Recent Developments

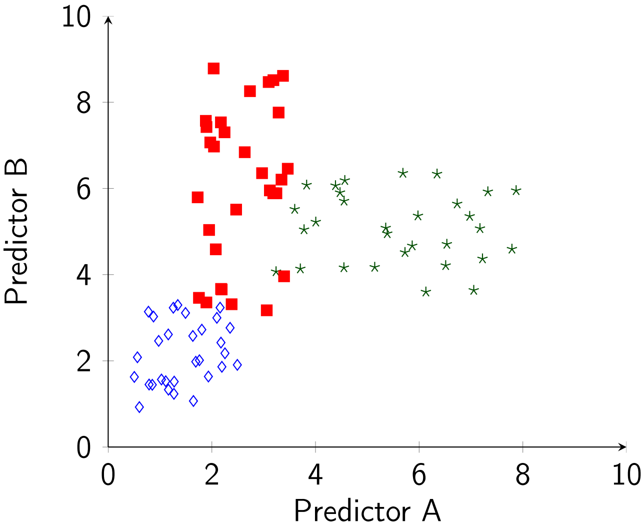
Nooreen Dabbish, Daqian Huang, and Shinjini Nandi

April 20, 2015

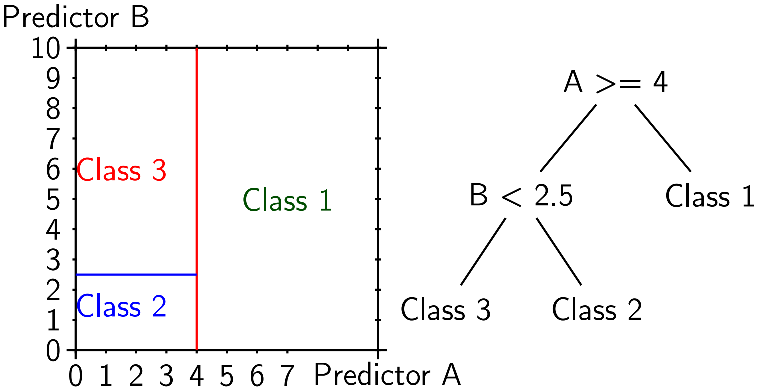
Outline

Bagging and Classification
- Classification Background
- Bagging basics
- Bagging in action
- Conclusions

Our two-predictor dataset



A classification scheme



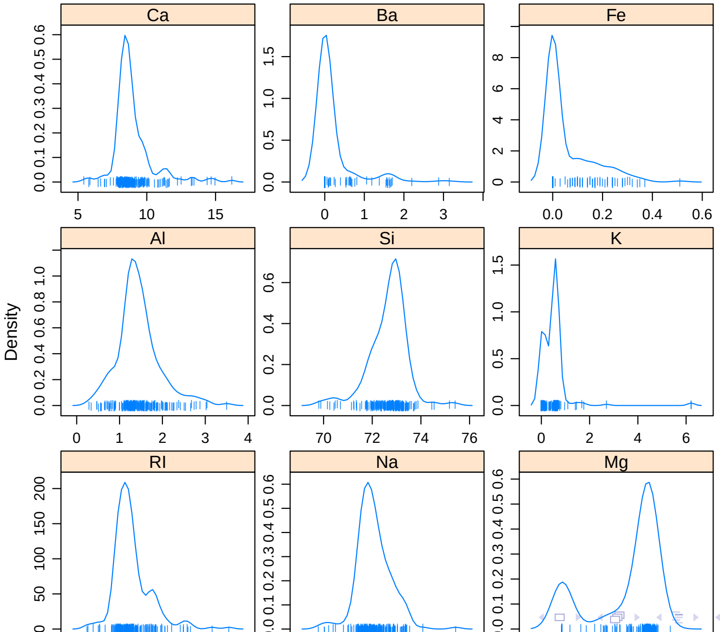
- Classification and Regression Tree (CART) method of Breiman et al. (1984)
- Starts with initial dataset and finds split value to minimize SSE

