# Decision Tree Learning from Scratch

Stefano Fochesatto

MATH 692 Mathematics for Machine Learning

2/10/2022

## Presentation Outline

- What is a Decision Tree?
- Training the Tree.
- Code Demo.
- Advantages and Pitfalls.
- Application in Ensemble Models.

# What is Decision Tree?

- Supervised Learning.
- A flowchart where internal nodes represent a test for the data.
- Leaf nodes apply classification label or regression.
- Derived from recursive partitioning.
- All nodes represent some partition of the data.
- We will discuss classification mainly,

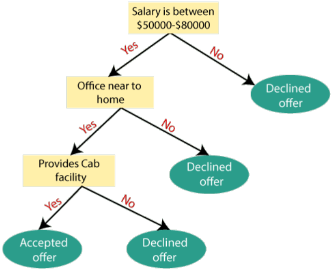



**Figure:** Decision Tree Example

# What is Decision Tree?

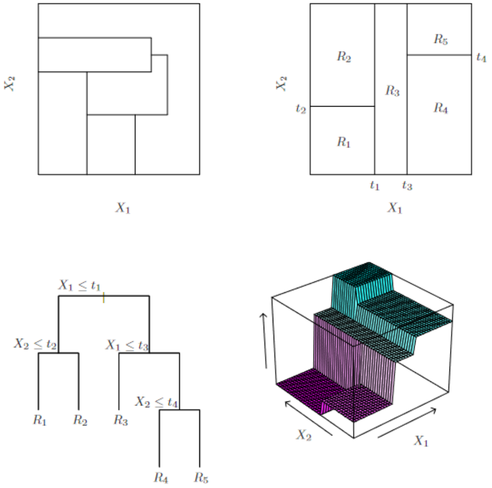



**Figure:** E.S.L. Friedman, Hastie, Tibshirani

## What is Decision Tree?

- Small Demo building a decision tree.

## Training the Tree.

- Methods don’t guarantee the optimal solution (Greedy).
- Top-Down Induction of Decision Trees (R.Quinlan)
- There are several algorithms for training.
  - CART (L.Breimann et al., 1984)
  - ID3 and C4.5 (R.Quinlan 1983, 93)
  - C5.0 (R.Quinlan)
  - and many more...

## Information Theory

- Consider the following splits,

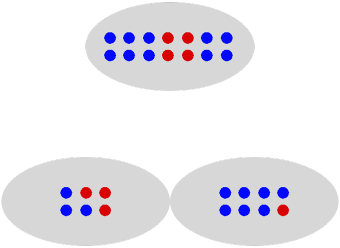


**Figure:** Example Split # 1

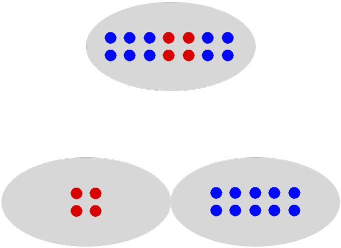


**Figure:** Example Split # 2

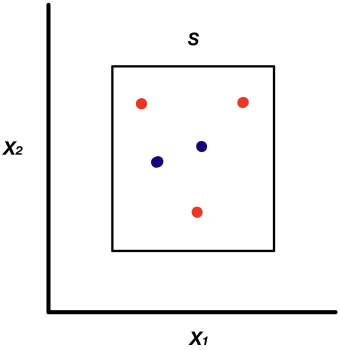
- To optimize our tree we need to be able to quantify the quality of a split (or generalized node)?

# Information Theory

- Let $X$ be a discrete random variable which represents a node $S$'s predicted class.
- Let $p(x)$ be the probability mass function.
- Consider the following example,

$$p(\textcolor{red}{\bullet}) = \frac{\#\text{ of } \textcolor{red}{\bullet}}{\text{Total } \#\text{ data in } S} = \frac{3}{5},$$

$$p(\textcolor{blue}{\bullet}) = \frac{\#\text{ of } \textcolor{blue}{\bullet}}{\text{Total } \#\text{ data in } S} = \frac{2}{5}.$$



## Information Theory

- From the previous slide we know that $X$ is a d.r.v with $n$(classes) possible outcomes and pmf $p(x)$. Now consider the following,

- In information theory, the unit of information ascribed to an outcome $x \in [n]$ is a log measure of 1/p(x),

$$I = \log_2\left(\frac{1}{p(x)}\right) = -\log_2(p(x)).$$

- Events that are rare have more information, events that are common have less information.

