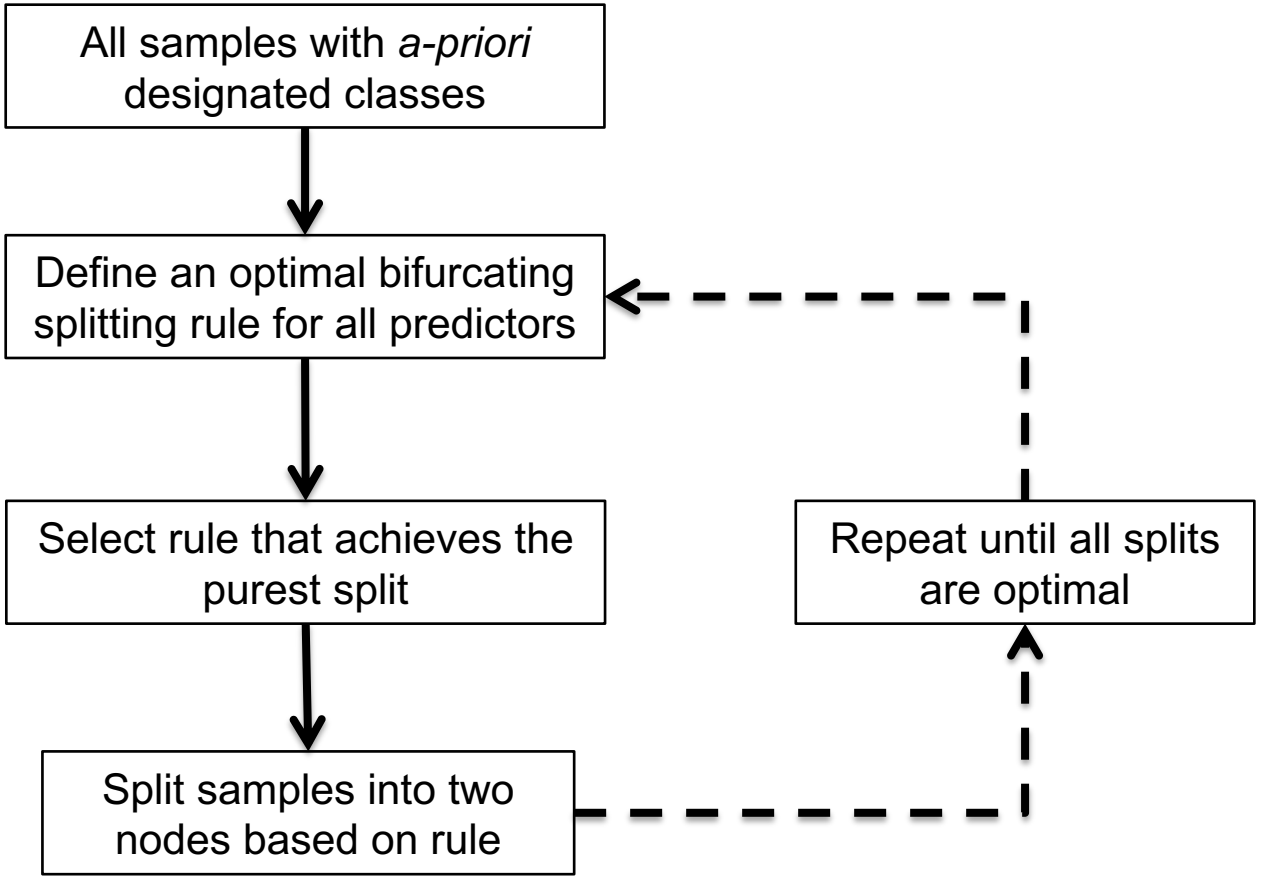
An Introduction to Random Forests

SIO R-Users Group
February 26, 2020

Eric Archer
Southwest Fisheries Science Center
eric.archer@noaa.gov
858-546-7121

Classification and Regression Trees (CART)

