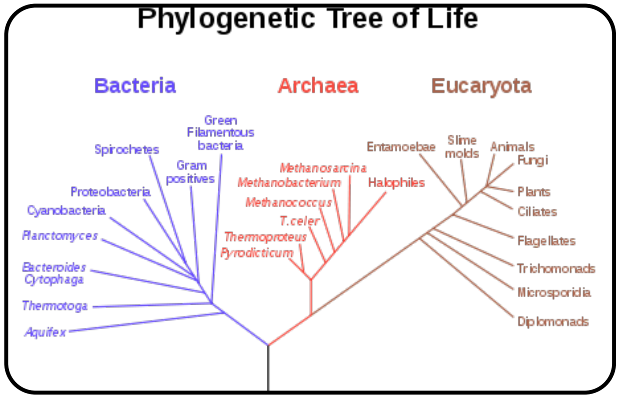
Classification and Regression Trees

Bob Stine
Dept of Statistics, Wharton School
University of Pennsylvania

Wharton
Department of Statistics

Trees

- Familiar metaphor
  - Biology
  - Decision tree
  - Medical diagnosis
  - Org chart



- Properties
  - Recursive, partitioning items into unique leaf
  - Increasing specialization

- Convey structure at-a-glance
- How to grow a tree from data?
  - What rules identify the relevant variables, split rules?

Trees as Models for Data

- Different type of explanatory variable
  - Decision rules replace typical predictors
  - Implicit equation uses indicator functions

$$X \Rightarrow I_{x \leq c} \;\; \& \;\; I_{x > c}$$

  - Software builds these from training data
- Process
  - Find rule to partition data
  - Fits are averages of subsets
  - Use validation data to decide when to stop
- Models as averages
  - All models average, just question of which cases

Old Idea

- Binning data
  - Use categorical variables to define bins
  - Each observation goes into a bin
  - Prediction
    - average of cases in bin
    - most common category in bin
- Classify new case
  - No equation: Use score for the matching bin
- Trade-offs
  - Good: avoid assuming additive, transformations
  - Bad: Some bins may be nearly empty, sparse

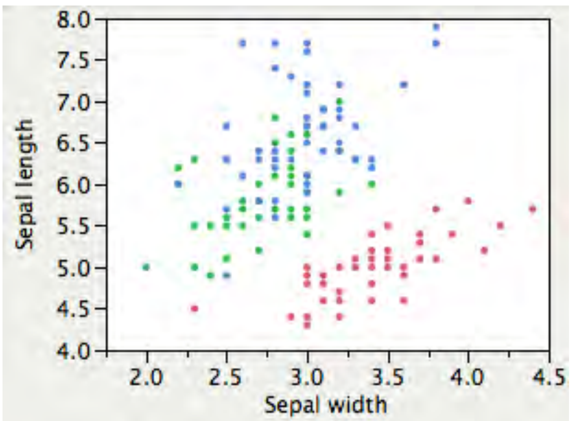
    Need lots of data to fill a contingency table with several axes
  - Issues: Which characteristics? Which attributes?

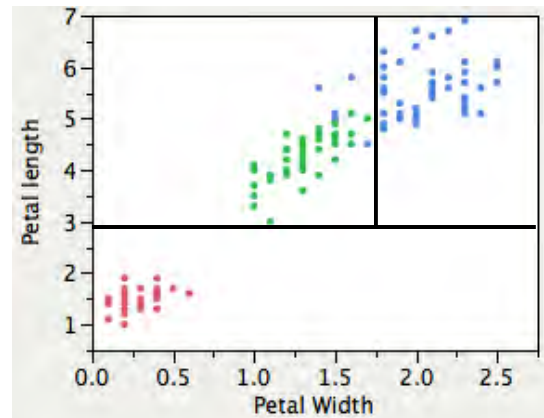
bias
vs
variance

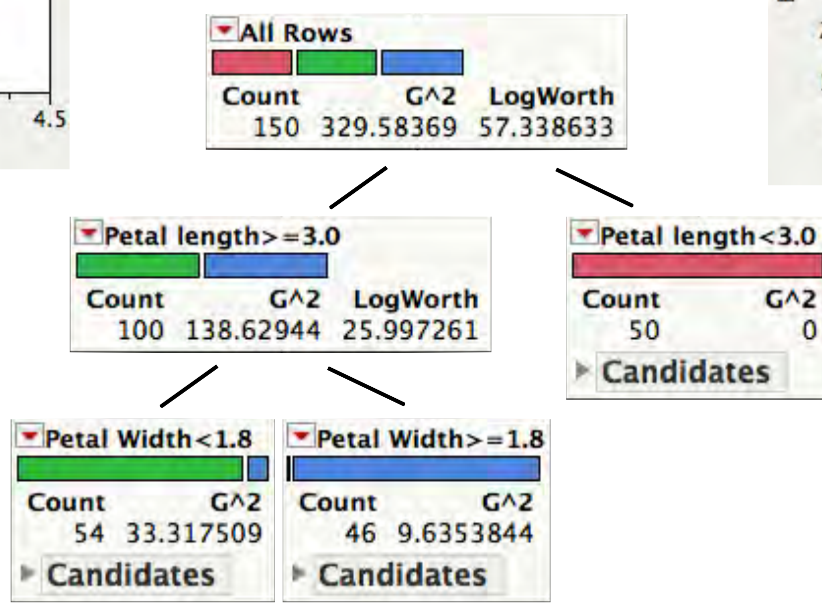
Classical Example

Classification tree

- Fisher's iris data
  - Classification tree: categorical response
  - 50 flowers from 3 species of iris
  - four variables: length and width of sepal and petal







G^2 = - 2 log likelihood
= 2 entropy

Stop?

