Introduction to Machine Learning. Lec.7 Random Forest

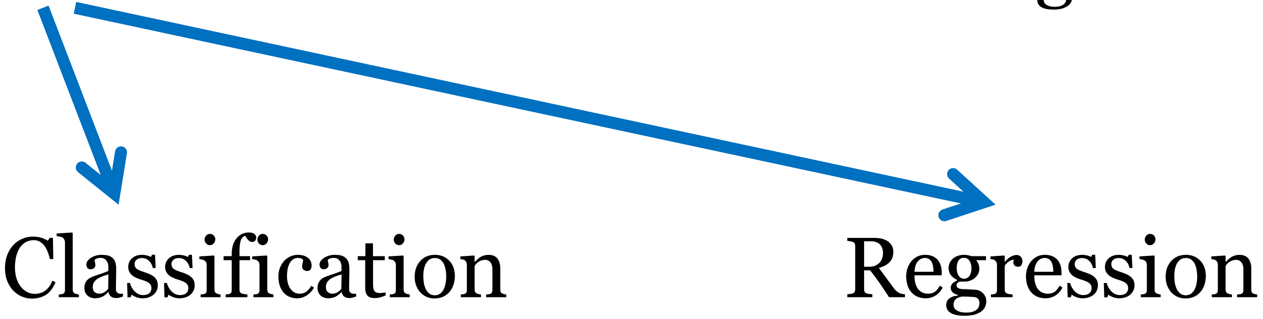
Aidos Sarsembayev, IITU, 2018

CART

- CART – is a classification and regression trees

CART

- CART – is a classification and regression trees

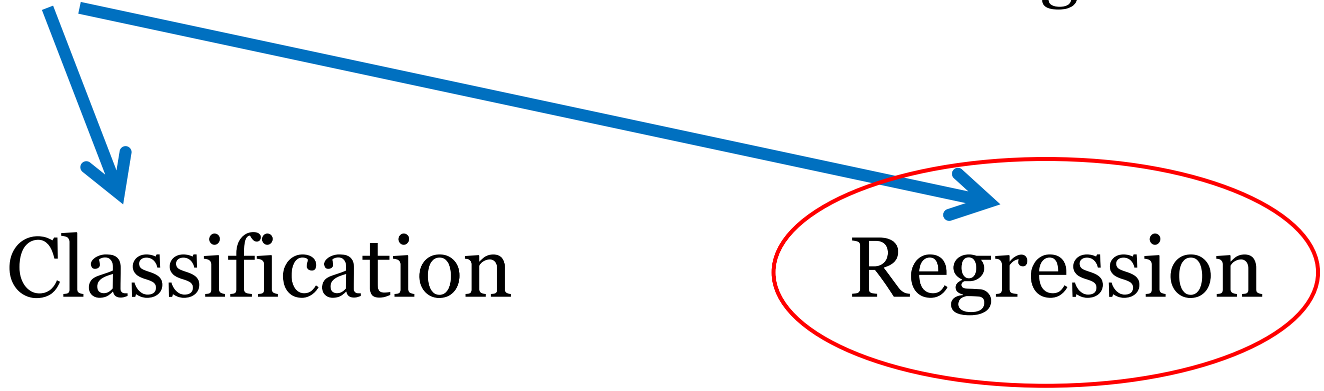


Classification

Regression

CART

- CART – is a classification and regression trees

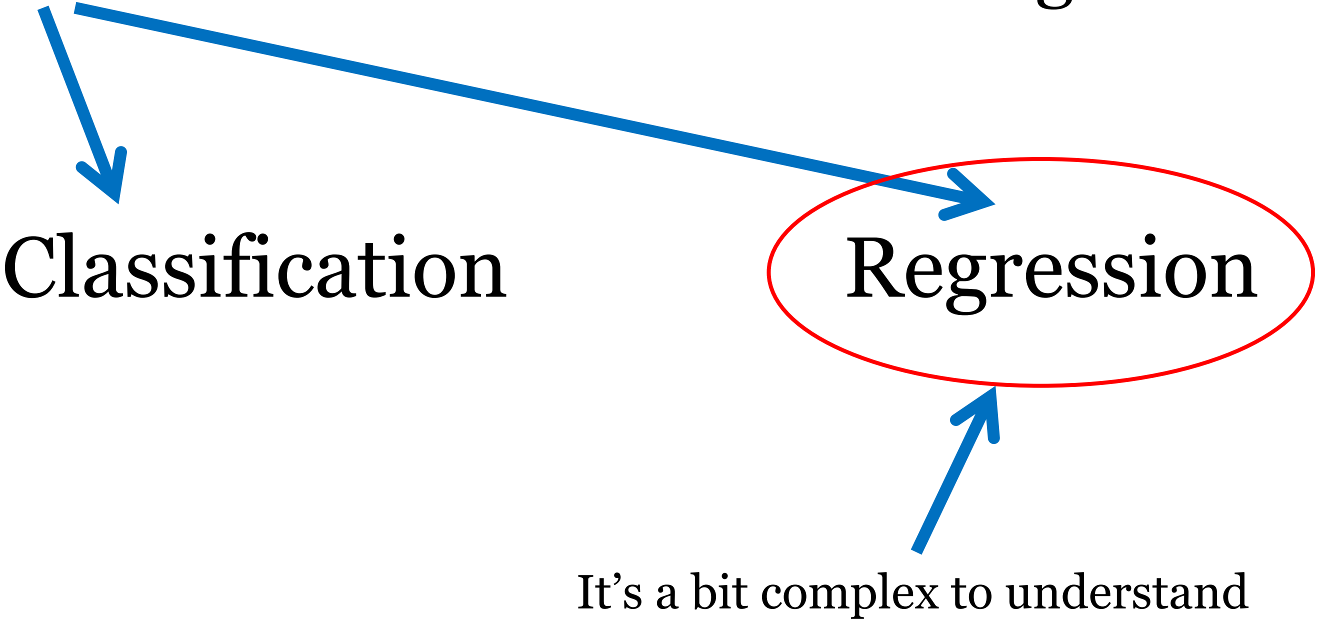


Classification

Regression

CART

- CART – is a classification and regression trees



Classification

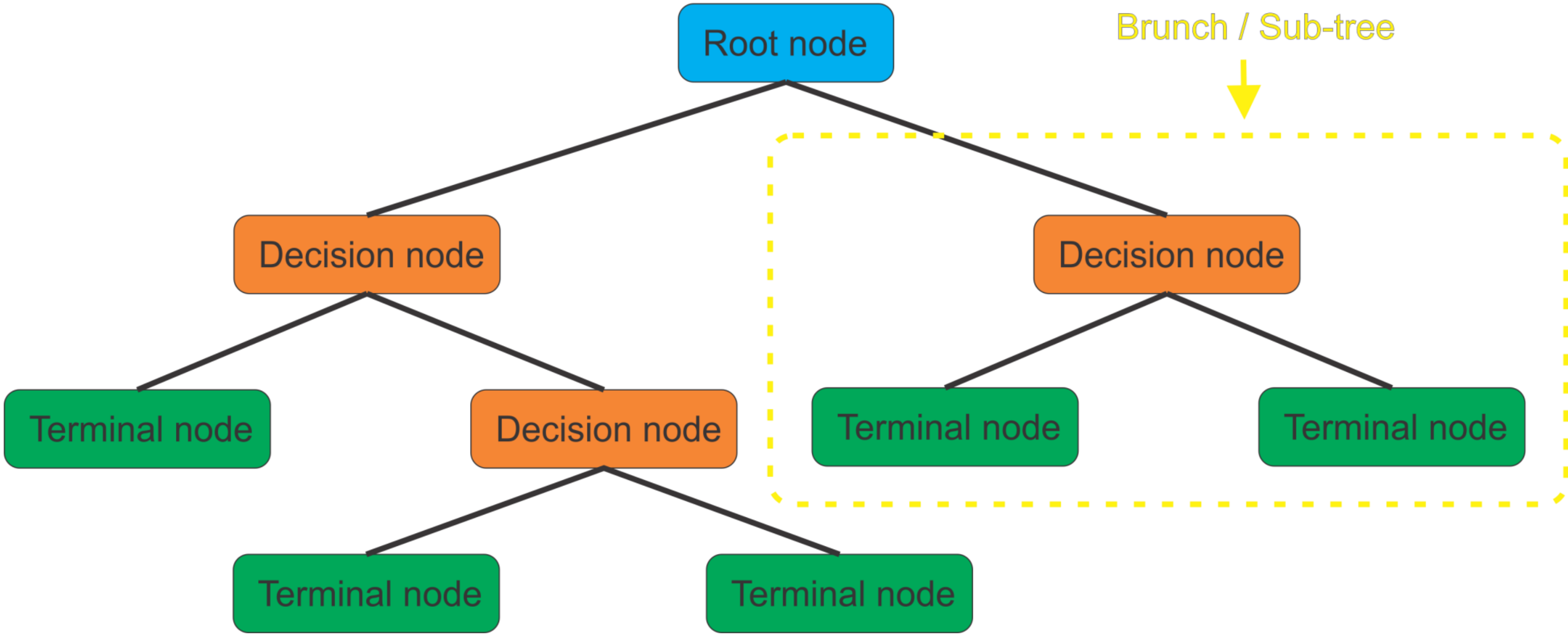
Regression

It’s a bit complex to understand

Regression & Classification

- As Decision Trees (DT), Random Forest (RF) also solves regression and classification problems

DT


Root node
Brunch / Sub-tree
Decision node
Decision node
