Introduction to Classification and Regression Trees (CART)

Ryan Miller

Motivating example (well-switching)

- In the early 2000s a team of US and Bangladeshi scientists measured the arsenic levels of wells in Arahazar upazila, Bangladesh

Motivating example (well-switching)

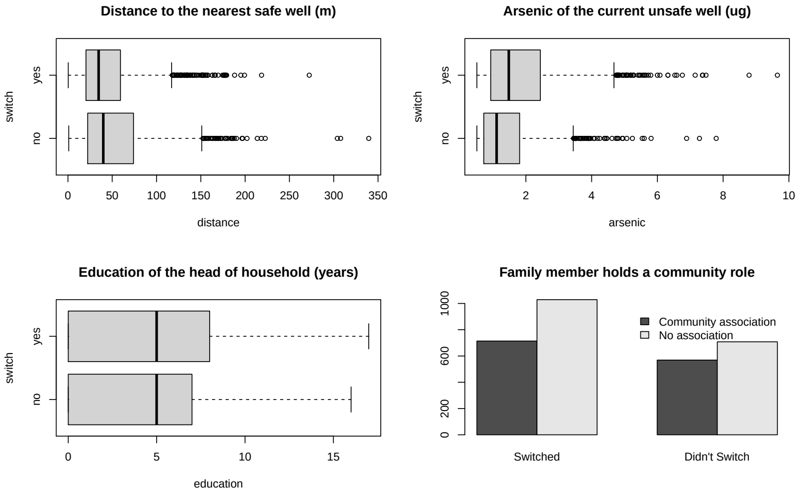
- In the early 2000s a team of US and Bangladeshi scientists measured the arsenic levels of wells in Arahazar upazila, Bangladesh
- Wells with arsenic concentrations $\geq 0.5\ \mu g$ were deemed "unsafe", and families using them were given information encouraging them to switch to a "safe" well

Motivating example (well-switching)

- In the early 2000s a team of US and Bangladeshi scientists measured the arsenic levels of wells in Arahazar upazila, Bangladesh
- Wells with arsenic concentrations $\geq 0.5\ \mu g$ were deemed "unsafe", and families using them were given information encouraging them to switch to a "safe" well
- After several years, researchers returned to the region and determined which of these families had actually switched
  - Their goal was to understand how various factors influenced the likelihood of a family switching from an unsafe well to a safe one

Motivating example (well-switching)

We'll consider four explanatory variables:



Initial analysis

1. How does study design impact our approach?

Initial analysis

1. How does study design impact our approach? These are observational data, so any bivariate relationships might be confounded
2. What existing models/methods might be appropriate?

Initial analysis

1. How does study design impact our approach? These are observational data, so any bivariate relationships might be confounded
2. What existing models/methods might be appropriate? The outcome is binary (switched or didn't switch), so *logistic regression* is a good choice

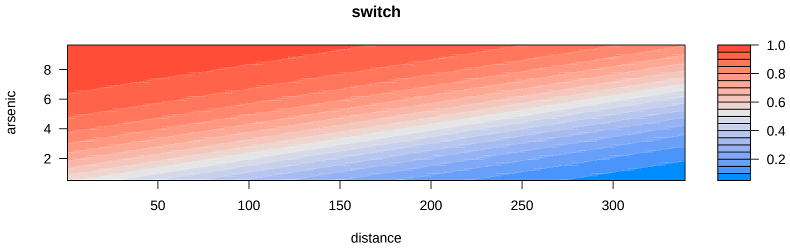
Logistic regression

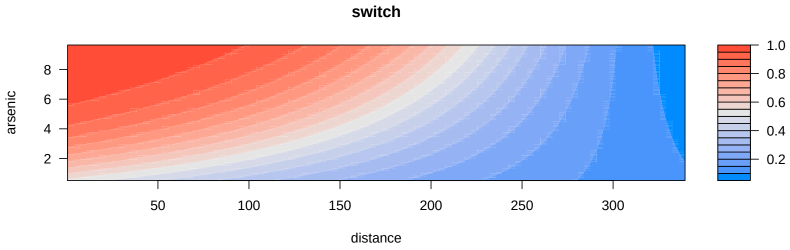
Consider two logistic regression models:

1. `switch ~ distance + arsenic`
2. `switch ~ distance + arsenic + distance:arsenic`

The plots on the next slide are created using the `visreg` package and they display each model's *estimated probability* of a switch (redder = higher chance of switching)

Logistic regression





Which plot corresponds to which model?