Splits are parallel to plot axes

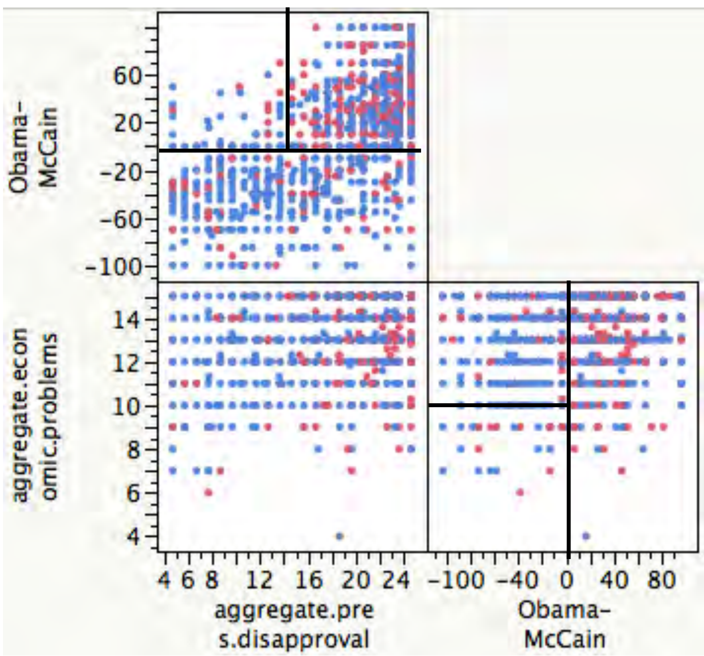
Splitting rules are not unique
Tree version of collinearity

Example

Regression tree with dummy response

- ANES 2008
  - Regression tree: numerical response
  - Favor or oppose gay marriage
  - X's: Obama-McCain, PresDiapproval, Econ Problem

Gay Marriage
- Favor
- Oppose



Use "Select Rows" command in the tree nodes

All Rows

		LogWorth	Difference
Count	1539	29.072897	0.26554
Mean	0.3482781		
Std Dev	0.4765795		

Node shows average of response (here percentage) for its cases

Stop?

RSquare	RMSE	N	Number of Splits	AICc
0.075	0.458175	1539	3	1975.14

Recursive Partitioning

- Recursive, binary splits   CART™
  - Start with all cases in one group, the root node
    Tree grows upside down
  - Split a current group to make homogeneous
    May split same group several times
  - Continue until objective is reached
- Comments
  - Recursive: once cases are split, never rejoin
  - Greedy: immediate step rather than look ahead
    Very fast, even with many features
  - Invariant of order-preserving transformations
  - Rules are not unique (as in collinearity in regr)
  - Interactions

Growing Tree

- Search for best splitting variable
  - Numerical variable

    Partition cases $X \leq c$ and $X > c$, all possible c

    Consider only numbers c that match a data point (ie, sort cases)
  - Categorical variable

    Partition cases into two mutually exclusive groups

    Lots of groups if the number of labels k is large ($2^{k-1}-1$ splits)
- Greedy search
  - One-step look ahead (as in forward stepwise)
  - Find next variable that maximizes search criterion, such as level of significance or R^2.
  - Criterion depends on response: numerical or categorical

Splitting Criteria

- Numerous choices
- Log-likelihood for classification tree
  - Recall -2 log likelihood ≈ residual SS in OLS
  - G^2 is node's contribution to -2 log likelihood

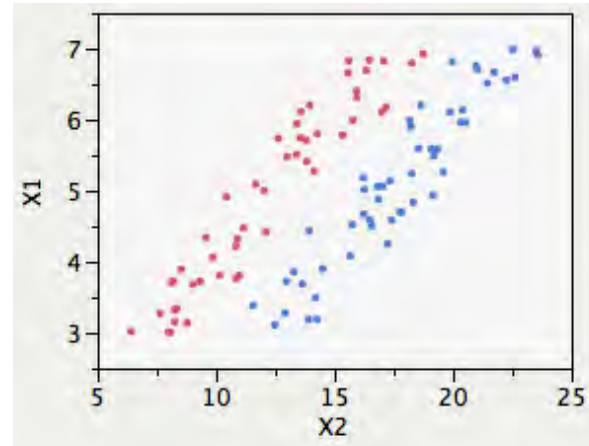
    Related to the entropy of the current partition
    (entropy measures randomness)
  - $G^2 = 0$ for node that is homogenous

    perfect fit, no value in trying to split further (entropy = 0)
- Log worth
  - JMP version of the p-value of a split
- Cross-validation
  - Use a tuning sample to decide how many splits

Common Limitations

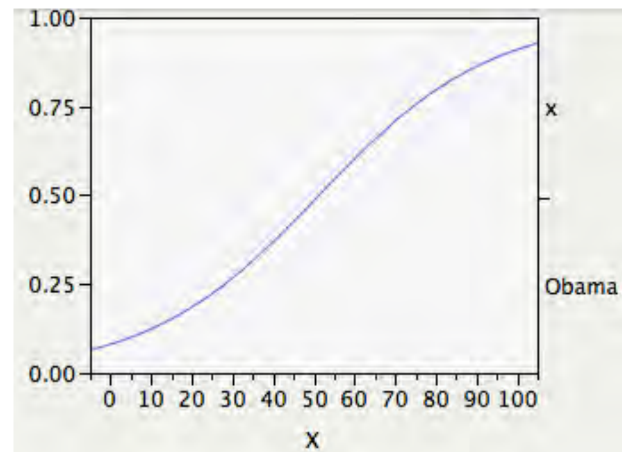
- Splits are parallel to axis
  - Binary split on an observed variable
  - Some tree methods allow splits on linear combination

    slower to fit since many more possible splits



- Discrete fit
  - Piecewise constant fit

    Lots of splits on one variable indicate trend



- Greedy search

  Vert fast but can miss the best partition

  Common advice: over-fit then prune back

  As used for AIC, BIC in regr

- Over-fitting

Example: ANES

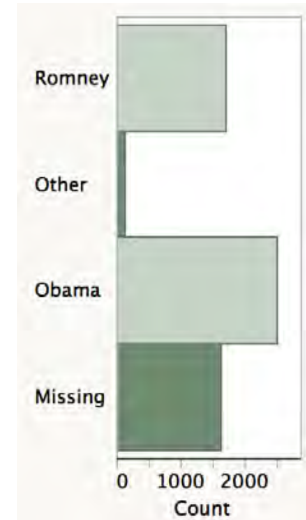
- Classify those who did not vote
  - Use 3-level validation variable

