Week – IV
Tree based ML Models

ML Bootcamp 2021

GDG Cloud EDMONTON

AlbertaAI

Careera Analytics Lab

Slido: #gdg_ml_bootcamp_w4

Table of contents

01 **Decision Trees**

What are they?

02 **Math behind DT**

How does it work internally?

03 **Ensemble learning**

The idea of voting to classify

04 **Random forest**

Random forest and its capabilities

05 **Advanced algorithms**

Gradient boosting techniques & XGBoost

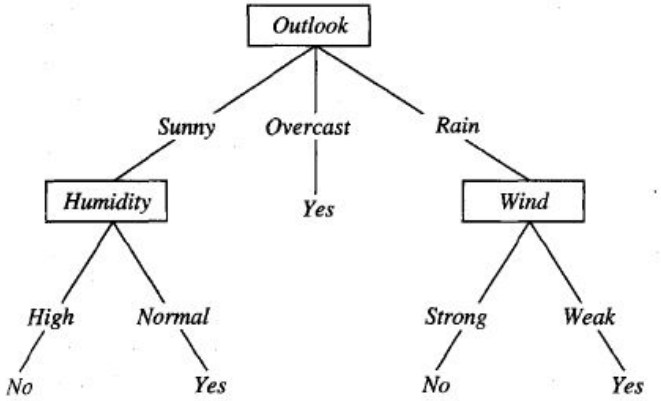
06 **Hands-on**

Lets start practicing!

Slido: #gdg_ml_bootcamp_w4

Decision Trees

- A machine learning algorithm where the learned function is in the form of a tree
- At each node of the tree, it must make a decision on the choice of the attribute (feature/column)
- The methodology behind choosing a certain attribute(feature/column) creates multiple algorithms as mentioned below:
  - **ID3** ← **Our focus for today!**
  - C4.5
  - CART(Classification and Regression Trees)



How are the attributes chosen?

- Principles from the subject of Information Theory are used, specifically “Entropy”
- Entropy characterizes the uncertainty(number of states) of a given system
- More precisely, it is the average number of bits needed to send a piece of information

$$Entropy(S) = \sum -p_i \log_2 p_i$$

Here S is the system, or the collection of examples and then p_i is the proportion of examples belonging to class i

Information Gain – ID3 Algorithm

1. To choose the best attribute(feature) out of all the given data, we use Information Gain, which uses entropy, to make the Decision

$$Information\ Gain = Entropy(S) - \sum_{i=Values(Attribute)}^{V} \frac{S_i}{S} Entropy(S_i)$$

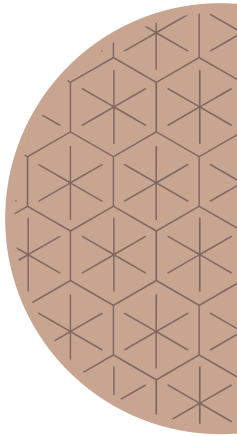
- The attribute that provides the highest information gain is chosen
- In other words, which attribute will lead to a reduction in entropy and increase in clarity

A short example

Outlook	Temperature	Humidity	Wind	Tennis
Sunny	Hot	High	Weak	No
Overcast	Hot	High	Weak	Yes
Sunny	Cold	Normal	Weak	Yes
Rain	Mild	High	Strong	No
Rain	Mild	Normal	Weak	Yes
Sunny	Hot	High	Strong	Yes
Overcast	Hot	Normal	Weak	Yes
Rain	Cool	Normal	Strong	No

Decision Tree Algorithms

- ID3 – Uses Information Gain as the testing criteria
- C4.5 – Uses Gain ratio as the testing criteria
- CART - Uses Gini index as the testing criteria
- And more…



Ensemble Learning

- Instead of a single algorithm making predictions on your dataset, why don't we consider a collection of algorithms and then make a collective decision?