# Information Theory

- We want to know the expected information at a node over all outcomes (classification classes),

$$\mathbb{E}(I) = \sum_{i=1}^{n} p(i)\log_2\left(\frac{1}{p(i)}\right) = -\sum_{i=1}^{n} p(i)\log_2(p(i)).$$

- We want to find the split which maximizes $\Delta I$ or information gain,

$$\Delta I = \mathbb{E}(I)_{Parent} - \sum w(i)\mathbb{E}(I)_{Children}.$$

- Where $w(i)$ is the size of the child node relative to the parent node.
- Sometimes we only care about the difference between splits.

$$\Delta I = 1 - \sum w(i)\mathbb{E}(I)_{Children}.$$

## Training the Tree

- For the CART algorithm the Gini Impurity is used to evaluate the quality of a node,

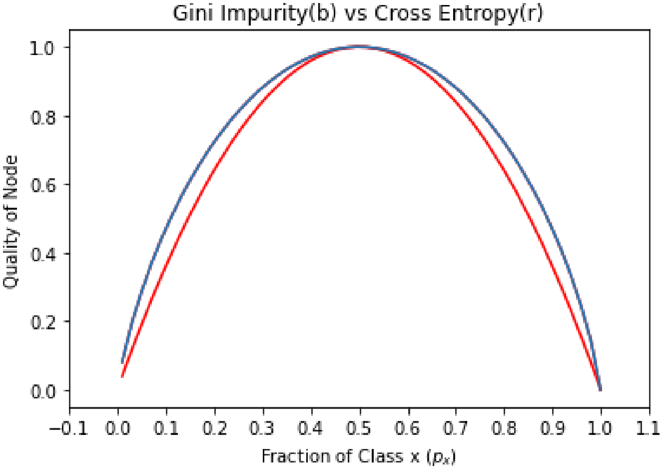
$$Gini = 1 - \sum_{i=1}^{n} p(i)^2.$$

- Again evaluating a split we take a weighted sum,

$$Gini_{split} = \sum w(i) Gini_{children}.$$

- Both methods are largely the same, Gini is preferred for predictive performance and computational complexity.

# Training the Tree



## Training the Tree

- A Very Naive Algorithm:
  - Search through each feature, threshold pair to find the optimal split for the current partition.
  - Partition the data.
  - Recurse.
  - Exit through hyperparameter or complete classification of training data.

- Code Demo

## Advantages and Pitfalls.

- Mimics human decision making.
- Can handle numerical and categorical data.
- Is an open-box model.
- Has naive runtime of $O(mn^2 log(n))$.
- It is robust to colinearity.
- Built-in metric for feature importance (with caveats)
- Very robust when boosted and bagged.

## Advantages and Pitfalls.

- Will be easily outperformed by other methods against linear decision boundaries,

