XGBoost

Part 2: Bagging and decision trees

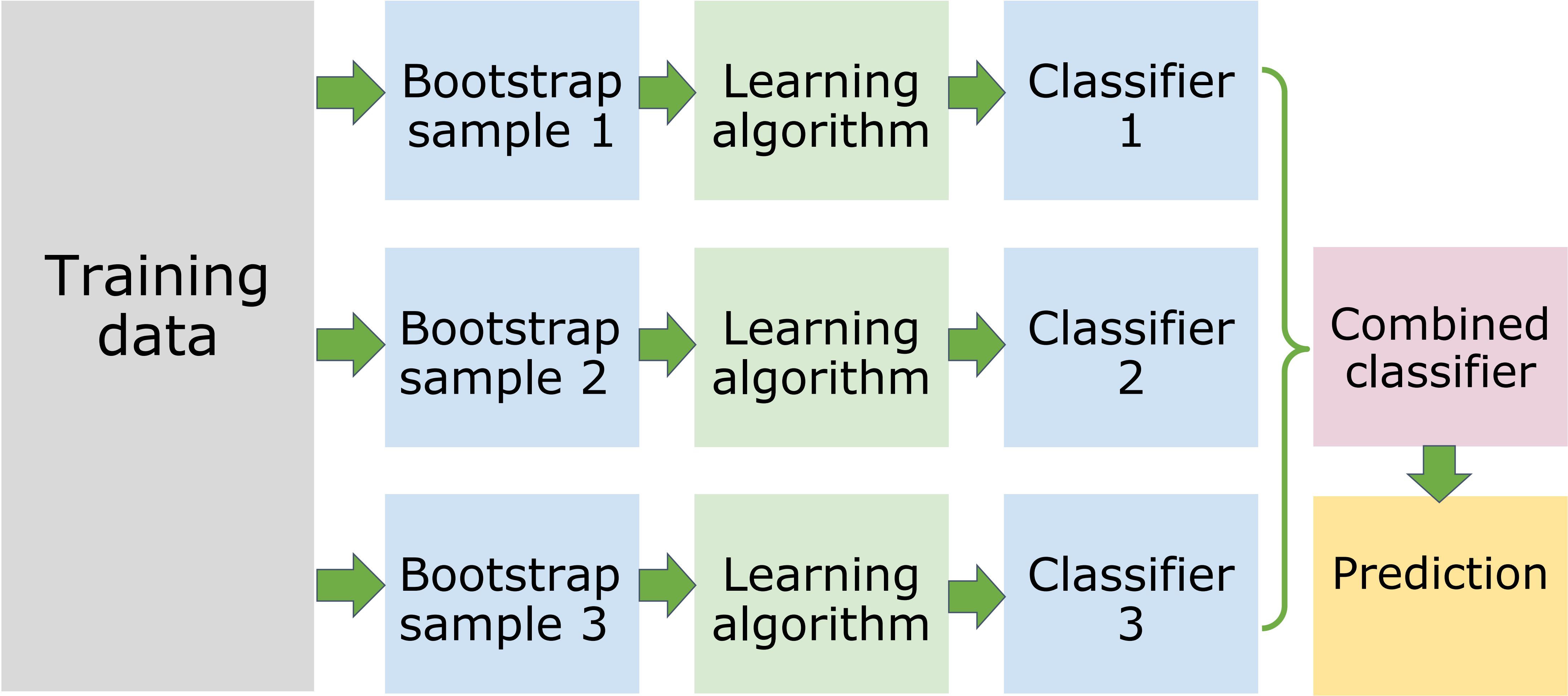
By: Noureddin Sadawi, PhD

University of London

Bagging (Bootstrap aggregation)

- Bootstrap sampling: a sampling method where a sample is randomly selected out of a set of points, using the **replacement** method.
  - In other words, the same point can be selected more than once in the same random sample.
  - If it is done without replacement, the subsequent selections will be dependent on the previous ones and this makes the sampling non-random.
- Aggregation: if we draw multiple samples and train a model on each sample, aggregation is combining these models for the final prediction.