Terminal node
Decision node
Terminal node
Terminal node
Terminal node
Terminal node

Disadvantages of DT

- Doesn’t generalize data well (a.k.a. overfitting)
- Can be unstable because of small variations of data (a.k.a. variance). It needs to be lowered by methods called bagging and boosting

- A better way is to train multiple trees

DT vs. RF

- DT – is a single decision tree
- RF – lots of trees…a forest of trees

DT vs. RF

- DT – is a single decision tree
- RF – lots of trees…a forest of trees

- The more trees we have in the forest, the more robust is our prediction and thus, the higher our accuracy

The ways of building RF

- Information Gain
- Gini index approach
- Other DT algorithms

RF classification

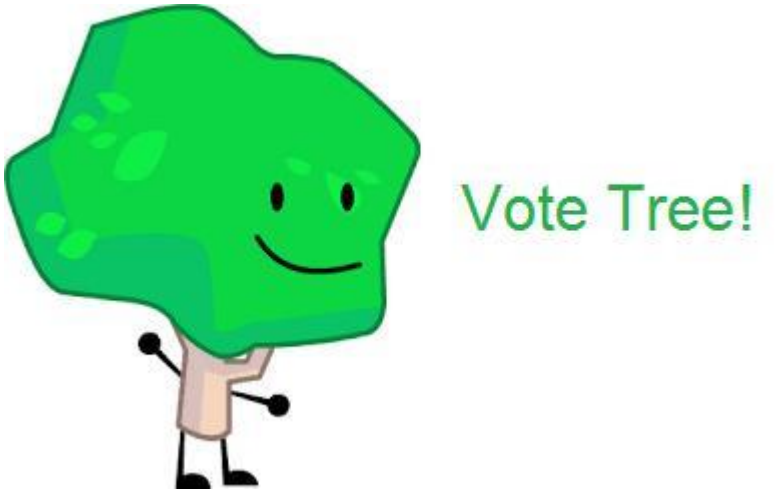
- In case of RF classification, each tree classifies a new object OR… votes for it
- The forest chooses the object having the most votes in order to make final decision (classification)

RF regression

- In case of RF classification, each tree classifies a new object OR… votes for it
- The forest chooses the object having the most votes in order to make final decision (classification)
- **In case of regression – takes the average of the outputs by different trees**

Voting trees

Voting trees



Advantages of RF

- Supports both – classification and regression
- Handles the missing data and maintains the accuracy when data is missing
- Doesn’t overfit the model
- Handles large datasets with high dimensionality

Disadvantages of RF

- Good for classification, less good for regression.

