Choice in the Number of Divisions or Leaves

The smaller Gini's index or the cross entropy, the better.

The classification error rate is often considered not sensitive enough.

The algorithm is applied until

- ✓ a predetermined number of leaves is reached;
- ✓ the number of observations per leaf is not "too small."

"Not enough leaves" is a sign of lack of fit: the classification is too simple or inappropriate.

"Too many leaves" is no better because overfitting often results in the classification tree performing poorly on validation data.

Pruning Classification Trees

A common strategy consists in letting the tree grow and then to prune it in order to reach a compromise between the degree of adjustment to the training data and the dimension $|T|$ of the tree T, measured by its number of leaves.

This compromise is obtained by minimizing

$$\sum_{j=1}^{|T|} c_j + \alpha\,|T|,$$

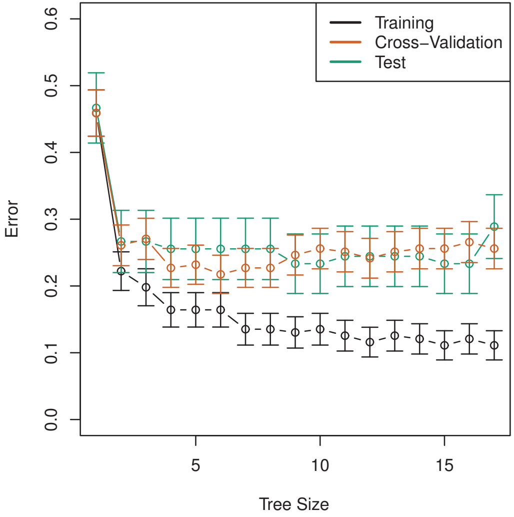
where $\alpha \geq 0$ is a hyperparameter which controls the compromise selected by cross validation.

k-Fold Cross Validation

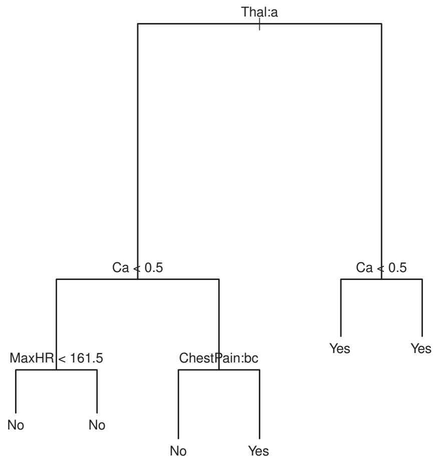
1. Remove n/k observations and adjust the model with the remaining $n - n/k$ observations.
2. Predict the n/k observations initially removed and measure the classification error.
3. Put the n/k removed observations back in the dataset and start again with another set of n/k observations.
4. Repeat Steps 1–3 until all the observations have been removed and predicted once; add the prediction errors.
5. Repeat Steps 1–4 for several values of α, which amounts to do this for all possible tree sizes) and choose α so as to minimize the error.

(or choose the smallest tree for which the error is "almost minimal.")

Error Rate as a Function of the Method Used



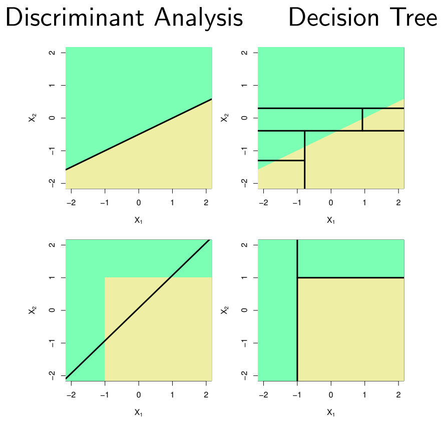
Optimal Tree for the Health Data



Advantages of Classification Trees

- ✓ No prior assumption on the data distribution
- ✓ Robust to outliers
- ✓ Easy to interpret
- ✓ Takes into account possible interactions between variables
- ✓ Select the important variables implicitly
- ✓ Lead to non-linear models

No Model is Ideal



Disadvantages of Classification Trees

- ✓ Require a large number of observations
- ✓ Results tend to be unstable
- ✓ Computationally demanding algorithms

Possible Improvements to Classification Trees

When dealing with large datasets, it is observed empirically that there are advantages to

combining the predictions of several classification rules, even weak ones.

An ensemble method consists in basing a single prediction from those given by several different models.

This computer-intensive technique results in complex classification rules.

Choice of Combination Rule

There are many ways of combining predictions from various models, e.g.,

- ✓ average of the probabilities prescribed by the various models;
- ✓ majority rule;
- ✓ "variable selection" type of approach similar to what is done in regression, but for the purpose of selecting models, *not* variables;
- ✓ bagging, boosting, random forests.

Bagging (1–2)

Bagging consists in

1. drawing B bootstrap samples;
2. constructing a classification tree for each of them;
3. using the B trees to predict the class to which a new observation should be assigned;
4. assigning the new observation to the most often predicted class.