    ≈4000 observed Obama/Romney, exclude others

    0 = training, 1 = tuning, determines tree size

    2 = test sample
  - Big assumption: same rules apply to those who voted and did not vote
- Predictive features to consider
  - Avoid direct Obama/Romney specific questions

    Keep the problem more challenging
  - Demographics
  - Missing indicators



Missing	1610	0.27
Obama	2496	0.42
Other	118	0.01
Romney	1692	0.28
Total	5916	1.00

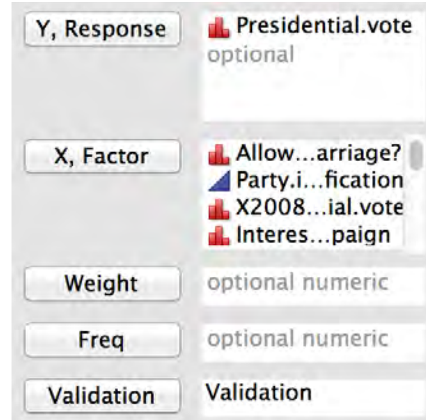
sample weights?

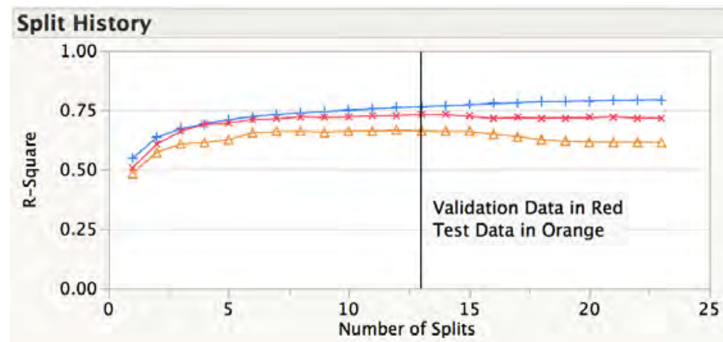
Fitting the Tree

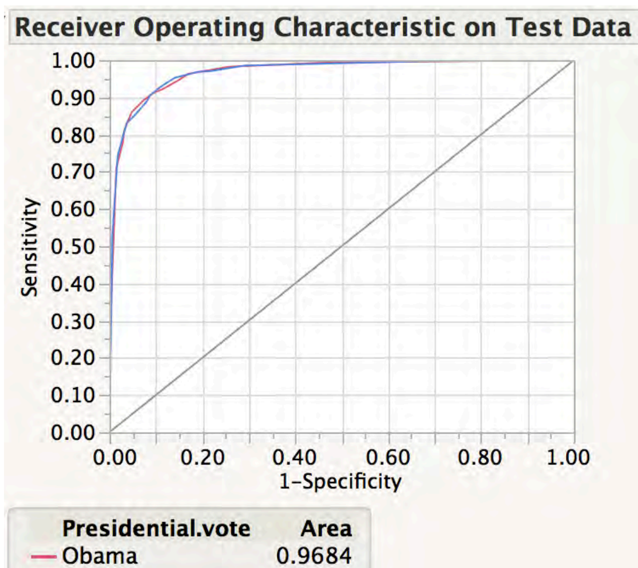
- Running options
  - Minimum split size 25

    Avoid leaves with few cases
  - Nice interface

    Can force splits at any location
- Validation properties







	RSquare	N	Number of Splits
Training	0.764	2167	13
Validation	0.730	699	
Test	0.663	1322	

Column Contributions

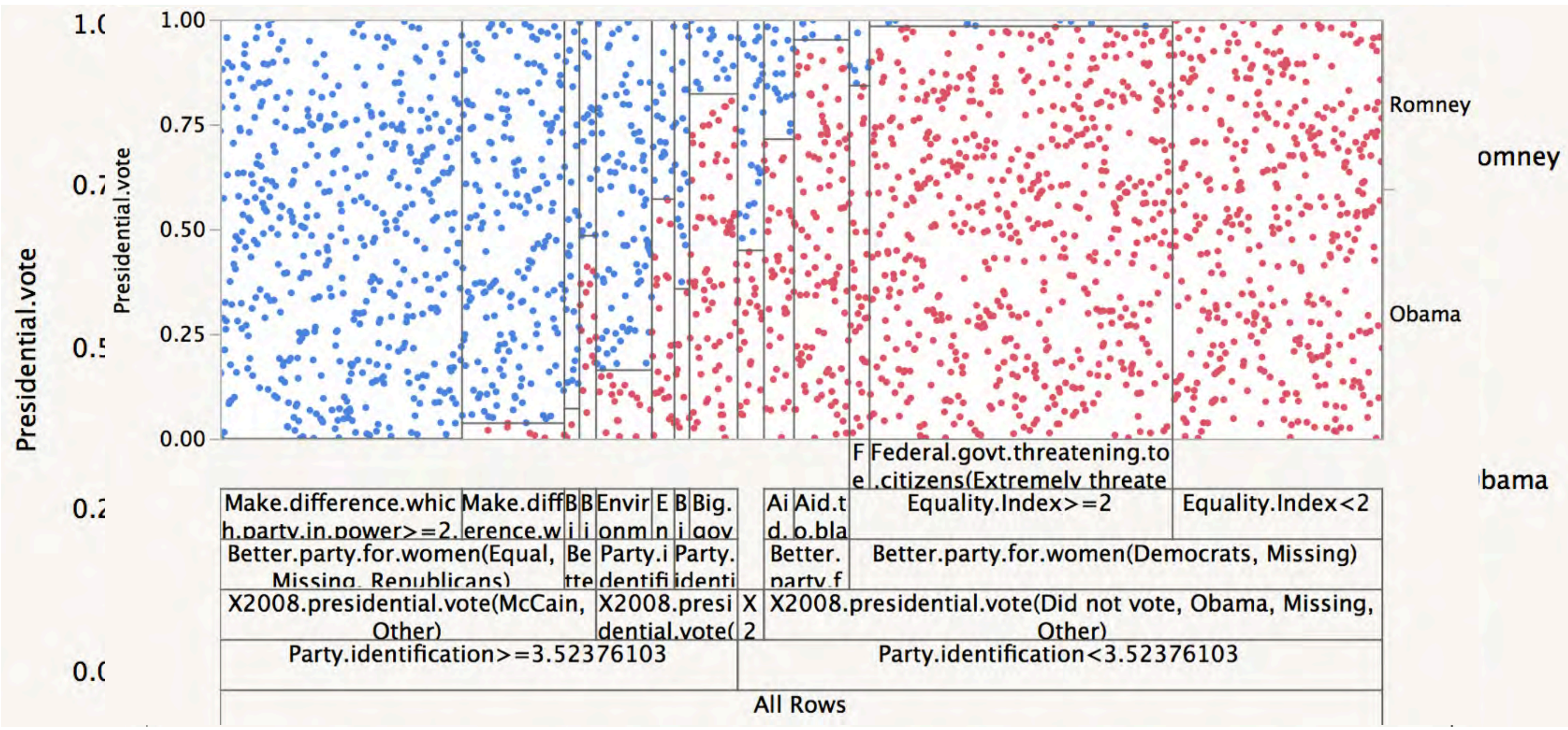
Term	Number of Splits	G^2	Portion
Party.identification	2	1652.00843	0.7380
~~X2008.presidential.vote~~	2	362.331905	0.1619
Better.party.for.women	2	101.85863	0.0455
Big.government.index	2	34.4754097	0.0154
Environmental.protection.scale	1	23.3682291	0.0104
Make.difference.which.party.in.power	1	17.1555426	0.0077
Aid.to.blacks.scale	1	17.1072163	0.0076
Equality.Index	1	15.1433622	0.0068
Federal.govt.threatening.to.citizens	1	14.9227152	0.0067
Allow Gay Marriage?	0	0	0.0000
Interest.in.campaign	0	0	0.0000
Media Frequency	0	0	0.0000