Glass dataset: introduction

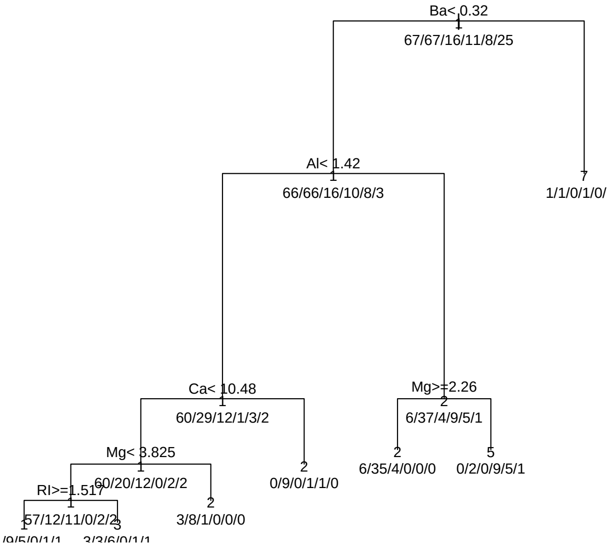
- One of the datasets analyzed in Breiman (1996)
- UCI Machine Learning repository

from help(Glass)
Description:
A data frame with 214 observation containing examples of the chemical analysis of 7 different types of glass. The problem is to forecast the type of class on basis of the chemical analysis. The study of classification of types of glass was motivated by criminological investigation. At the scene of the crime, the glass left can be used as evidence (if it is correctly identified!).

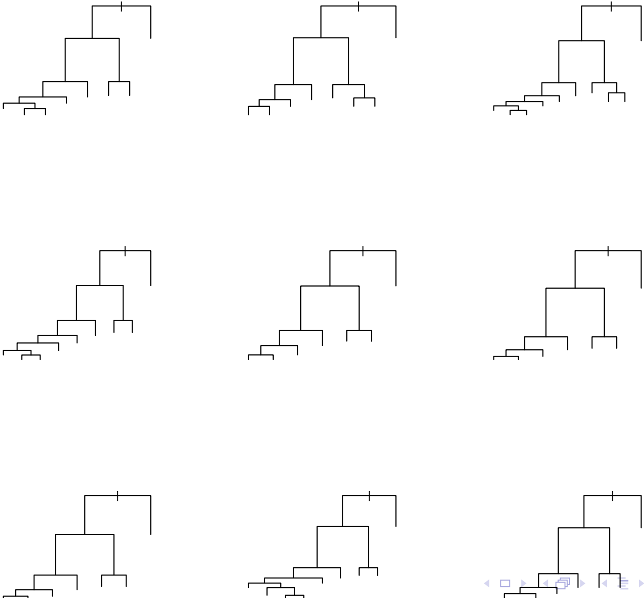
Glass dataset: preliminary analysis



Glass dataset: single classification tree



Different Learning Sets Create Different Trees



Bagging definition

- "bagging" bootstrap aggregation
  uses resampling as a smoothing device
  - prediction and nonparametric classification problems
  - useful when basic algorithm is unstable after small data perturbations
- Notation
  - data $d = \{(x_j, y_j), j = 1, \ldots, n\}$
  - response y numerical or class
  - predictor $x \equiv (x^{(1)}, \ldots, x^{(p)})$
  - predictor formula $m_0(x|d_n)$

Bagging overview

- empirical bagged predictor, acts as asmoother which reduces variance.
- can reduce MSE of predictor by 50%
- bagging by voting: pick the class which is choosen most often in R resamples
- "boosting" attaches weights to the data according to difficulty clasifying.