Bagging (Bootstrap aggregation)



Training data

Bootstrap sample 1

Learning algorithm

Classifier 1

Bootstrap sample 2

Learning algorithm

Classifier 2

Bootstrap sample 3

Learning algorithm

Classifier 3

Combined classifier

Prediction

Bagging pros and cons

Pros:

- Allows several weak learners to join efforts and outperform a single more powerful learner.
- Reduces variance, and therefore cuts down overfitting.

Cons:

- Difficult to interpret overall model.
- Can be computationally expensive.

Decision tree

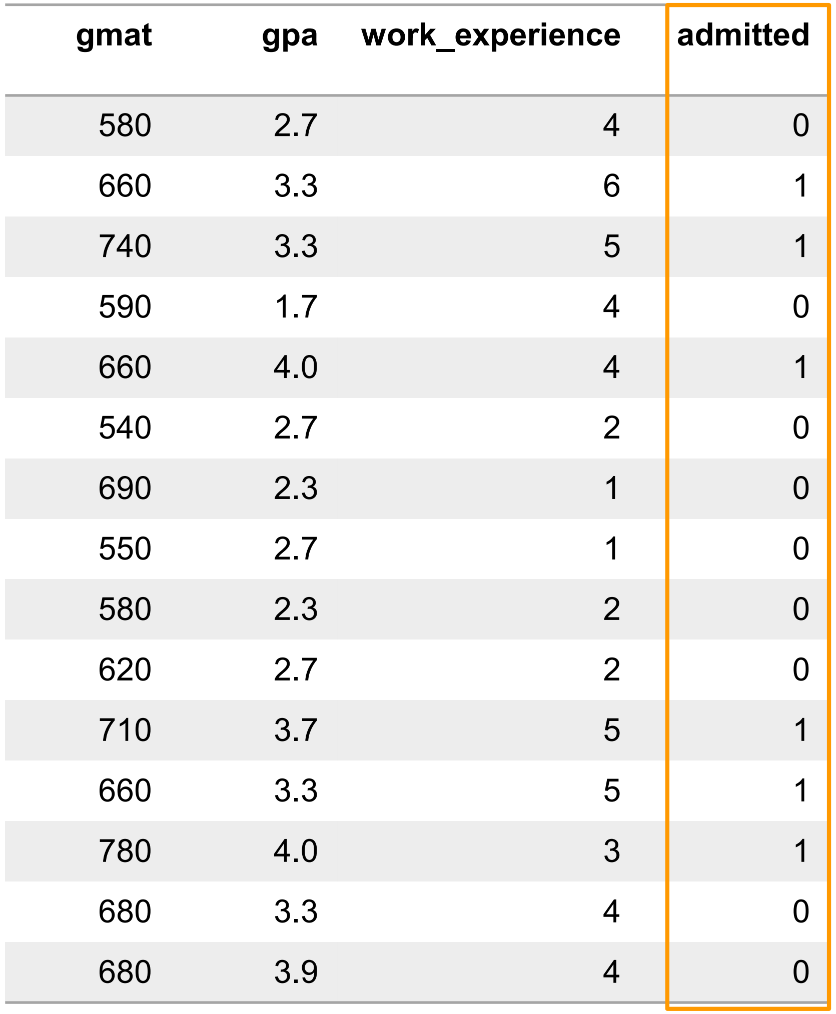
- The decision tree builds classification or regression models in the form of a tree structure.
- It breaks down a dataset into smaller and smaller subsets, while at the same time an associated decision tree is incrementally developed.
- The final result is a tree with decision nodes and leaf nodes.
- The algorithm at its core (called ID3) employs a top-down, greedy search through the space of possible branches with no backtracking.
- It uses entropy and information gain to select the best feature to split on at each level.