What happens if this feature is not used?

Would be a nice feature for NN as well!

Mosaic Plot

- Summary of tree
  - Thin bins have few cases
  - Less flat means better splits

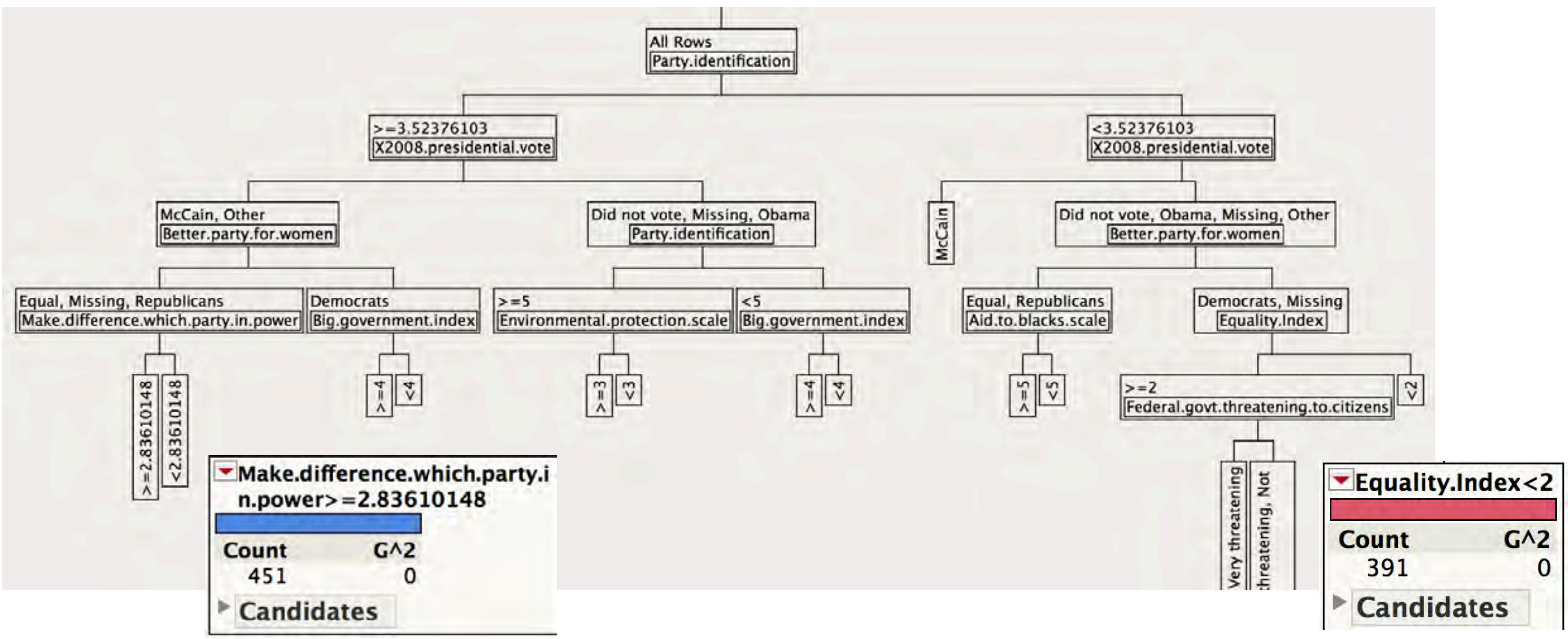


Romney Obama

Estimated Tree

- Note variables that define first splits
  - Feeling thermometer differences, several splits
  - Race, but only for some
  - Voting behavior
- Some leaves are very homogeneous
  - No point in further splitting

Very 'parallel' structure



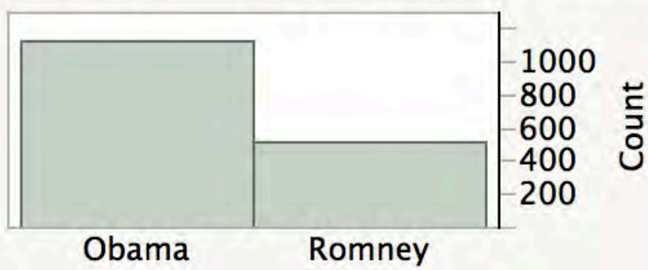
Classify Missing

- Majority vote
  - 'Drop case' into estimated tree
  - Classify based on the preponderance of cases
- Results
  - Save tree prediction formula (not predicteds)