Bagging schematic

1. Resample the data: $d \to d_1^\star, \ldots, d_R^\star$
2. Construct predictors: $m_0(x|d_1^\star), \ldots, m_0(x|d_R^\star)$
3. Bagged predictor is an average:

$$\hat{m}_B(x|d) = \frac{1}{R}\sum_{r=1}^{R} m_0(x|d_r^\star)$$

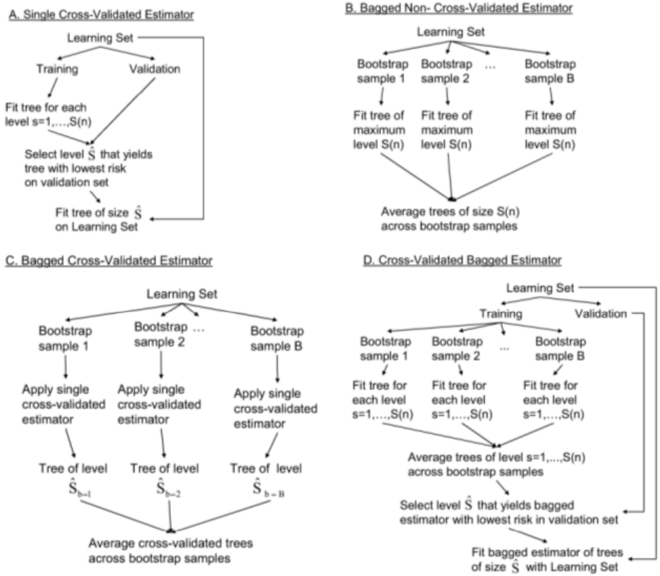
Approximates $m_B(x|d) = E^\star\{m_0(x|D^\star)\}$.

Classification scheme stability (Breiman 1994)

Stable	Unstable
k-nearest neighbor	Neural nets
	Classification trees
	Regression trees
	Subset selection in linear regression

- Bagging is useful in unstable cases
- Typical number of bootstrap replicates is ~ 50

Cross-Validation (Petersen et al. 2008)



Misclassification rate reduction with bagged classification trees (Breiman 1996)

Table 1 Missclassification Rates (Percent)

Data Set	$\bar{e}_S$	$\bar{e}_B$	**Decrease**
waveform	29.0	19.4	33%
heart	10.0	5.3	47%
breast cancer	6.0	4.2	30%
ionosphere	11.2	8.6	23%
diabetes	23.4	18.8	20%
glass	32.0	24.9	22%
soybean	14.5	10.6	27%

Screening of predictor variables

- Hard thresholding example
  - linear regression formula with screening of predictor variables

$$m_0(x|d) = \sum_{i=1}^{p} \hat{\beta}_i I(|\hat{\beta}_i| > c_i) x^{(i)}$$

- Bagged predictor does "soft thresholding."

$$m_B(x|d_n) = \sum_{i=1}^{p} E^\star\{\hat{\beta}_i^\star I(|\hat{\beta}_i^\star| > c_i)\} x^{(i)}$$

Buhlmann and Yu (2002)

- Bagged indicator example shows how bagging converts a hardthresholding desicion to softthresholding (smoothens)

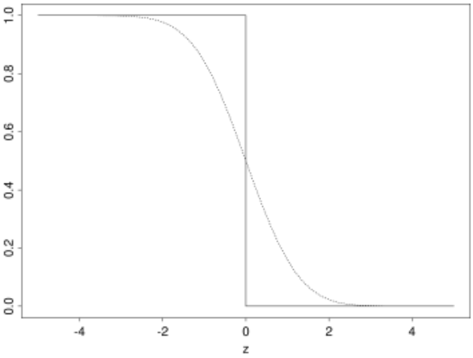


FIG. 1. *Indicator predictor from* (2.6) *at* $x = x_n(0)$ *as in* (2.2) *or Proposition* 2.1. *Function* $g(z) = \mathbb{1}_{[z \le 0]}$ *(solid line) and* $g_B(z)$ *(dotted line) define the asymptotics of the predictor and its bagged version* (*see Proposition* 2.1).

Advantages of Bagging

- Bagging is useful for unstable predictors
  - aggregation reduces variance and makes predictions more stable
- Useful for high-dimensional case
  - also shown to be successful in smaller problems (Buja and Stuetzle 2000)
- Out-of-bag: use samples that were not selected by bootstrap to measure predictive performance

Bagging Pitfalls

- Less interpretable than single model
- Poor classification predictors can become worse (Breiman 1996)
- Increase computational complexity/demand