Non-linear Regression with Decision Trees and $k$NN
Decision Trees: A Review

To build the tree:
- Start with all observations at the root of the tree
- Score all remaining variables using the current set of observations
- Split the current set of observations based on the variable with the best score
- Repeat until all groups of observations are pure enough, or until the tree is sufficiently deep

To classify:
- Classify a new observation by walking the tree according to its feature values
- Assign the new observation a class by majority vote, among all observations at the new observation’s leaf
Regression Trees

To build the tree:

- Start with all observations at the root of the tree
- Score all remaining variables using the current set of observations
- Split the current set of observations based on the variable with the best score
- Repeat until all groups of observations are pure enough, or until the tree is sufficiently deep

To classify:

- Classify a new observation by walking the tree according to its feature values
- Assign the new observation a value by averaging the values of all observations at the new observation’s leaf
Regression Trees: Scoring

- To build decision trees, we scored variables based on an impurity measure. The more they distinguished among classes, the better!
- To build regression trees, we will instead score variables based on their variance, and we will aim to minimize variance.

\[
\text{GOOD: } \text{Score} = \text{var}(1, 1.2) + \text{var}(2.2, 2.3) = 0.04
\]

\[
\text{BAD: } \text{Score} = \text{var}(1, 2.2) + \text{var}(1.2, 2.3) = 1.325
\]
Regression Trees in R

Pretty much the same as decision trees!

`rpart` will detect whether you’re predicting real values or categorical ones.

```r
> attach(iris)

> rpart(Sepal.Length ~ Sepal.Width + Petal.Length + Petal.Width)
```
Regression with \( k \text{NN} \)

Same ideas as decision trees: Instead of majority vote, average!

- Average the values of all the neighbors equally
- Or, assign greater weight to the values of closer neighbors
Unweighted

\[(3.2 + 1.3 + 0.5) / 3 = 1.6\]

Weighted

\[\frac{\left(\frac{1}{1.2} \cdot 3.2\right) + \left(\frac{1}{3} \cdot 1.3\right) + \left(\frac{1}{4.3} \cdot 0.5\right)}{\frac{1}{3} + \frac{1}{1.2} + \frac{1}{4.3}} = 2.3\]