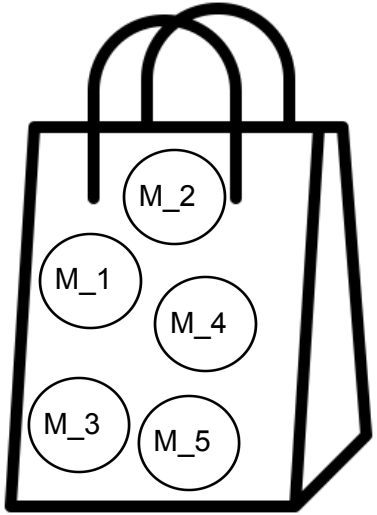
- Two most popular approaches: **Bagging** and **Boosting**

- The Netflix challenge that hyped up the idea of Ensemble Learning: https://blog.echen.me/2011/10/24/winning-the-netflix-prize-a-summary/

Bagging

Bagging_for_classification():

1. Create Bootstraps samples from your given dataset: $D_1, D_2, D_3, \ldots$
2. Train a set of Machine Learning algorithms(2 or more) on these different samples
3. Combine the predictions of multiple algorithms by taking a majority vote



$\{1,1,0,0,0\}$

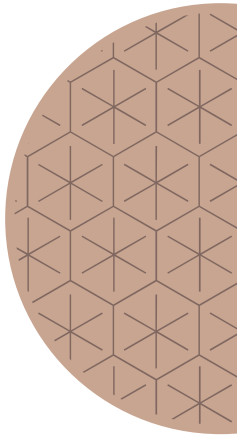
Random Forest

- It is a combination of multiple decision trees
- Each tree learns a from **Random** set of datasets, each one generated based on a randomly selected features and rows
- Final prediction is based on a combination of predictions from all the Decision Trees

Boosting

- Combine a bunch **Weak Learners** to make a **Strong Learner**

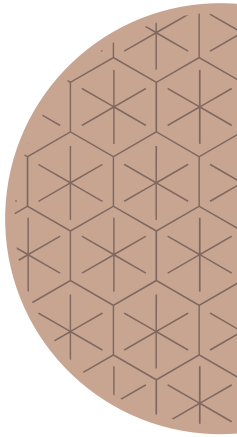
- Boosting_framework():
  - Train your classifier on the dataset and make predictions
  - Compute the weighted errors for datapoint based on correct/incorrect prediction
  - Iterate



Gradient Boosting Algorithm

Steps:

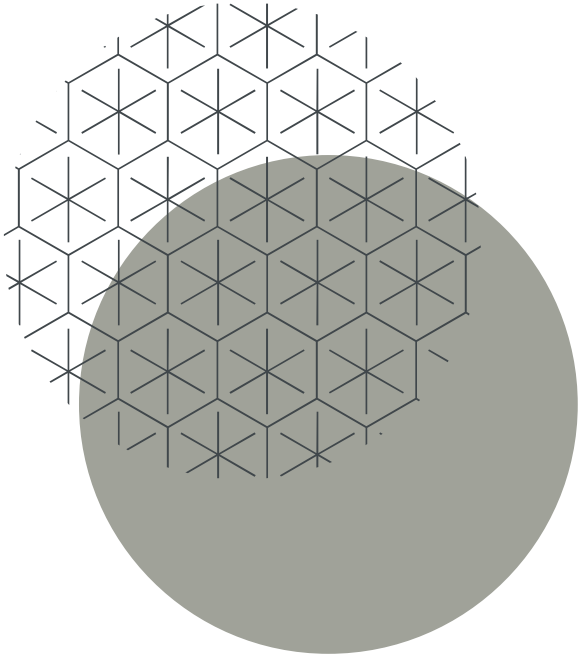
1. Initialize a base learner f maybe a random classifier
2. Start Iterations for k=1→K:
    a. Compute the residuals
    b. Train new learner M_k that is trained on predicting the residuals
    c. Update the learner f based on the new learner M_k

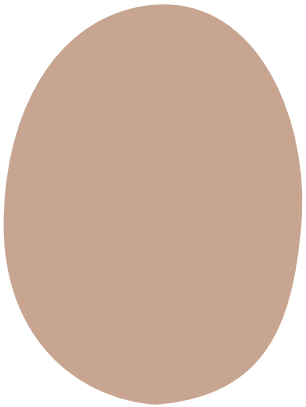


XGBoost Algorithm

- eXtreme Gradient Boosting algorithm
- This algorithm is same as the Gradient boosting algorithm we have seen before
- It has been engineered for better speed, accuracy and distributed computing
- One of the most popular algorithms used on Kaggle
- https://xgboost.readthedocs.io/en/latest/

Lets start Practicing!





Slido: #gdg_ml_bootcamp_w3

Thanks

Do you have any questions?


slidesgo