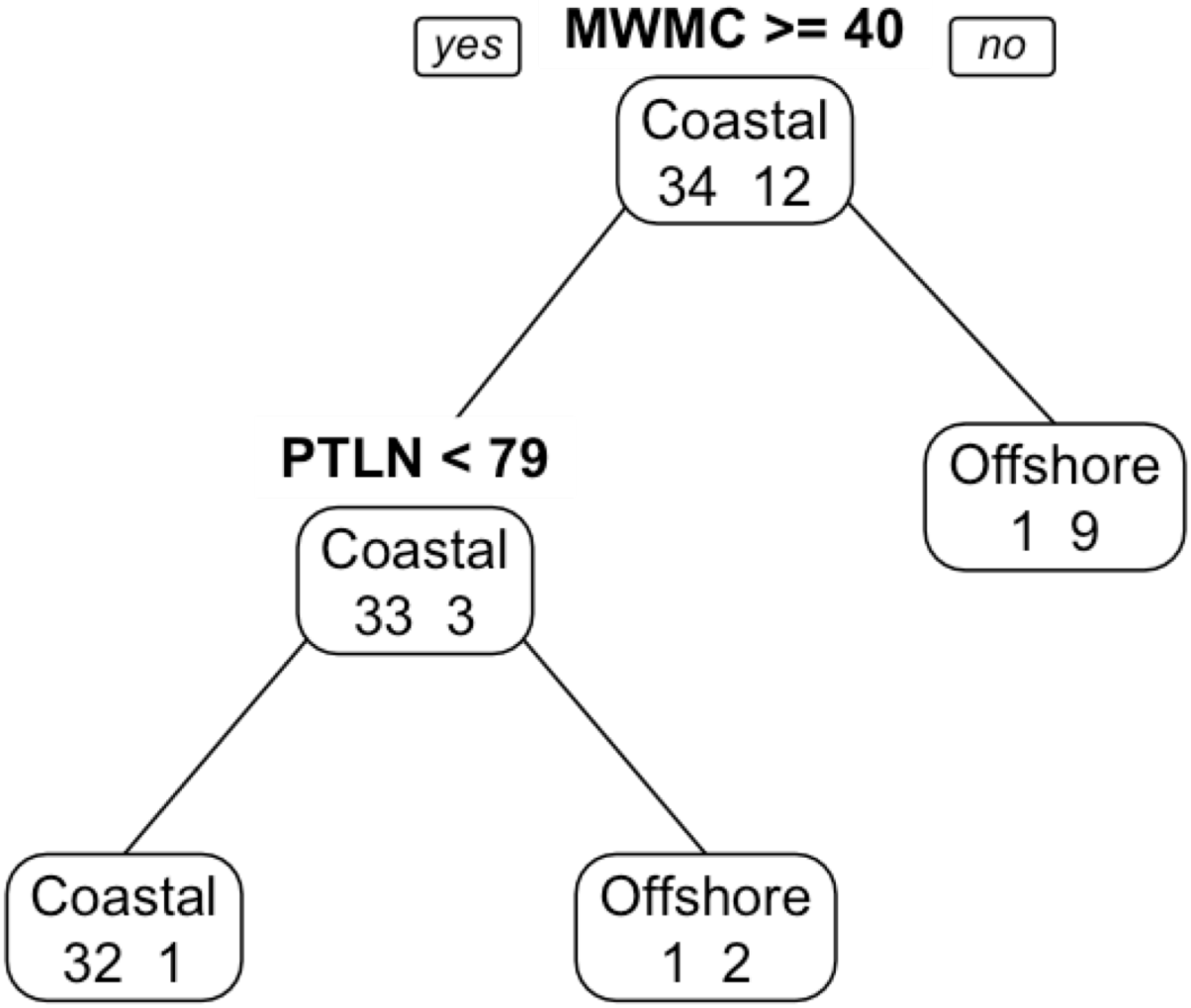
(Breiman et al. 1984; Therneau and Atkinson 1997; Hastie et al. 2009; rpart package)