    Get probabilities* as well as most likely choice
  - Distribution of predicteds for missing cases

Distributions Presidential.vote=Missing

Most Likely Presidential.vote



Frequencies

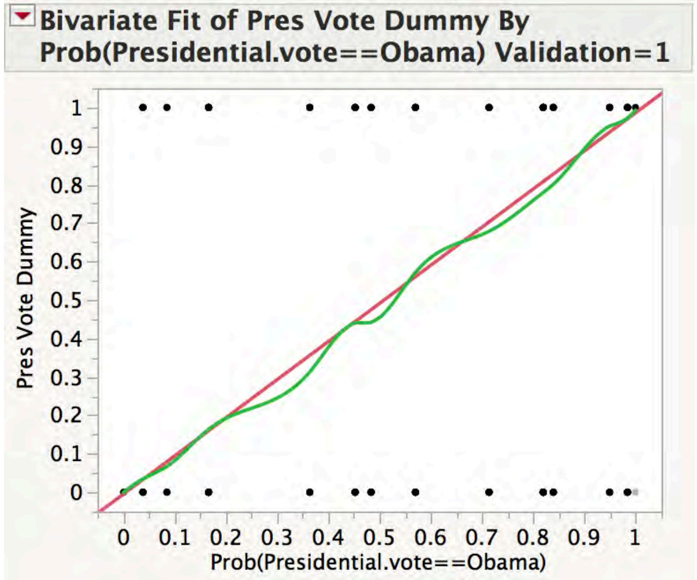
Level	Count	Prob
Obama	1107	0.68758
Romney	503	0.31242
Total	1610	1.00000
N Missing	0	

60/40 split among observed voters

*JMP smooths probabilities using prior from parent node.

Things to Improve?

- So few possible values
  - Number of leaf nodes determines the number of possible predictions; very discrete fit.
- Highly variable
  - Take a different subset and split points change
  - Fitted values, however, are likely similar
- Calibrated, but few possible values



Example from training sample

Averaging Trees

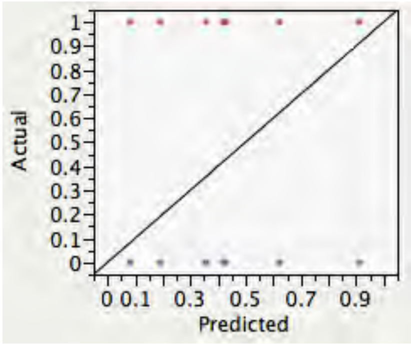
- Rather than average within a model, we can average over models
- Model averaging borrows strength
  - Fit collection of models
  - Predict by ‘majority vote’ or averaging
  - Question: How to get a collection of models?
- Boosting
  - Re-weight cases not fit well by current model
    (If numerical Y, fit next model to residuals of current model)
  - Simple models
- Bagging
  - Build trees (forest) using bootstrap samples
  - Complicated models, different sets of variables

Random Forest

- Problem with trees
  - 'Grainy' predictions, few distinct values

    Each final node gives a prediction
  - Highly variable

    Sharp boundaries, huge variation in fit at edges of bins



- Random forest
  - Cake-and-eat-it solution to bias-variance tradeoff

    Complex tree has low bias, but high variance.

    Simple tree has high bias, but low variance.
  - Fit ensemble of trees, each to different BS sample
  - Average of fits of the trees
  - Increase independence of trees by forcing different variables in the different trees

    Often need relatively big tree to capture interesting structure

Random Forest

- Fit using random forest
  - Classification tree has only few leaves

    Very coarse predictions of voting behavior (though maybe enough)
  - Forest has more branches, more variables

bottom left of dialog

Method	Bootstrap Forest ▾

- Summary of forest
  - More variables used

Number of trees in the forest	200
Number of terms sampled per split:	49
Bootstrap sample rate:	1
Minimum Splits Per Tree:	10
Maximum Splits Per Tree	2000
Minimum Size Split:	25
☑ Early Stopping	

Target Column:	Presidential.vote	Training rows:	2167
Validation Column:	Validation	Validation rows:	699
		Test rows:	1322
Number of trees in the forest:	46	Number of terms:	196
Number of terms sampled per split:	49	Bootstrap samples:	2167
		Minimum Splits Per Tree:	10
		Minimum Size Split:	25

Column Contributions

Term	Number of Splits	G^2	Portion
Party.identification	112	22516.6355	0.3627
X2008.presidential.vote	65	12031.6696	0.1938
Better.party.for.women	59	7245.49179	0.1167
Big.government.index	42	4303.97312	0.0693
Ideology	32	1975.53171	0.0318
Health.insurance.plan.scale	36	1710.34517	0.0275
Economy.past.year	35	1698.3951	0.0274
Unemployment.past.year	44	1336.7817	0.0215
Tea.Party.support	40	1264.07236	0.0204
Offshore.drilling	26	910.504349	0.0147
Race	38	860.210261	0.0139
Moral.Traditionalism	28	821.10427	0.0132

Forest Results

- Confusion matrix

Confusion Matrix

Actual		Predicted
Training	**Obama**	**Romney**
Obama	1248	40
Romney	64	815

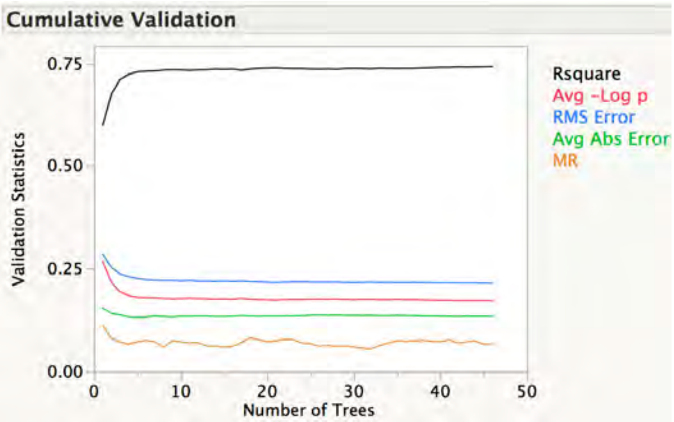
Actual		Predicted
Validation	**Obama**	**Romney**
Obama	412	19
Romney	20	248

