AI & Robotics

Random Forests

PXL AI & ROBOTICS LAB

Goals

The **junior-colleague**

- can explain bagging in the context of random forests in their own words
- can explain bootstrap sampling
- can explain a random forest in their own words
- can explain the importance of random sampling in the context of random forests
- can describe the habit of overfitting in context of decision trees
- can describe why the different decision trees in a random forest need to be as uncorrelated as possible
- is able to sum up and explain 5 advantages and 2 disadvantages of random forests

Ensemble methods

- Combining classifiers
- Methods:
  - **Bagging**
  - Boosting
  - Voting
  - Stacking

Bagging

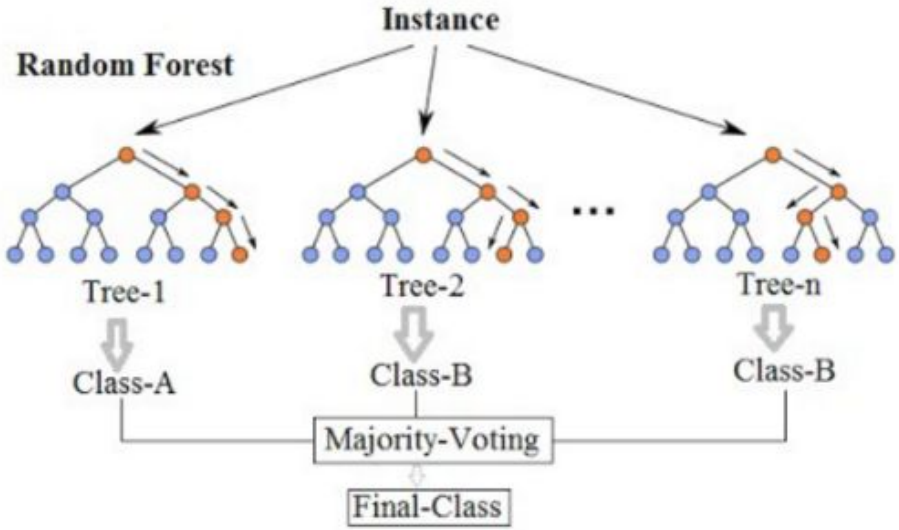
- Bootstrap aggregating
- Apply "bootstrapping"
  - From training set T of size n, draw sample T_i of size n (with replacement)
  - Do this m times -> m different training sets from T
- Learn model from each different data set T_i

