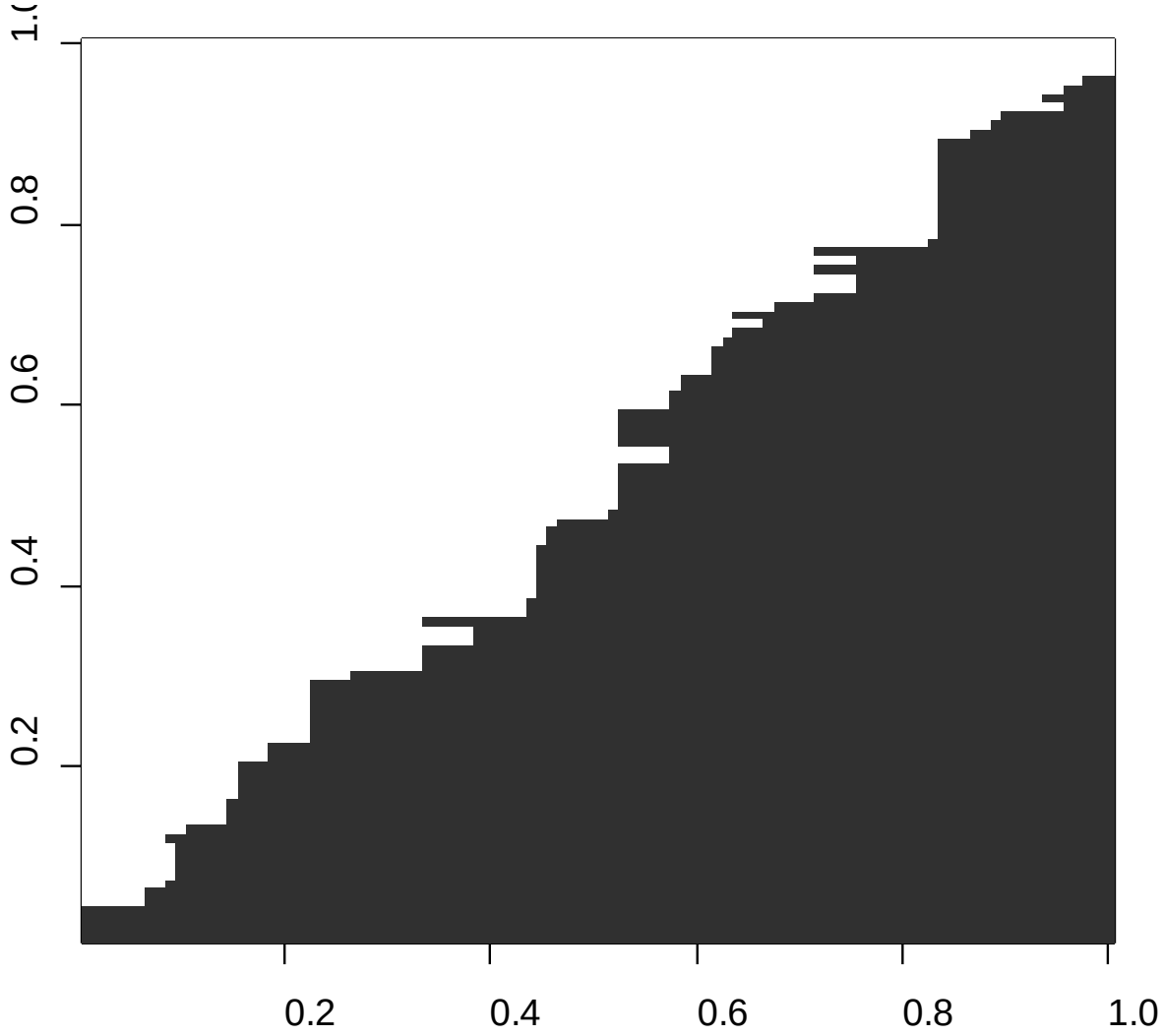
Single tree:



25 Averaged Trees:



25 Voted Trees:



Bagging (Bootstrap Aggregating)

Breiman, "Bagging Predictors", *Machine Learning*, 1996.

Fit classification or regression models to bootstrap samples from the data and combine by voting (classification) or averaging (regression).

Bootstrap sample $\Rightarrow$ $f_1(x)$

Bootstrap sample $\Rightarrow$ $f_2(x)$

Bootstrap sample $\Rightarrow$ $f_3(x)$

…

Bootstrap sample $\Rightarrow$ $f_M(x)$

MODEL AVERAGING

Combine $f_1(x),\ldots, f_M(x) \Rightarrow f(x)$

$f_i(x)$'s are "base learners"

Bagging (Bootstrap Aggregating)

- A bootstrap sample is chosen at random *with* replacement from the data. Some observations end up in the bootstrap sample more than once, while others are not included ("out of bag").
- Bagging reduces the *variance* of the base learner but has limited effect on the *bias*.
- It's most effective if we use *strong* base learners that have very little bias but high variance (unstable). E.g. trees.
- Both bagging and boosting are examples of "ensemble learners" that were popular in machine learning in the '90s.

Bagging CART

Dataset	# cases	# vars	# classes	CART	Bagged CART	Decrease %
Waveform	300	21	3	29.1	19.3	34
Heart	1395	16	2	4.9	2.8	43
Breast Cancer	699	9	2	5.9	3.7	37
Ionosphere	351	34	2	11.2	7.9	29
Diabetes	768	8	2	25.3	23.9	6
Glass	214	9	6	30.4	23.6	22
Soybean	683	35	19	8.6	6.8	21

Leo Breiman (1996) “Bagging Predictors”, Machine Learning, 24, 123-140.