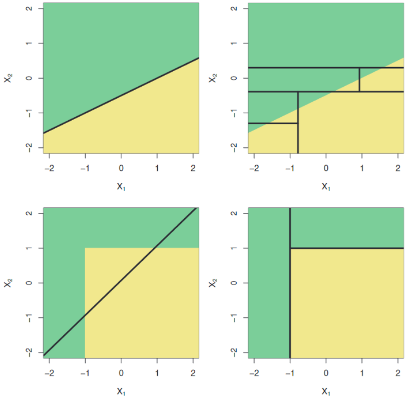



**Figure:** I.S.L. James, Witten, Hastie, Tibshirani

## Advantages and Pitfalls.

- Is prone to overfitting,

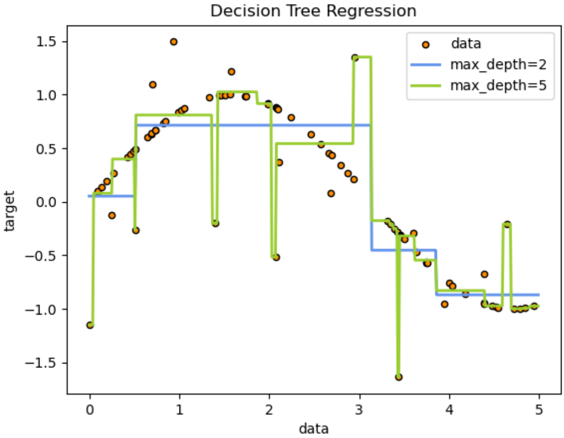



**Figure:** SKLearn Docs

## Dealing with Overfitting.

- Generally there are two ways to deal with overfitting.
  - Tuning Hyperparameters (pre-pruning)
    - max depth, min samples leaf, min samples split...
    - Grid Search Optimization =(
  - Cost Complexity Analysis (post-pruning)
    - Another optimization problem.
    - Grow Tree $T_0$ to Maximal Length,
    - Find the sub tree $T \subset T_0$ which minimizes the following,

$$C(T)_\alpha = \sum_{InternalNodes}^{|T|} Entropy + \alpha|T|$$

    - $\alpha$ is another parameter which is estimated using cross-validation.
- Note the Bias-Variance trade-off of pruning.

## Applications in Ensemble Models.

- Bagging (and RandomForest)
  - The underlying idea is model averaging (many to one).
  - Bootstrap the data (RandomForest means bootstrapping features).
  - Construct several full size decision trees.
  - When predicting average the results (majority vote).

- Individual models have very low bias.
- Averaging the models reduces the variance.

# Applications in Ensemble Models.

- Decision Tree Ensembles are good.

Journal of Machine Learning Research 15 (2014) 3133-3181 Submitted 11/13; Revised 4/14; Published 10/14

**Do we Need Hundreds of Classifiers to Solve Real World Classification Problems?**

**Manuel Fernández-Delgado** MANUEL.FERNANDEZ.DELGADO@USC.ES
**Eva Cernadas** EVA.CERNADAS@USC.ES
**Senén Barro** SENEN.BARRO@USC.ES
*CITIUS: Centro de Investigación en Tecnoloxías da Información da USC*
*University of Santiago de Compostela*
*Campus Vida, 15872, Santiago de Compostela, Spain*

**Dinani Amorim** DINANIAMORIM@GMAIL.COM
*Departamento de Tecnologia e Ciências Sociais- DTCS*
*Universidade do Estado da Bahia*
*Av. Edgard Chastinet S/N - São Geraldo - Juazeiro-BA, CEP: 48.305-680, Brasil*

**Figure:** 179 Classifiers, 17 families, 121 datasets

- Concluded that the best classifier over several datasets and metrics was an implementation of RandomForest.
- Six RandomForest classifiers and five SVM were among the top 20 classifiers.

## Applications in Ensemble Models.

- Decision Tree Ensembles are good.

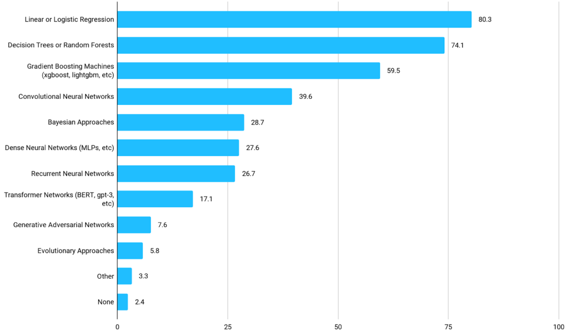



**Figure:** 2021 Kaggle survery: Over 25,000 Data Scientists and ML Engineers.

- Among ML practitioners decision trees are nearly as ubiquitous as regression.