Logistic regression

A likelihood ratio test can help us decide whether the interaction term is useful:

```
library(lmtest)
m1 <- glm(switch ~ distance + arsenic, data = Wells, family = "binomial")
m2 <- glm(switch ~ distance*arsenic, data = Wells, family = "binomial")
lrtest(m1, m2)
```

```
## Likelihood ratio test
##
## Model 1: switch ~ distance + arsenic
## Model 2: switch ~ distance * arsenic
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   3 -1965.3
## 2   4 -1963.8  1 3.0399    0.08124 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- We see *borderline evidence* in favor of the interaction
- However, the interaction would substantially complicate how we interpret the effect of each variable within the model...

Alternatives

Classification and regression trees (CART) is an algorithmic modeling approach that is well-suited for applications involving interactions between variables

1. We'll first walk through an example introducing the method on Fisher's iris data
2. I'll give you an opportunity to apply the method yourself to the well-switching data

The CART Algorithm

The CART algorithm involves a procedure called *recursive binary splitting*:

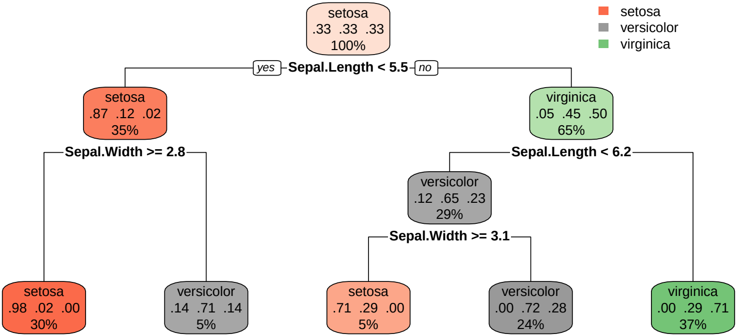
1) Starting with a "parent" node, search for a binary splitting rule that optimizes the *purity* of the "child" nodes
2) Check stopping criteria, if they aren't yet satisfied execute the split found in Step #1, then designate the resulting child nodes as parents and return to Step #2.

We'll get into the details in a few moments, but it's useful to see a fitted CART model first

CART example (Fisher's iris data)

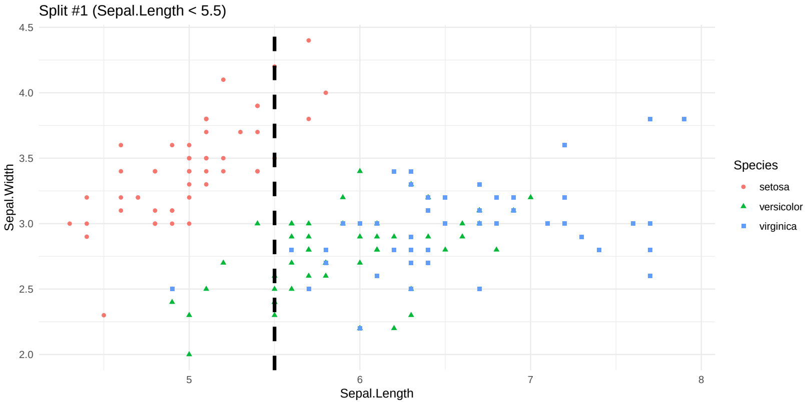
Fisher's iris dataset contains three species (setosa, versicolor, and virginica), the CART model below predicts species using sepal length and sepal width:

```r
fittedTree <- rpart(Species ~ Sepal.Length + Sepal.Width, data = iris)
rpart.plot(fittedTree, extra=104) # "extra = 104" adds some info to the tree
```

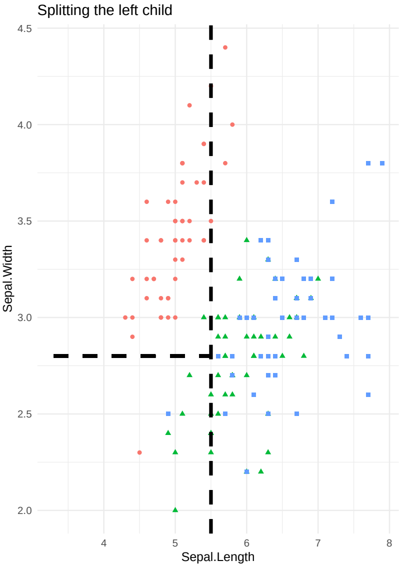


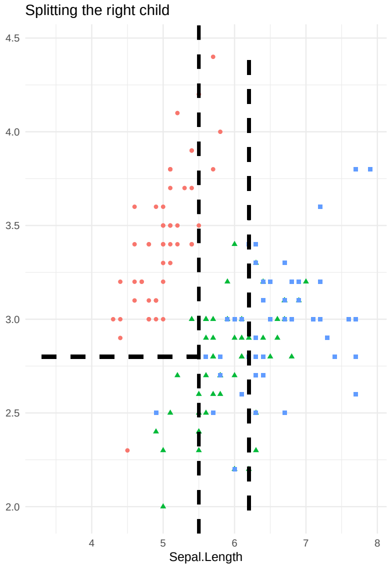
CART example (Fisher's iris data)

In this example it's easy to step through the process of recursive binary splitting:



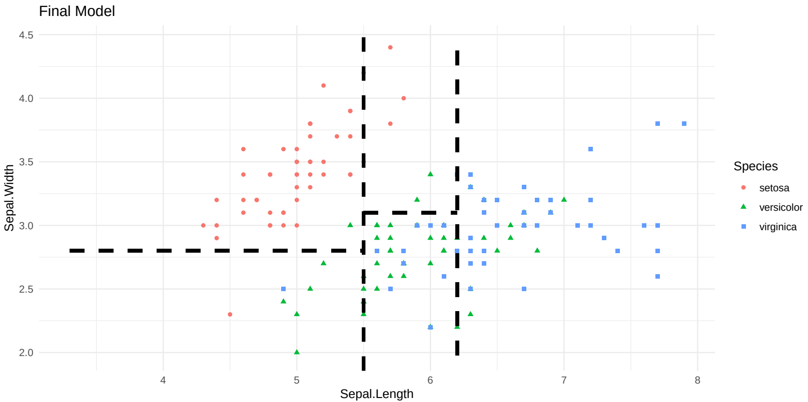
CART example (Fisher's iris data)





CART example (Fisher's iris data)

Shown below is the final model:



- The model's predicted probabilities are the proportion of each species in each partition
  - For example, an iris in the "upper left" partition is predicted to have a 98% chance (45/46) of being a setosa

Detail #1 - finding the optimal split

There are dozens of ways to quantify the *purity* of a node, but a few popular ones are:

- *Change in Gini Index*
  - Gini Index $= \sum_k p_k(1 - p_k)$ where p_k is the proportion of in category k (higher Gini = lower purity)
  - Gini Improvement = Gini Index for parent node minus a weighted average of the Gini Index for the two child nodes
- *Information Gain*: A more sophisticated theoretical construct that compares the divergence of two probability distributions

See: Breiman, L., J. H. Friedman, R. A. Olshen, and C. J. Stone. 1984. Classification and Regression Trees

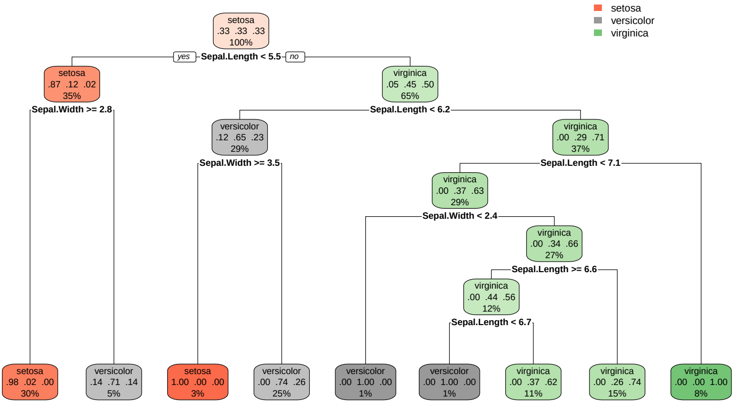
Detail #2 - deciding when to stop

Generally there are two important factors in determining when the CART algorithm stops:

1) The *complexity parameter*, *cp*, which defines a minimum relative improvement in purity that must be achieved in order for a split to be considered "worthwhile" (1% is the default in `rpart`)
2) The *minimum node size*, the minimum number of data-points that must belong to a node for it to be deemed eligible for splitting (20 is the in `rpart`)

Detail #2 - deciding when to stop

```
fittedTree2 <- rpart(Species ~ Sepal.Length + Sepal.Width, data = iris,
                     control = rpart.control(cp = 0.008, minsplit = 6))
rpart.plot(fittedTree2, extra=104)
```



Detail #3 - is CART better than logistic regression?

Ideas for how to compare CART and logistic regression?

- How about a statistical test, such as the likelihood ratio test?

Detail #3 - is CART better than logistic regression?

Ideas for how to compare CART and logistic regression?

- How about a statistical test, such as the likelihood ratio test?
  - No, the models aren't nested.

Detail #3 - is CART better than logistic regression?

Ideas for how to compare CART and logistic regression?

- How about a statistical test, such as the likelihood ratio test?
  - No, the models aren't nested.
- How about a model selection criterion like AIC or BIC?

Detail #3 - is CART better than logistic regression?

Ideas for how to compare CART and logistic regression?

- How about a statistical test, such as the likelihood ratio test?
  - No, the models aren't nested.
- How about a model selection criterion like AIC or BIC?
  - No, the CART model doesn't involve a likelihood.

Detail #3 - is CART better than logistic regression?

Ideas for how to compare CART and logistic regression?

- How about a statistical test, such as the likelihood ratio test?
  - No, the models aren't nested.
- How about a model selection criterion like AIC or BIC?
  - No, the CART model doesn't involve a likelihood.
- How about metrics like classification accuracy, AUC, or Cohen's kappa?

Detail #3 - is CART better than logistic regression?

Ideas for how to compare CART and logistic regression?

- How about a statistical test, such as the likelihood ratio test?
  - No, the models aren't nested.
- How about a model selection criterion like AIC or BIC?
  - No, the CART model doesn't involve a likelihood.
- How about metrics like classification accuracy, AUC, or Cohen's kappa?
  - Yes, but we should be careful to use *cross-validation* so we do not reward overfitting

Your turn

- The CART lab document begins by providing examples of several functions used to fit and visualize CART models
- Next, the lab includes a link and data dictionary for the well-switching data
- Your task to work through the subsequent questions that will guide you through an introductory CART analysis of the well-switching data
  - If you have extra time, there's a challenge question
  - Be prepared to share your answers and don't be afraid to ask questions