Outline

- Background.
- Trees.
- Bagging predictors.
- Random Forests algorithm.
- Variable importance.
- Proximity measures.
- Visualization.
- Partial plots and interpretation of effects.

Random Forests

Dataset	# cases	# vars	# classes	CART	Bagged CART	Random Forests
Waveform	300	21	3	29.1	19.3	**17.2**
Breast Cancer	699	9	2	5.9	3.7	**2.9**
Ionosphere	351	34	2	11.2	7.9	**7.1**
Diabetes	768	8	2	25.3	**23.9**	24.2
Glass	214	9	6	30.4	23.6	**20.6**

Leo Breiman (2001) “Random Forests”, Machine Learning, 45, 5-32.

Random Forests

Grow a **forest** of many trees. (R default is 500)

Grow each tree on an independent **bootstrap sample*** from the training data.

At each node:

1. Select ***m*** variables **at random** out of all ***M*** possible variables (independently for each node).
2. Find the best split on the selected ***m*** variables.

Grow the trees to maximum depth (classification).

Vote/average the trees to get predictions for new data.

*Sample N cases at random with replacement.

Random Forests

Inherit many of the advantages of CART:

- Applicable to both regression and classification problems. Yes.
- Handle categorical predictors naturally. Yes.
- Computationally simple and quick to fit, even for large problems. Yes.
- No formal distributional assumptions (non-parametric). Yes.
- Can handle highly non-linear interactions and classification boundaries. Yes.
- Automatic variable selection. Yes. But need variable importance too.
- Handles missing values ~~through surrogate variables.~~ Using proximities.
- ~~Very easy to interpret if the tree is small~~. NO!

Random Forests

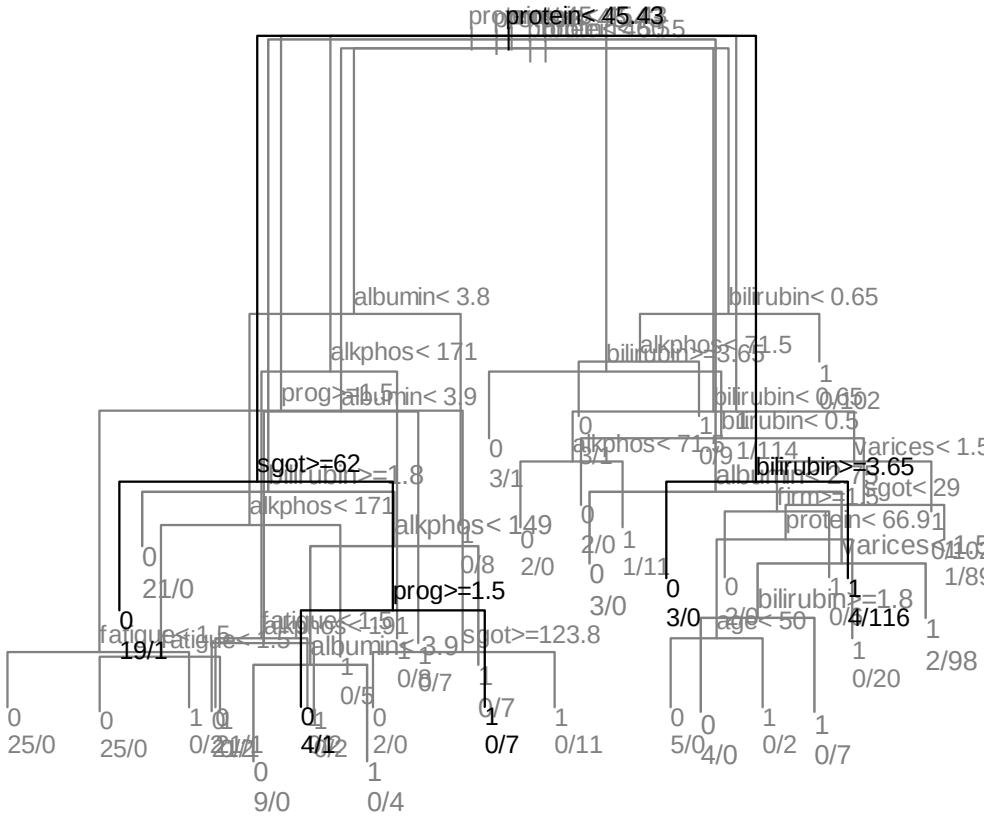
But do not inherit:

- The picture of the tree can give valuable insights into which variables are important and where.

NO!

- The terminal nodes suggest a natural clustering of data into homogeneous groups.

NO!



Random Forests

Improve on CART with respect to:

- *Accuracy* – Random Forests is competitive with the best known machine learning methods (but note the "no free lunch" theorem).

- *Stability* – if we change the data a little, the individual trees may change but the forest is relatively stable because it is a combination of many trees.

References

- Leo Breiman, Jerome Friedman, Richard Olshen, Charles Stone (1984) “Classification and Regression Trees” (Wadsworth).
- Leo Breiman (1996) “Bagging Predictors” Machine Learning, 24, 123-140.
- Leo Breiman (2001) “Random Forests” Machine Learning, 45, 5-32.
- Trevor Hastie, Rob Tibshirani, Jerome Friedman (2009) “Statistical Learning” (Springer).