Disadvantages of RF

- Good for classification, less good for regression. In case of regression, it cannot predict beyond the range of the train data.

Disadvantages of RF

- Good for classification, less good for regression.
  In case of regression, it cannot predict beyond the range of the train data.
  It also may overfit model when data is very noisy.

Applications of RF

- Banking sector
- Stock market
- Recommendation systems for sellers
- Disease classification in medicine
- Computer vision (Microsoft Kinect)
- Speech recognition

Pseudocode of RF algorithm

- Assume number of cases in the training set is N. Then, sample of these N cases is taken at random, but with replacement
  (the sample will be the training set for growing the tree)

Pseudocode of RF algorithm

- Assume number of cases in the training set is N. Then, sample of these N cases is taken at random, but with replacement
  (the sample will be the training set for growing the tree)
- If there are M input variables or features, a number m<M is specified such that at each node, m variables are selected at random out of the M. The best split on these m is used to split the node. The value of m is held constant while we grow the forest

Pseudocode of RF algorithm

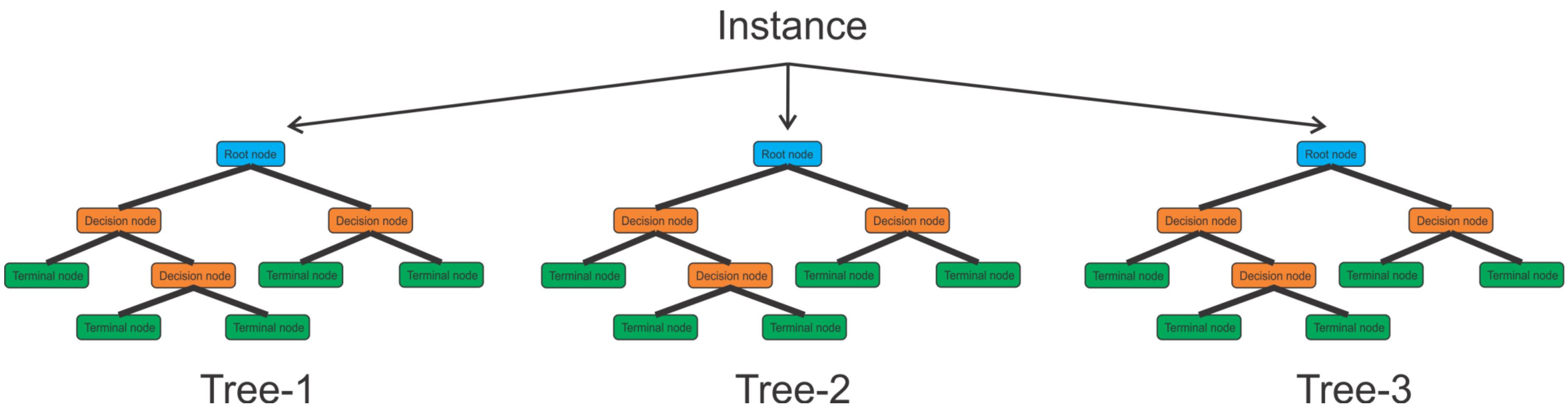
- Assume number of cases in the training set is N. Then, sample of these N cases is taken at random, but with replacement
  (the sample will be the training set for growing the tree)
- If there are M input variables or features, a number m<M is specified such that at each node, m variables are selected at random out of the M. The best split on these m is used to split the node. The value of m is held constant while we grow the forest
- Each tree is grown to the largest extent possible and there is no pruning

Pseudocode of RF algorithm

- Assume number of cases in the training set is N. Then, sample of these N cases is taken at random, but with replacement
  (the sample will be the training set for growing the tree)
- If there are M input variables or features, a number m<M is specified such that at each node, m variables are selected at random out of the M. The best split on these m is used to split the node. The value of m is held constant while we grow the forest
- Each tree is grown to the largest extent possible and there is no pruning
- Predict new data by aggregating the predictions of the n tree trees (i.e. majority votes for classification, average for regression)

Example of RF

- Let’s say we want to build a forest classifying an image. We want to know if there is or there is no cat on the image.



