Tree-based methods

5 Mar 2023

Table of contents

Tree-based methods 1
Classification and regression trees 2
CART: machinery 2
tree-splitting rule complexity 2
complexity pruning 2
categorical predictors 2
loss matrix 3
missing predictor variables 3
linear combination splits 3
spam example 3
MARS 4
random forests 4
boosting 4

Tree-based methods

- trees are a basic building block of modelign methods ($\sim$ linear regression)
- **greedy** partitioning of parameter space
- efficient updating rules instead of linear algebra
- better at categorical predictors, interactions, missing data
- bias-variance tradeoff, curse of dimensionality, need for hyperparameter tuning … **still apply**

Classification and regression trees

- recursive *binary* splitting
- builds basis of rectangular regions
  - predictions homogeneous within regions
  - could be expressed as indicator variables

CART: machinery

- splitting rule
  - regression: improve SSQ, deviance, …
  - improve misclassification error, Gini coefficient, deviance
  - Gini coefficient: $\sum_k \hat{p}_{mk}(1-\hat{p}_{mk})$ (weighted average $(1-p)$ loss)
  - deviance: $\sum \hat{p}_{mk}(-\log \hat{p}_{mk})$ (weighted average $-\log$ loss)

tree-splitting rule complexity

- only $O(Np)$!
- (more specifically $\sum(\#\text{unique } x_i)$)
- splits only happen at data point values

complexity pruning

- $1/N_m \sum_{x \in R_m} (y_i - \bar{y}_m)^2 + \alpha|T|$
- boils down to (total loss) + α size
- weakest-link pruning (greedy again): collapse least-useful splits first

categorical predictors

- to avoid combinatorial splitting problems, order categories by
  - frequency falling in outcome 1 (binary output)

  - mean response value
  - optimal split for Gini/deviance/cross-entropy/L2 loss
  - multicategory harder
- favors categorical vars with many categories ("such variables should be avoided" … ???)

loss matrix

- allow weighting of misclassification
- e.g. cost of false positive/negative, **or** value of sensitivity/specificity

missing predictor variables

- 'missing' category
- use **surrogate variables** (algorithm?? effects of other splitting variables are already computed?)
- is imputation better?

linear combination splits

- can do generalized discriminant analysis at each split
- weights, split point for $\sum a_j X_j \leq s$
- seems better (Loh and Vanichsetakul 1988) but Breiman and Friedman disagree (Breiman and Friedman 1988)
- highly empirical!

spam example

- 4600 messages, 57 predictors (48 word percentages; punctuation percentages; sequences of capitals)
- earlier: misclassification 7.6% from logistic regression, 5.5% from GAM
- CART: 9.3%
- weighted tree does slightly better at high specificity, but still $\ll$ GAM …

Loh, Wei-Yin, and Nunta Vanichsetakul. 1988. "Tree-Structured Classification via Generalized Discriminant Analysis." *Journal of the American Statistical Association* 83 (403): 715–25. https://doi.org/10.1080/01621459.1988.10478652.

Breiman, Leo, and Jerome H. Friedman. 1988. "Tree-Structured Classification Via Generalized Discriminant Analysis: Comment." *Journal of the American Statistical Association* 83 (403): 725–27. https://doi.org/10.2307/2289296.

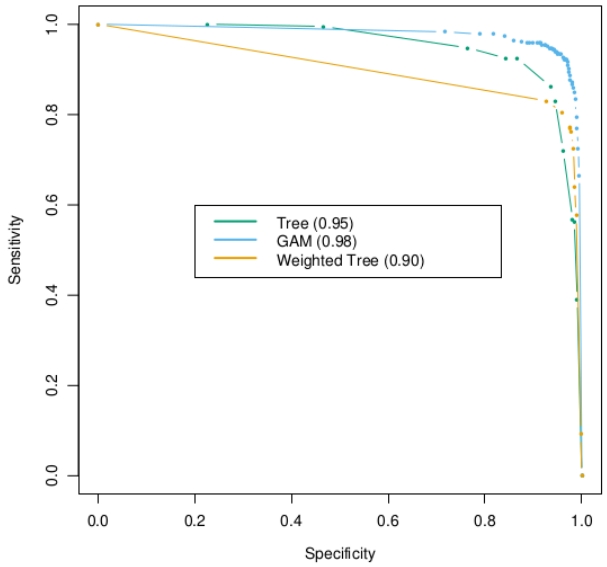

FIGURE 9.6. *ROC curves for the classification rules fit to the* `spam` *data. Curves that are closer to the northeast corner represent better classifiers. In this case the GAM classifier dominates the trees. The weighted tree achieves better sensitivity for higher specificity than the unweighted tree. The numbers in the legend represent the area under the curve.*

MARS

random forests

boosting