Tree Based Methods

Machine Learning EBC 4257

Ivan Ricardo

Maastricht University

April 14, 2022

Outline

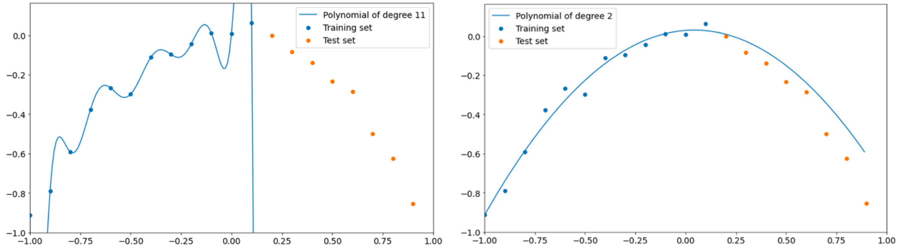
1. Bias/Variance Tradeoff
2. Cross Validation
3. Decision Trees
4. Impurity Measures
5. Bootstrapping
6. Spam Dataset

Bias/Variance Tradeoff

Bias/Variance Tradeoff

- Bias: the inability for a Machine Learning method to capture the true relationship of the variables
- Variance: How well the Machine Learning method adapts to testing data given the training data
- When we fit a model, there is typically a trade-off; high bias may not capture a relationship well but would better predict testing data (Hastie, Friedman, and Tisbshirani, 2017)

Overfitting



(Malato, 2020)

Cross Validation (CV)

- Used to compare different Machine Learning methods and get a sense of how well they perform in practice
- Leave-One-Out CV vs. K-fold CV

Cross Validation (CV)

- Used to compare different Machine Learning methods and get a sense of how well they perform in practice
- Leave-One-Out CV vs. K-fold CV
- LOOCV can be streamlined by calculating the leverage of the additional points, and this reflects the amount that an observation influences its own fit (James et al., 2013)

Decision Trees

- Regression vs Classification Trees

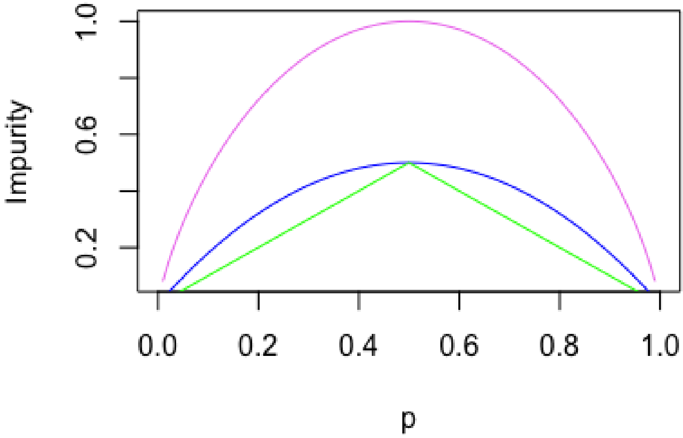
Decision Trees

- Regression vs Classification Trees
- If we have a dataset and we want to predict the presence of a disease given some symptoms, we can use a Classification Tree
- How do we know which variable to use as the root? branches? leaves?

Decision Trees

- Regression vs Classification Trees
- If we have a dataset and we want to predict the presence of a disease given some symptoms, we can use a Classification Tree
- How do we know which variable to use as the root? branches? leaves?
- We need measures for **impurity**

Impurity Measures



Indices

Misclassification Error

$$\frac{1}{N_m}\sum_{i\in R_m} I(y_i \neq k(m)) = 1 - \hat{p}_{mk(m)}$$

Gini index

$$\sum_{k\neq k'} \hat{p}_{mk}\hat{p}_{mk'} = \sum_{k=1}^{K} \hat{p}_{mk}(1-\hat{p}_{mk})$$

Cross Entropy

$$-\sum_{k=1}^{K} \hat{p}_{mk}\log \hat{p}_{mk}$$

What are the differences?

Problems with Decision Trees

Problems with Decision Trees

- When splitting a predictor having q possible unordered values, there are $2^{q-1} - 1$ possible partitions of the q values into two groups, and the computations become prohibitive for large q. Hastie, Friedman, and Tisbshirani, 2017

Problems with Decision Trees

- When splitting a predictor having q possible unordered values, there are $2^{q-1} - 1$ possible partitions of the q values into two groups, and the computations become prohibitive for large q. Hastie, Friedman, and Tisbshirani, 2017
- A loss matrix may be required if we value one classification more than another