Instance

Root node

Decision node

Decision node

Terminal node

Decision node

Terminal node

Terminal node

Terminal node

Terminal node

Tree-1

Root node

Decision node

Decision node

Terminal node

Decision node

Terminal node

Terminal node

Terminal node

Terminal node

Tree-2

Root node

Decision node

Decision node

Terminal node

Decision node

Terminal node

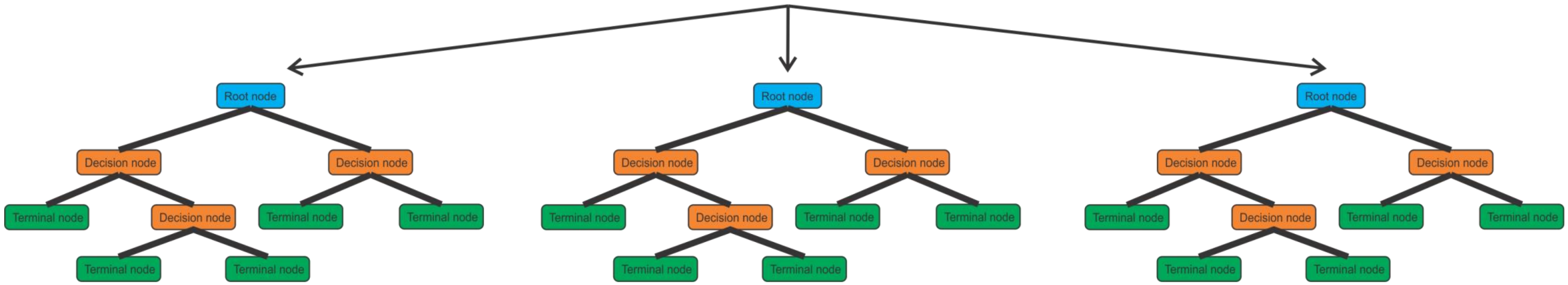
Terminal node

Terminal node

Terminal node

Tree-3

Instance



Tree-1

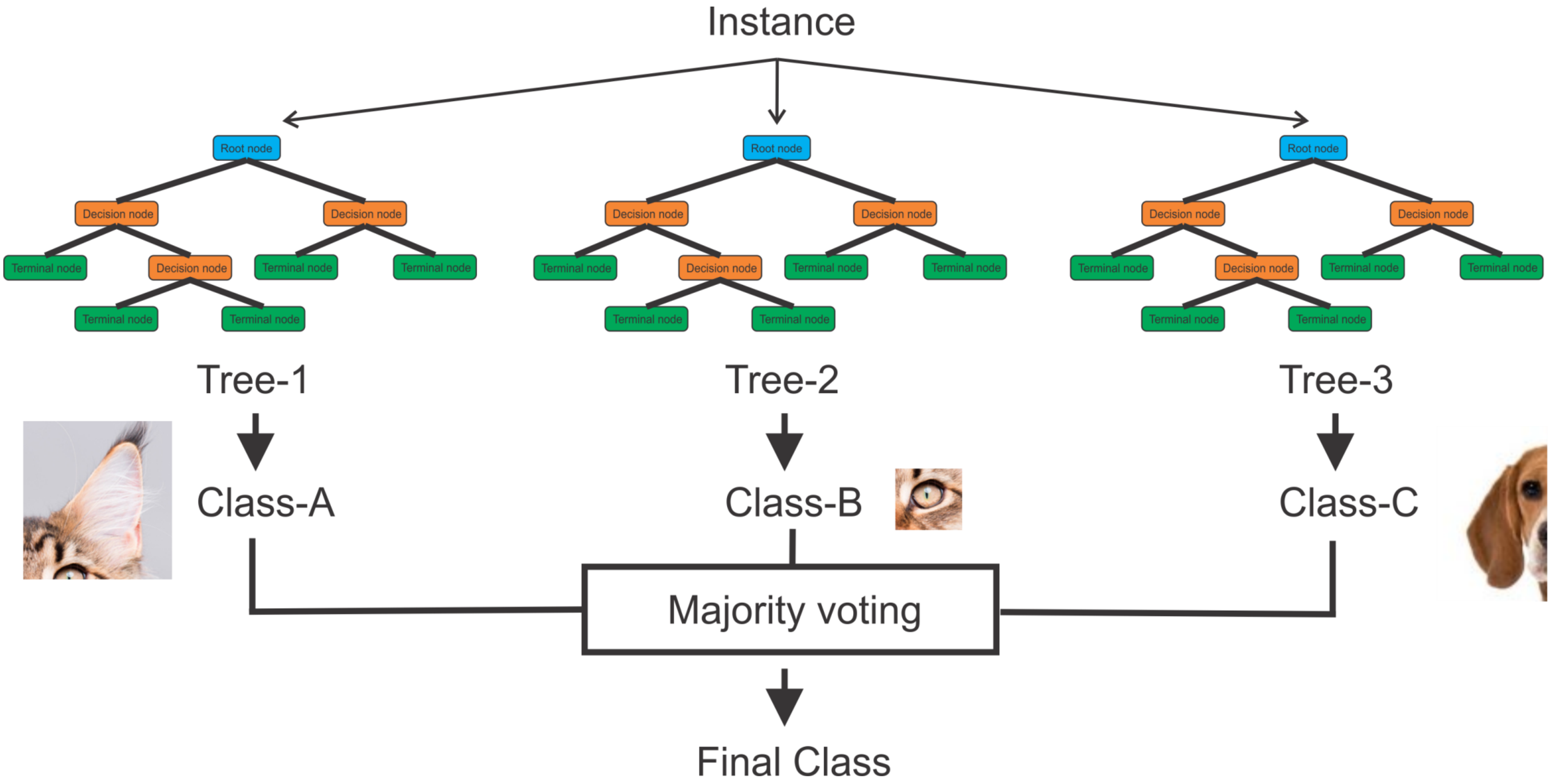
Class-A

Tree-2

Class-B

Tree-3

Class-C



Instance

Root node

Decision node

Decision node

Terminal node

Decision node

Terminal node

Terminal node

Terminal node

Terminal node

Tree-1

Class-A

Root node