Actual		Predicted
Test	**Obama**	**Romney**
Obama	732	45
Romney	47	498

- Goodness of fit summary

Overall Statistics

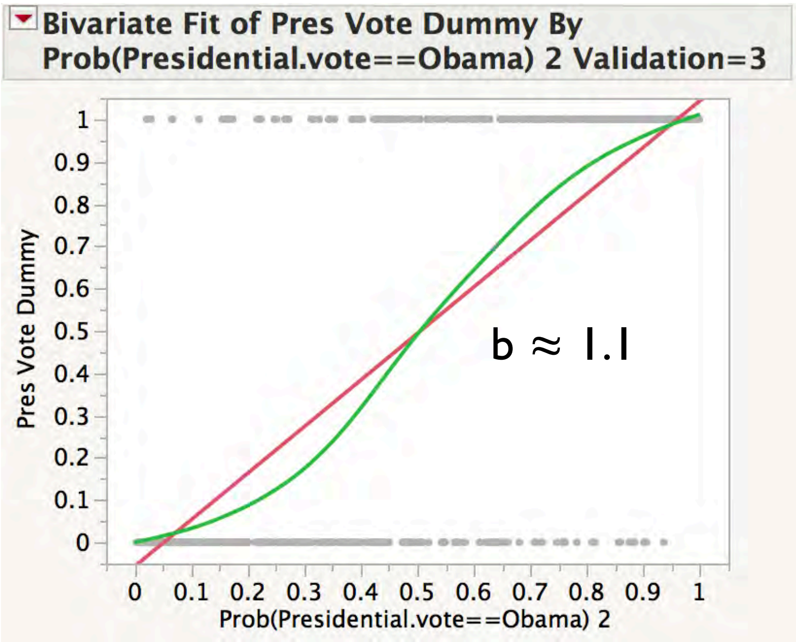
Measure	Training	Validation	Test	Definition
Entropy RSquare	0.7620	0.7419	0.7097	1-Loglike(model
Generalized RSquare	0.8674	0.8528	0.8325	(1-(L(0)/L(model
Mean -Log p	0.1607	0.1718	0.1967	$\sum -\mathrm{Log}(\rho[j])/n$
RMSE	0.2038	0.2142	0.2339	$\sqrt{\sum(y[j]-\rho[j])^2/n}$
Mean Abs Dev	0.1252	0.1335	0.1415	$\sum \lvert y[j]-\rho[j]\rvert/n$
Misclassification Rate	0.0480	0.0558	0.0696	$\sum (\rho[j]\neq\rho Max)/n$
N	2167	699	1322	n

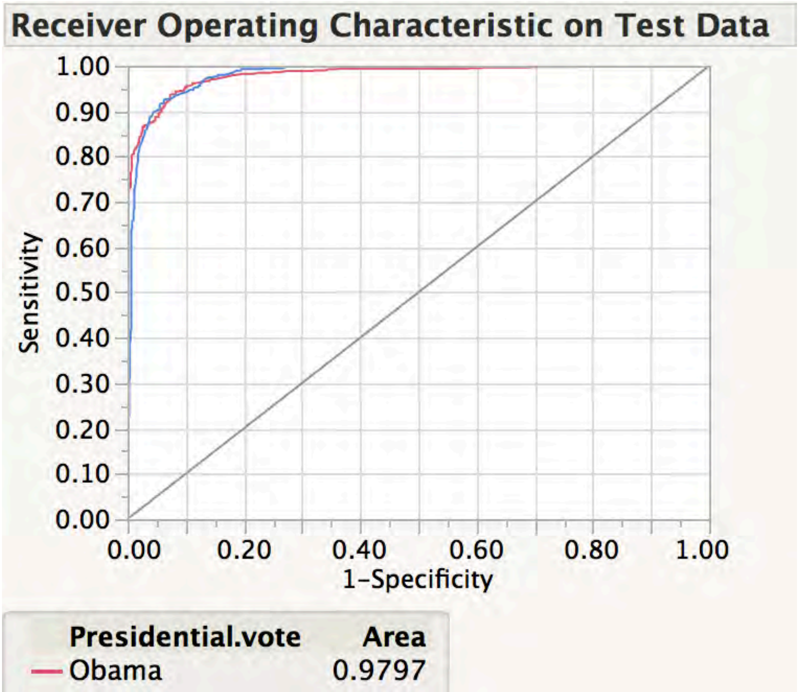


progress as forest grows

Calibration Plot

- Test sample results for random forest
- Richer set of predictions
  - linear, but not with slope 1
- Smooth ROC





Boosting

- General method for improving predictive model
  - Build additive sequence of predictive models (ensemble)

    Final prediction is accumulated over many models.
  - Start with initial predictive model
  - Compute residuals from current fit
  - Build model for residuals
  - Repeat

Original method called Adaboost

- Implication: Use simple model at each step
  - Weak learner: 'stump' (one split), few splits
  - Next response = (current response) - (learning rate) x fit

    0.1 or smaller
- Weaknesses
  - Loss of 'interpretability', at what gain?

Boosted Trees

Only in JMP Pro

Method: Boosted Tree

- Different way to get multiple trees
  - Simple models
  - Refit to training sample, but put more weight to cases not fit well so far
- Uses many variables without random exclusion

Gradient-Boosted Trees Specification

Number of Layers:	200
Splits Per Tree:	3
Learning Rate:	0.01
Overfit Penalty:	0.0001
Minimum Size Split:	5

- [x] Early Stopping
- [ ] Multiple Fits over splits and learning rate:

Max Splits Per Tree	3
Max Learning Rate	0.1

Specifications

Target Column:	Presidential.vote	Number of training rows:	2167
Validation Column:	Validation	Number of validation rows:	699
Number of Layers:	200	Number of test rows:	1322
Splits Per Tree:	3		
Learning Rate:	0.01		
Overfit Penalty:	0.0001		

Column Contributions

Term	Number of Splits	G^2	Portion
Party.identification	81	206582.042	0.2654
Post.stratified.sample.weight	231	143276.078	0.1841
X2008.presidential.vote	56	141790.62	0.1821
Better.party.for.women	51	105168.516	0.1351
Big.government.index	33	61440.0214	0.0789
Economy.past.year	42	36496.9081	0.0469
Health.insurance.plan.scale	22	28999.8764	0.0373
Unemployment.past.year	24	15180.9567	0.0195
Environmental.protection.scale	19	13967.6236	0.0179
Moral.Traditionalism	24	13416.6216	0.0172
Defense.spending	4	5879.29482	0.0076
Regulation.of.business	7	2672.81659	0.0034
Economy.next.year	3	1802.83762	0.0023

Boosting Results

- Confusion matrix

Confusion Matrix

Variation in choice of test sample?

