KJ 8.2

Group 3

2024-07-03

# 8.2 Tree Granularity Bias

**Use a simulation to show tree bias with different granularity.**

Decision trees can develop biases based on the characteristics of the data they receive. One example of this is the granularity of the data. Granularity refers to the level of detail in the data. For example, a variable with low granularity might have only a few categories, while a variable with high granularity might have many categories.

Due to the nature of decision trees, if they receive data with low granularity, they might develop a bias towards variables with higher granularity. This is because the tree will be more likely to split on variables with more categories, as they provide more opportunities for the tree to find patterns. This can lead to the tree ignoring variables with low granularity, even if they are more informative.

Here, I will use a simulation to demonstrate how decision trees can develop biases based on the granularity of the data. I will create a dataset with two variables: one with low granularity and one with high granularity. I will then fit a decision tree to the data and visualize the splits made by the tree. By comparing the splits made by the tree on the two variables, I will show how the tree can develop a bias towards variables with higher granularity.

set.seed(1125)  
  
# Create a dataset with two variables: one with low granularity and one with high granularity  
sim <- data.frame(  
 low\_granularity = rep(1:10, each = 10),  
 high\_granularity = rep(1:50, each = 2),  
 y = rnorm(100)  
)  
  
# Fit a decision tree to the data  
tree <- rpart(y ~ low\_granularity + high\_granularity, data = sim)  
  
# Check the importance of the variables  
varImp(tree)

## Overall  
## high\_granularity 0.27893751  
## low\_granularity 0.07637142

The data generated above was created with no correlation between the y variable and either of the two predictor variables. However, the decision tree model developed a bias towards the high granularity variable, as evidenced by the higher importance score (0.28) compared to the low granularity variable (0.07). This bias is due to the higher granularity of the high\_granularity variable, which provided more opportunities for the tree to find patterns and make splits.

Next, I will recreate the dataset with the same number range for each of the variables as before but with the same granularity for both variables. I will then fit a decision tree to the data and check the importance of the variables to see if the tree develops a bias towards one of the variables.

set.seed(1125)  
  
# Recreate the tree with the same granularity for both variables  
sim\_same\_granularity <- data.frame(  
 low\_granularity = rep(1:10, each = 1),  
 high\_granularity = seq(1, 50, 10),  
 y = rnorm(100)  
)  
  
# Fit a decision tree to the data  
tree\_same\_granularity <- rpart(y ~ low\_granularity + high\_granularity, data = sim\_same\_granularity)  
  
# Check the importance of the variables  
varImp(tree\_same\_granularity)

## Overall  
## high\_granularity 0.05576726  
## low\_granularity 0.05908178

In this case, the importance of each variable was almost the same. This reinforces the idea that the previous ‘importance’ was merely a result of the granularity allowing it to pick up on false patterns in the noise of the data and not the actual informativeness of the variables.

## Conclusion

This simulation demonstrates how decision trees can develop a bias towards features with higher granularity, even when those features are not necessarily more informative. When presented with data containing features of varying granularity, the tree prioritizes splitting on features with more categories, leading to an underestimation of the importance of features with fewer categories. This highlights the importance of considering data granularity when interpreting decision tree models. It might be beneficial to explore techniques like feature scaling or binning to ensure features have comparable granularity before feeding them into a decision tree algorithm. Additionally, relying solely on variable importance scores from decision trees can be misleading. Complementary methods for feature selection, like correlation analysis, can help identify truly informative features regardless of their granularity.