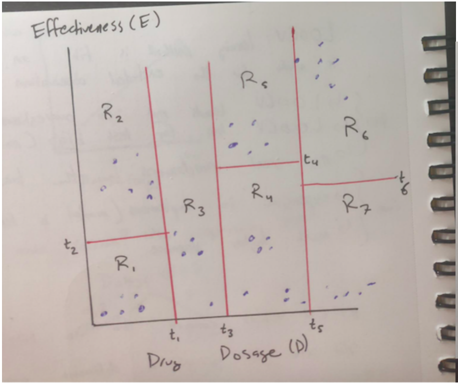
Problems with Decision Trees

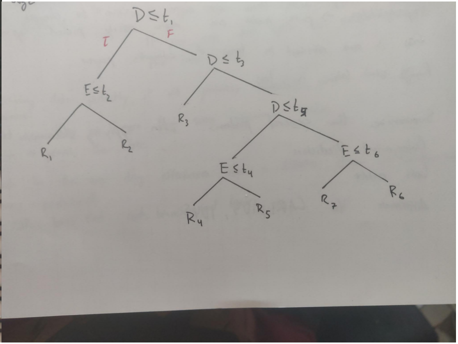
- When splitting a predictor having q possible unordered values, there are $2^{q-1} - 1$ possible partitions of the q values into two groups, and the computations become prohibitive for large q. Hastie, Friedman, and Tisbshirani, 2017
- A loss matrix may be required if we value one classification more than another
- We may have to discard missing predictor values. Potential fixes are appending a "missing" category or constructing a surrogate variable

CART, ID3, and C4.5

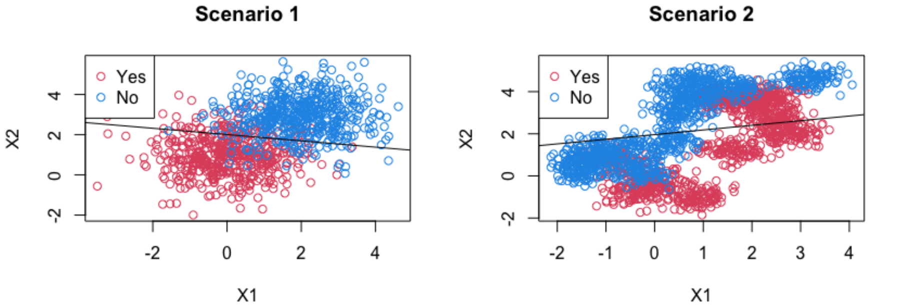
- Iterative Dichotomizer (3): used for **binary classification**
- Classification **and** Regression Trees: uses a mix of categorical and numerical data in order to form the tree
- ID3 Loss function: selects the split based on Information gain
- CART loss function selects splits to minimize Gini impurity

Partition Design for a Decision Tree





Scenarios 1 and 2



Why are Decision Trees considered unstable procedures?

Why are Decision Trees considered unstable procedures?

- Going back to the Bias/Variance tradeoff, Decision Trees are considered high variance but low bias models.
- For descriptive analysis, this is ideal as we may have full information and can categorize every point given to us.
- For predictive analysis, this leads to **overfitting**, and would result in terrible predictions if we were to obtain new points

Application of Bias Variance in Decision Trees

Cross Validation Revisited

- What are the advantages and disadvantages of k-fold cross validation relative to the validation set approach? LOOCV?

Cross Validation Revisited

- What are the advantages and disadvantages of k-fold cross validation relative to the validation set approach? LOOCV?
- Validation set approach is the simplest method, splitting the data into two parts (training and testing), and then evaluating the machine learning methods by comparing the error rates in the testing set
- k-fold cross validation utilizes multiple partitions in the data when splitting into training and testing set
- LOOCV is the most extreme case, using a single datapoint as a testing set, while all the other points are the training set
- Bias/Variance tradeoff for Cross Validation

Bootstrapping

- How do we use bootstrapping in Machine Learning?

Bootstrapping

- How do we use bootstrapping in Machine Learning?
- We can estimate the accuracy of a statistic of interest
- We can also use the bootstrap to estimate prediction error of our training set (pg. 226 of ISLR)
- We fit the model on a set of bootstrap samples and then keep track of how well it predicts the original training set
- What is the problem with this approach?

Cross Validation to obtain optimal level of Tree Complexity

Spam Dataset interpretation

- What is the learning problem?

Spam Dataset interpretation

- What is the learning problem?
- Considered a **Classification problem**
- Task: predict whether an observed email is spam
- Performance: % of emails correctly classified as spam
- Experience: Database of words that are commonly found in spam emails with given classifications

Analysis for SPAM dataset using Decision trees

References

Hastie, T., Friedman, J., & Tisbshirani, R. (2017). *The elements of statistical learning: Data mining, inference, and prediction*. Springer.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning: With applications in r* (Second). Springer.

Malato, G. (2020). An example of overfitting and how to avoid it. https://towardsdatascience.com/an-example-of-overfitting-and-how-to-avoid-it-f6739e67f394