Boosted:

Actual / Predicted: Training	Obama	Romney
Obama	1242	46
Romney	78	801

Actual / Predicted: Validation	Obama	Romney
Obama	410	21
Romney	23	245

Actual / Predicted: Test	Obama	Romney
Obama	735	42
Romney	54	491

⇐ Boosted

Forest:

Actual / Predicted: Training	Obama	Romney
Obama	1248	40
Romney	64	815

Actual / Predicted: Validation	Obama	Romney
Obama	412	19
Romney	20	248

Actual / Predicted: Test	Obama	Romney
Obama	732	45
Romney	47	498

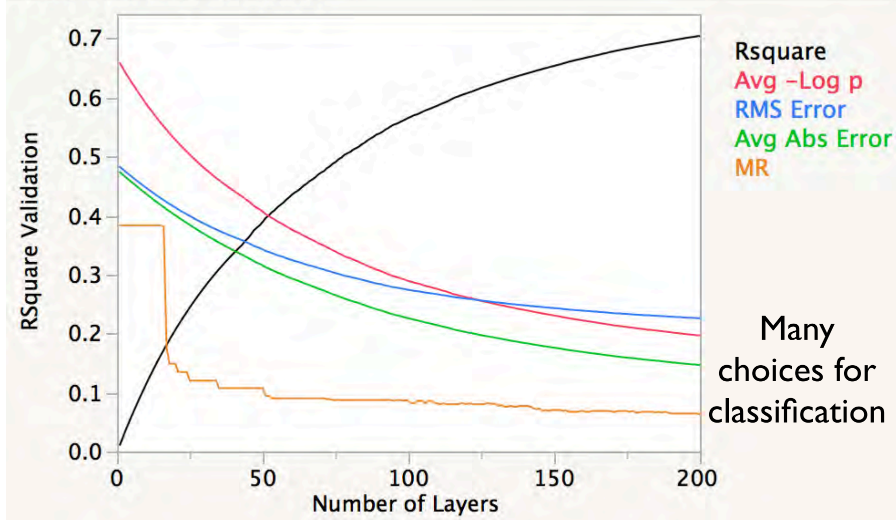
⇐ Forest

- Goodness of fit summary

Overall Statistics

Measure	Training	Validation	Test	Definition
Entropy RSquare	0.7241	0.7046	0.6859	1-Loglike(model)/Loglike
Generalized RSquare	0.8421	0.8271	0.8156	(1-(L(0)/L(model))^(2/n))
Mean -Log p	0.1863	0.1966	0.2128	∑ -Log(ρ[j])/n
RMSE	0.2159	0.2258	0.2390	√ ∑(y[j]-ρ[j])²/n
Mean Abs Dev	0.1403	0.1467	0.1514	∑ |y[j]-ρ[j]|/n
Misclassification Rate	0.0572	0.0629	0.0726	∑ (ρ[j]≠ρMax)/n
N	2167	699	1322	n