CA Coastal & Offshore *Tursiops*

46 samples

29 cranial measurements



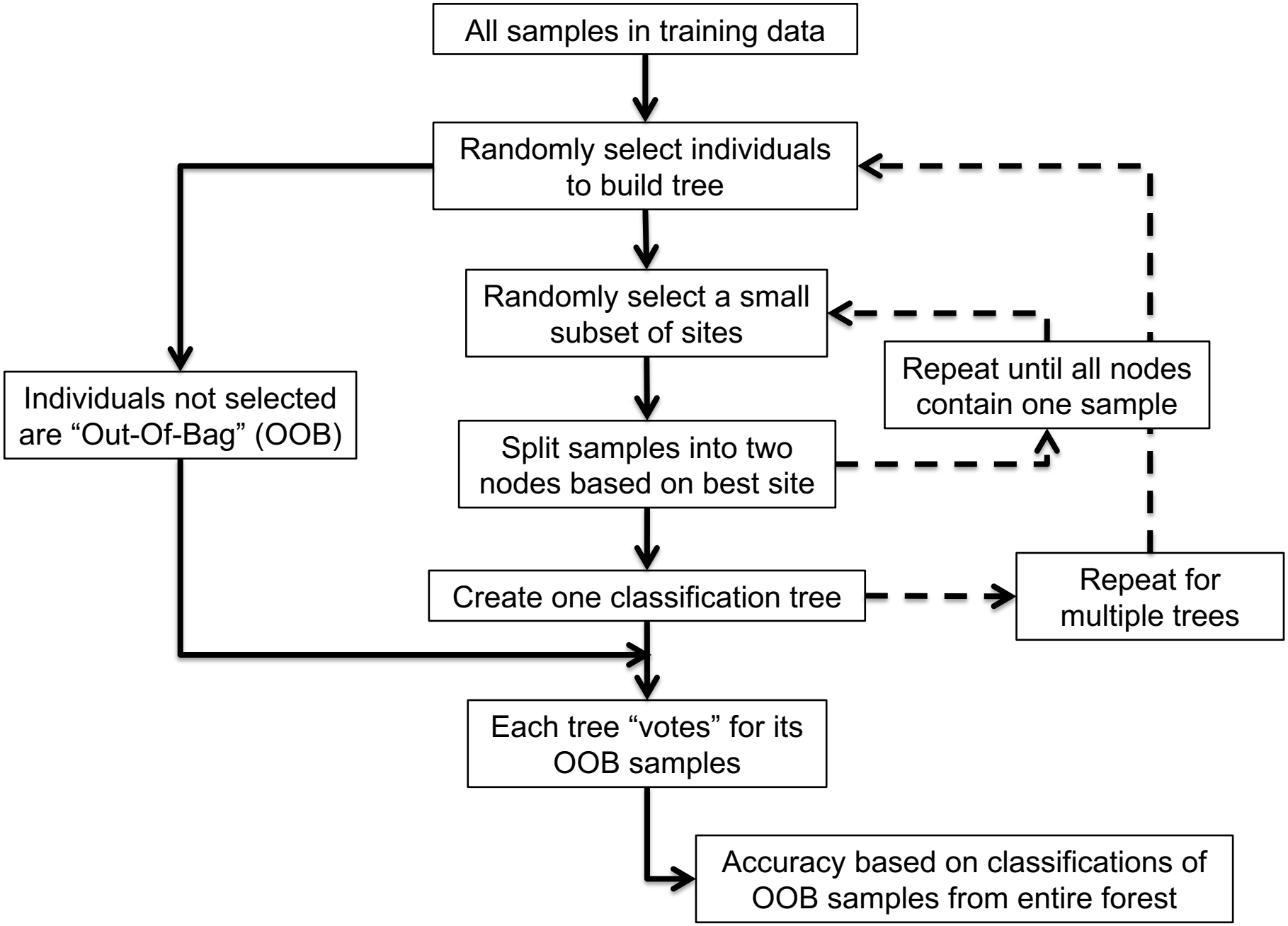
Disadvantages of CART

- Does not use combinations of predictors
- Tree can be deceptive – predictors can be “masked”
- Tree structures may be unstable – a change in the sample may give different trees
- Tree is optimal at each split, but may not be globally optimal
- Prediction accuracy of trees unknown without valid test data set

Random Forests (RF)

(Breiman 2001; Cutler et al. 2007; Berk 2008; Hastie et al. 2009)

- Ensemble method based on CART to create an algorithm to classify unknown samples.
- Is internally validated and produces by-class estimates of classification error.
- Identifies diagnostic characters.
- Uses of all types of data (continuous/discrete, ordered/un-ordered).
- Permits balancing of classification errors and weighting of classification probabilities based on prior knowledge.
- May be able to take advantage of large amounts of data collected on a number of weak, and possibly correlated, predictors
- Does not require a complete understanding of the processes behind the data.


All samples in training data
Randomly select individuals to build tree
Randomly select a small subset of sites
Split samples into two nodes based on best site
Repeat until all nodes contain one sample
Create one classification tree
Repeat for multiple trees
Individuals not selected are “Out-Of-Bag” (OOB)
Each tree “votes” for its OOB samples
Accuracy based on classifications of OOB samples from entire forest

OOB error rate

Fraction of *training* samples that were misclassified when they were OOB.

		Predicted			
		Stratum A	*Stratum B*	Total	OOB error rate
True	*Stratum A*	59	11	70	16%
	Stratum B	3	570	573	1%
	Total			643	Overall = 2%

Unknown assignment error

Fraction of trees in forest voting for a sample

Unknown	*Stratum A*	*Stratum B*
1	0.4	0.6
2	0.51	0.49
3	0.001	0.999
4	0.9	0.1

Disadvantages of Random Forests

- No easily visualized model
- Can overfit with very noisy data
- Variable importance can be biased towards correlated measures - see *cforest* for conditional implementation (Strobl et al 2008)

References

Berk, R.A. 2008. Statistical Learning from a Regression Perspective. Springer, New York.

Breiman, L. 2001. Random forests. Machine Learning 45: 5-32.

Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J. 1984. Classification and Regression Trees. Chapman and Hal/CRC.

Chen C., Liaw, A., Breiman, L., 2004. Using random forest to learn unbalanced data. Technical Report 666, Statistics Department, University of California Berkeley.

Cutler, D.R., Edwards, T.C., Beard, K.H., Cutler, A., Hess, K.T., Gibson, J., Lawler, J.J. 2007. Random forests for classification in ecology. Ecology 88:2783–2792.

Hastie, T., Tibshirani, R. Friedman, J. 2009. The Elements of Statistical Learning, Second Edition. Springer.

Liaw, A. and Wiener, M. 2002. Classification and Regression by randomForest. R News 2(3), 18—22.

Shi, T. and Horvath, S. 2006. Unsupervised learning with random forest predictors. Journal of Computational and Graphical Statistics 15: 118-138.

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).

Strobl, C., Boulesteix, A-L., Kneib, T., Augustin, T., and Zeileis, A. 2008. Conditional variable importance for random forests. BMC Bioinformatics 9:307.


Questions?