Bagging (2–2)

For Step 3, i.e., predict a “new” observation, there is no need to resort to a validation set.

Instead, one can call instead on “Out of Bag” (OOB) estimation.

For each of the B bootstrap classification trees, the idea is to predict the observations that were *not* selected in the bootstrap sample.

Each observation is then assigned to the class that was *most often* predicted for it.

Once this is done, it only remains to compute the classification error rate.

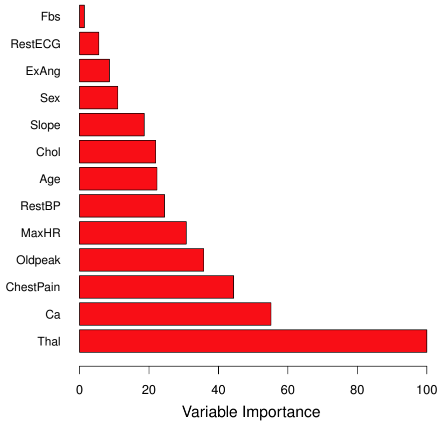
Interpretation of “Bagging”

Bagging increases predictive stability. However, it becomes more difficult to interpret the importance of each variable in the classification process.

One possible way to proceed is as follows:

1. add the reduction in Gini’s index over all divisions based on a given variable;
2. compute the average of this reduction over all trees;
3. express the average reduction for each variable as a percentage of the maximum average reduction observed over the set of all variables.

Example (cont'd)



Random Forest Methodology

This methodology is very similar to bagging. It also considers B classification trees derived from B bootstrap samples.

The idea is the following:

- ✓ when constructing each of the B trees, choose at random $m < p$ of the variables $X_1, \ldots, X_p$ before each division;
- ✓ these m variables are the only ones considered for the optimal division.

A frequent choice is to take $m \approx \sqrt{p}$.

The purpose of this approach is to reduce correlation between the trees. Indeed, the B trees derived by standard bagging are often very similar.

Boosting

Boosting is designed to create successive trees by giving more weight to observations that were poorly classified in previous trees.

It is important to use simple and poorly performing trees and to let boosting choose the weights to be allotted to the different trees.

There are many variants to this approach, e.g.,

- ✓ adaptative boosting (AdaBoost);
- ✓ gradient boosting.

Fun fact: Gradient boosting implemented via algorithm XGBoost (eXtreme Gradient Boosting) is one of the strategies most often used by winners of "Kaggle" type competitions in recent years.

Investigate Work (1–2)

No algorithm guarantees the best classifier in every case.

All problems are different and in each case, one must find the best way to proceed by trial and error.

Nevertheless, certain general principles apply, e.g.,

- ✓ start with exploratory data analysis (e.g., descriptive data analysis on each variable, PCA, CA, clustering, etc.);
- ✓ take avantage of whatever expert knowledge may be available;
- ✓ check whether some methods have been particularly successful in previous similar analyses.

Investigate Work (2–2)

✓ The main challenge is often the dimension of the problem: the number of possible models/methods is humongous and grows rapidly with the number of variables available.

✓ It is sometimes helpful to reduce the dimension of the problem, e.g., via PCA, CA, unsupervised classification: the techniques are applied to a subset of variables and the resulting scores serve as predictors for the classification algorithms...

✓ There is no magic recipe for the choice of subset...

✓ Check on data analysis competition websites (e.g., Kaggle, KD Nuggets, etc.) and look at the approaches used by the winners of problems similar to yours...

Ensemble Methods

- ✓ When a large sample is available (sample size n in the tens of thousands), boosting and similar methods generally work well.
- ✓ It is important to bind together several weak classifiers in boosting or bagging.
- ✓ One can resort to a majority vote or a weighted average of 2–3 strong classifiers to try and improve the model.

Statistics, Machine Learning... or Both? (1–3)

- ✓ Statistical methods usually lead to interpretable and meaningful classifiers (one can tell how each variable affects the chances of falling in one class or another).
- ✓ Statistical methods allow one to adjust predictions to situations where the sampling does not entirely reflect the target population (e.g., through the introduction of sampling weights).
- ✓ Statistical methods are not very good at handling situations involving a very large number of variables and observations.
- ✓ Statistical methods are best in situations where in-depth analysis is possible, e.g., using *prior* knowledge.

Statistics, Machine Learning... or Both? (2–3)

- ✓ Machine learning algorithms lead to good classification rules even when the underlying class formation phenomenon is very complex. However, one must have sufficient training data.
- ✓ If such is the case, *prior knowledge* of the phenomenon to be modeled is not as crucial as for classical statistical methods.
- ✓ The model predictions are only valid if the training sample is representative of the population for which they were designed.
- ✓ Although one can identify which variables play an important role in classification, it is usually difficult to interpret their effect properly.

Statistics, Machine Learning... or Both? (3–3)

In practice, we usually face a mixture of all these situations.

Data scientists are therefore people who are proficient with statistical methods and machine learning methods!