Cumulative Validation

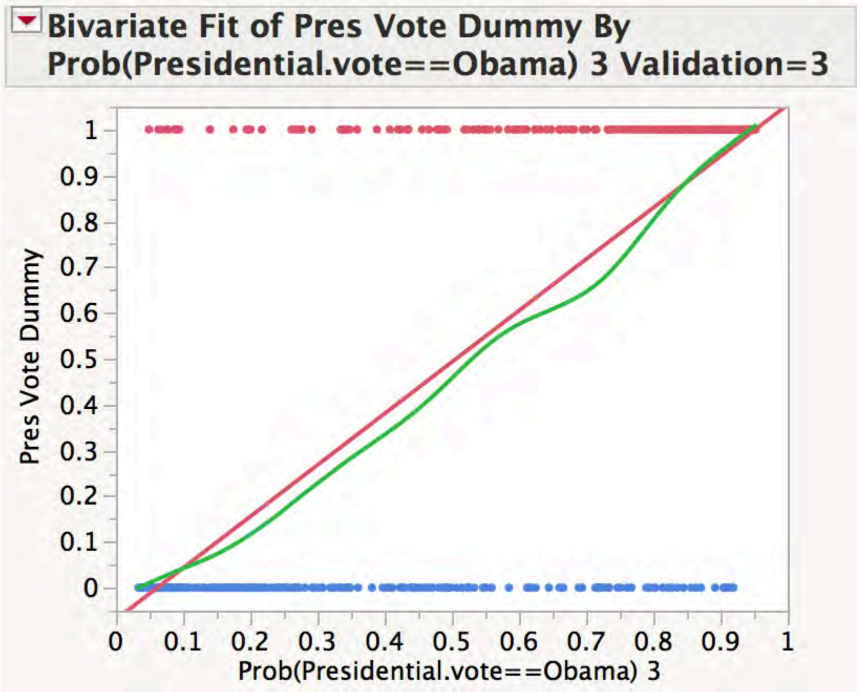


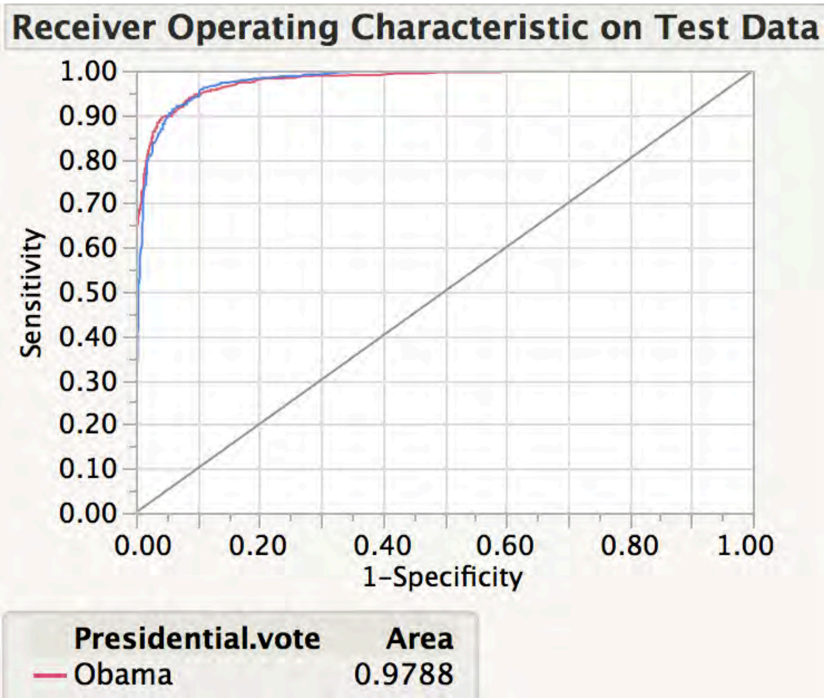
Many choices for classification

should have run longer!

Calibration Plots

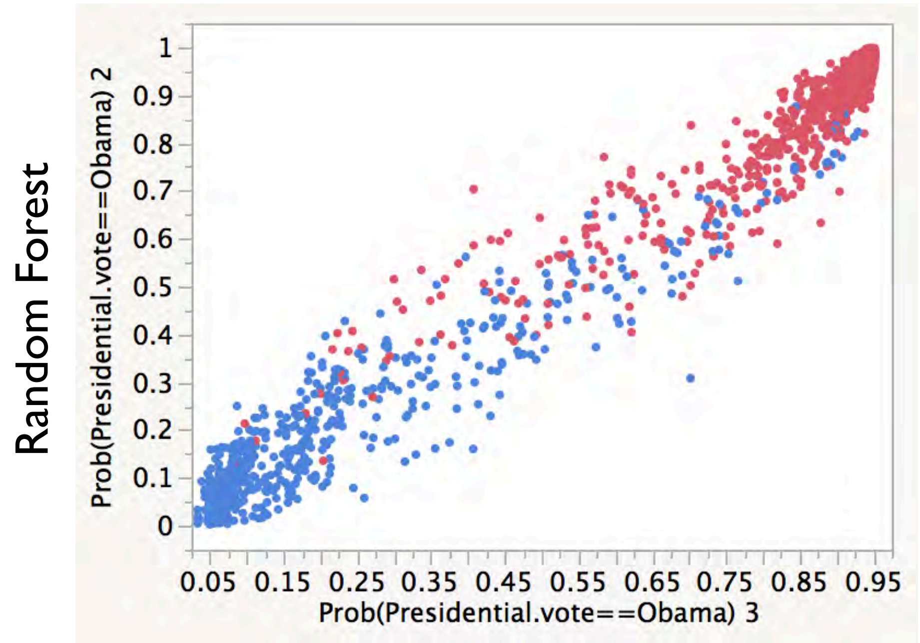
- Results for test sample with boosting
- Similar benefits obtained by forest
  - Boosting is a bit more predictive





Comparison of Predictions

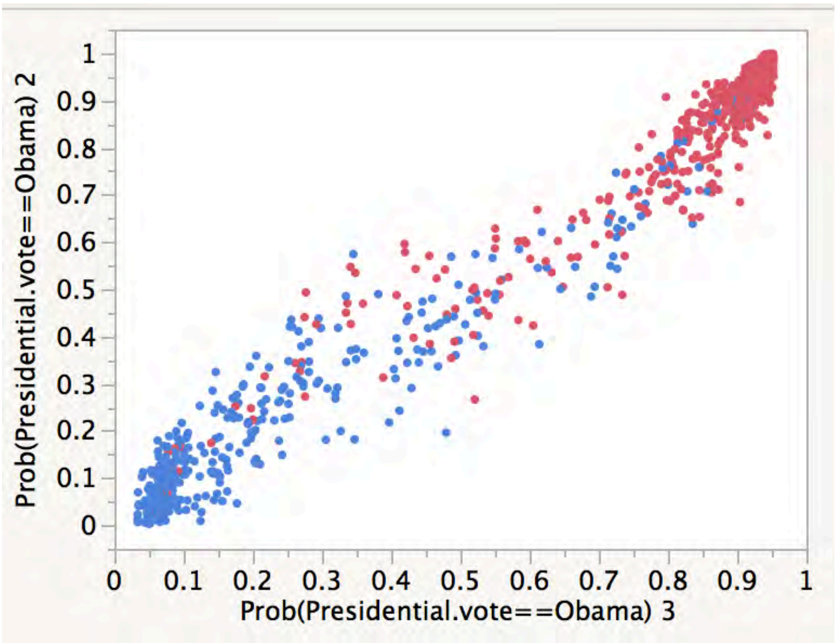
Training Sample



Boosted Tree

r=0.99

Test Sample



Boosted Tree

r=0.99

Wharton
Department of Statistics

Take-Aways

- Classification and regression trees
  - Partition cases into homogeneous subsets
    Regression tree: small variation around leaf mean
    Classification tree: concentrate cases into one category
  - Greedy, recursive algorithm
    Very fast
  - Flexible, iterative implementation in JMP
    Also found in several R packages (such as ‘tree’)
- Model averaging
  - Boosting, bagging smooth predictions
  - Borrow strength
- Over-fitting
  - Control with cross-validation
  - Analogous to use of CV in tuning Neural Net

Some questions to ponder...

- How does a tree indicate the presence of an interaction between factors?
- What does it mean when a tree splits many times on the same variable?

  How might you remedy this problem?
- Why is it important (at least 2 reasons) to avoid categorical variables with many categories in trees?
- What does it mean to describe a tree as defined by recursive and binary cuts?

  Why do it this way?

Next Time

- Thursday
  - Newberry Lab day for nets and trees
- Friday
  - Kernel methods and random projection
  - Text mining
  - Comparisons and summary