Decision node

Decision node

Terminal node

Decision node

Terminal node

Terminal node

Terminal node

Terminal node

Tree-2

Class-B

Root node

Decision node

Decision node

Terminal node

Decision node

Terminal node

Terminal node

Terminal node

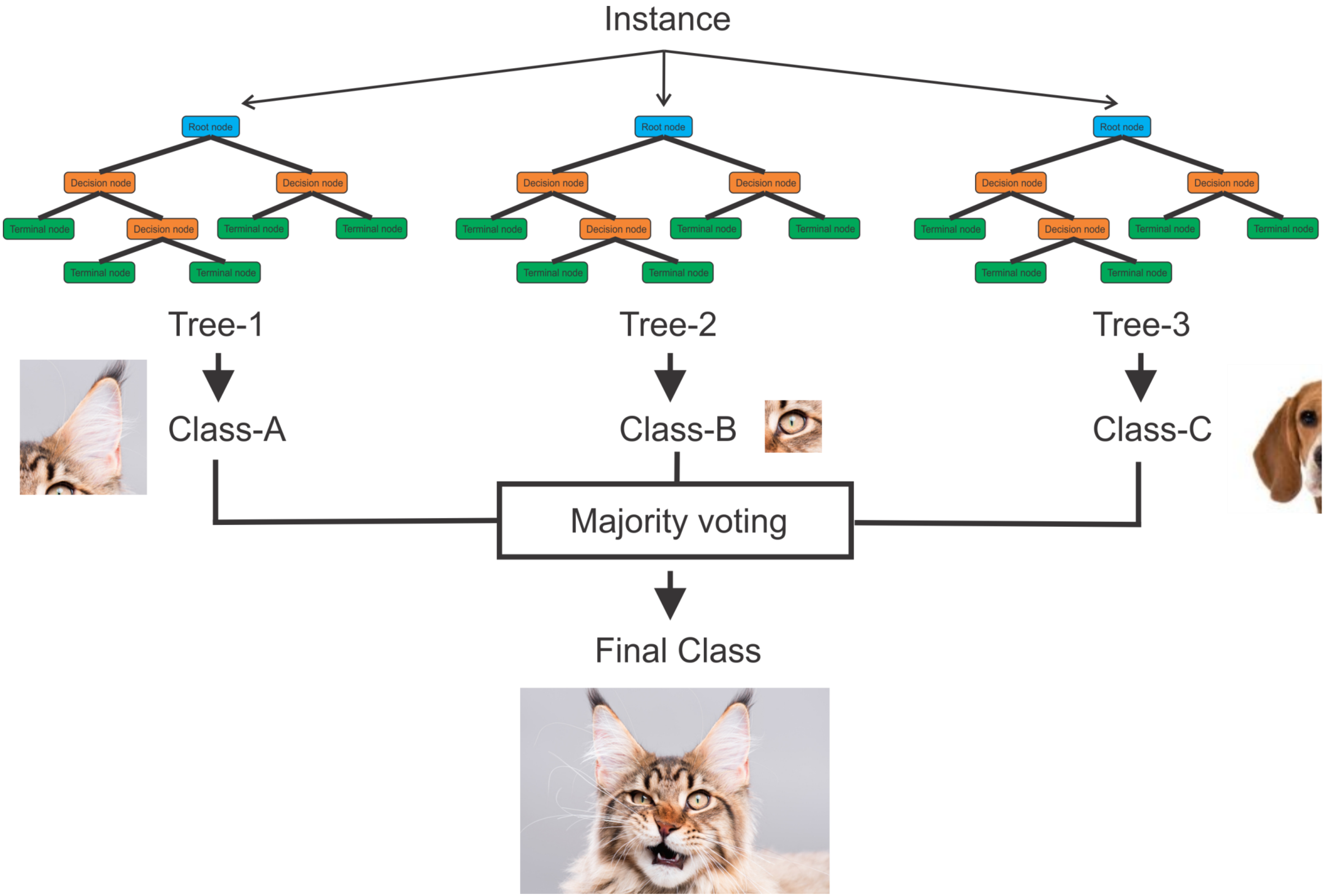
Terminal node

Tree-3

Class-C

Majority voting

Final Class



Instance

Root node

Decision node

Decision node

Terminal node

Decision node

Terminal node

Terminal node

Terminal node

Terminal node

Tree-1

Root node

Decision node

Decision node

Terminal node

Decision node

Terminal node

Terminal node

Terminal node

Terminal node

Tree-2

Root node

Decision node

Decision node

Terminal node

Decision node

Terminal node

Terminal node

Terminal node

Terminal node

Tree-3

Class-A

Class-B

Class-C

Majority voting

Final Class

RF is Ensemble ML method

- The RF algorithm takes into account the majority voting
- The majority voting consists of multiple decision trees
- This is why RF belongs to Ensemble methods

Ensemble method

- Ensemble methods are divide-and-conquer type of methods
- The main principle is to form a strong learner out of a group of weak learners

Some few terms

- Bagging
  is an ML ensemble meta-algorithm designed to improve the stability, reduce variance and helps to avoid overfitting
- Boosting
  is an ML ensemble meta-algorithm for primarily reducing bias, and also variance. It also converts weak learners into strong ones.

- More about them later.