Decision tree

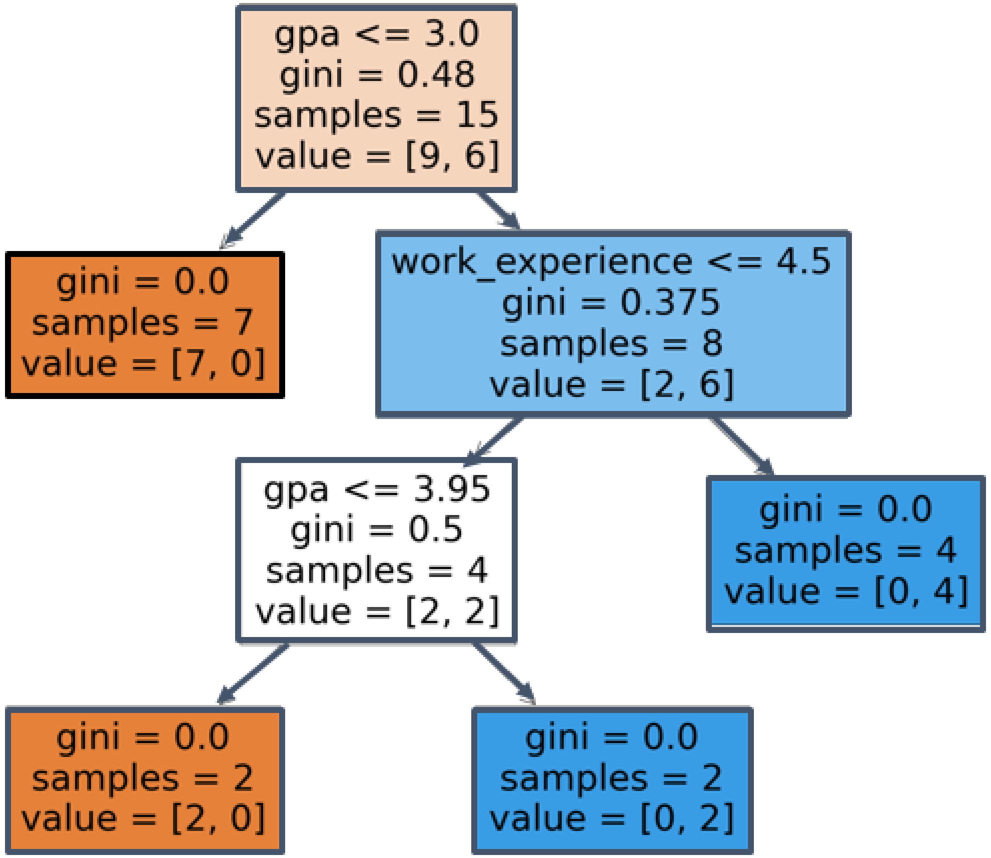
- A decision node (e.g. Outlook) has two or more branches (e.g. Sunny, Overcast and Rainy).
  - Leaf node (e.g. Yes) represents a classification or decision.
- The topmost decision node in a tree (which corresponds to the best splitting predictor) is called root node.
- Decision trees can handle both categorical and numerical data.

Outlook	Temp	Humidity	Windy	Play Golf?
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Rainy	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Sunny	Mild	High	False	No
Sunny	Cool	Normal	False	Yes
Rainy	Mild	Normal	False	Yes
Sunny	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Rainy	Mild	High	True	No

Decision tree example



gmat	gpa	work_experience	admitted
580	2.7	4	0
660	3.3	6	1
740	3.3	5	1
590	1.7	4	0
660	4.0	4	1
540	2.7	2	0
690	2.3	1	0
550	2.7	1	0
580	2.3	2	0
620	2.7	2	0
710	3.7	5	1
660	3.3	5	1
780	4.0	3	1
680	3.3	4	0
680	3.9	4	0



gpa <= 3.0
gini = 0.48
samples = 15
value = [9, 6]

gini = 0.0
samples = 7
value = [7, 0]

work_experience <= 4.5
gini = 0.375
samples = 8
value = [2, 6]

gpa <= 3.95
gini = 0.5
samples = 4
value = [2, 2]

gini = 0.0
samples = 4
value = [0, 4]

gini = 0.0
samples = 2
value = [2, 0]

gini = 0.0
samples = 2
value = [0, 2]

Random forest (RF)

- A classical bagging algorithm (an ensemble of decision trees).
- Builds several decision trees and combines them together to form a more accurate predictive model.
- In RF, only a random subset of the features is used to build each tree (often the size of this subset is sqrt(num of features)).
    - In other words, trees have different root nodes and so on.
- This usually results in a better model as the diversity is wider!