=> Different learners have different inductive bias

=> Each model makes their own mistakes

- Average over the results

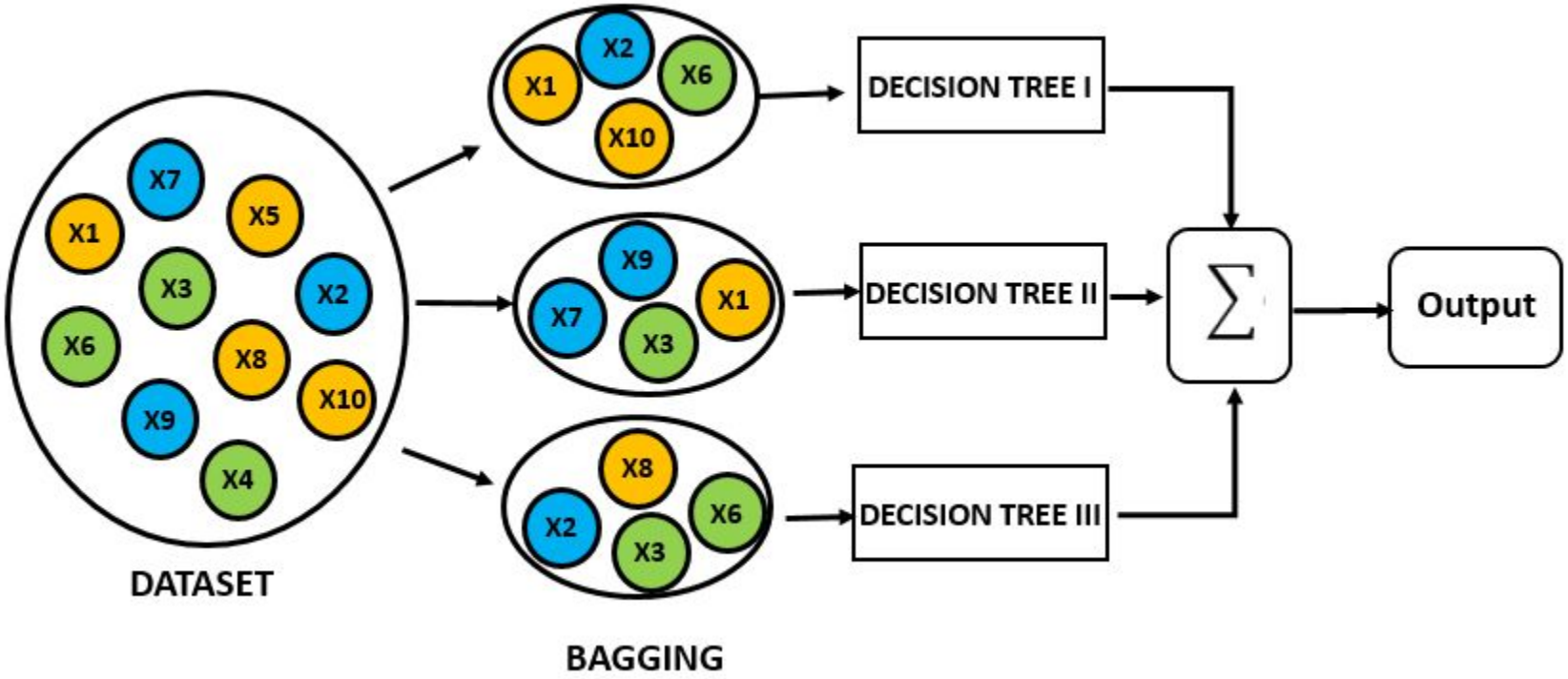
Random forests



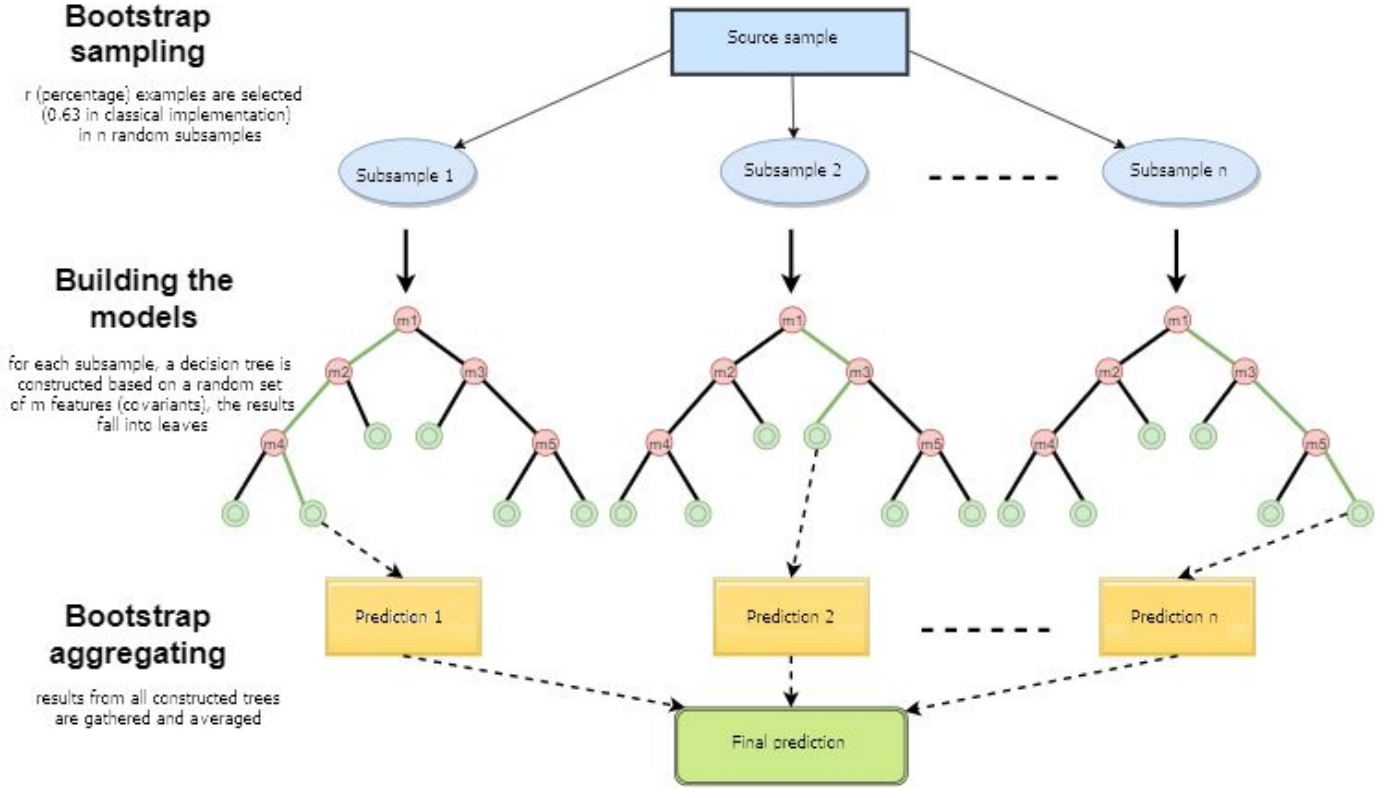
- General purpose Machine Learning model for structured data
- Good at lots of different problems
- Method: Decision Tree Bootstrap Aggregation

=> Correct for decision trees' habit of overfitting to their training set

Random forests



Random forests



Random forests

- You want the decision trees in your random forest to be as uncorrelated as possible
- Remember Bootstrapping!

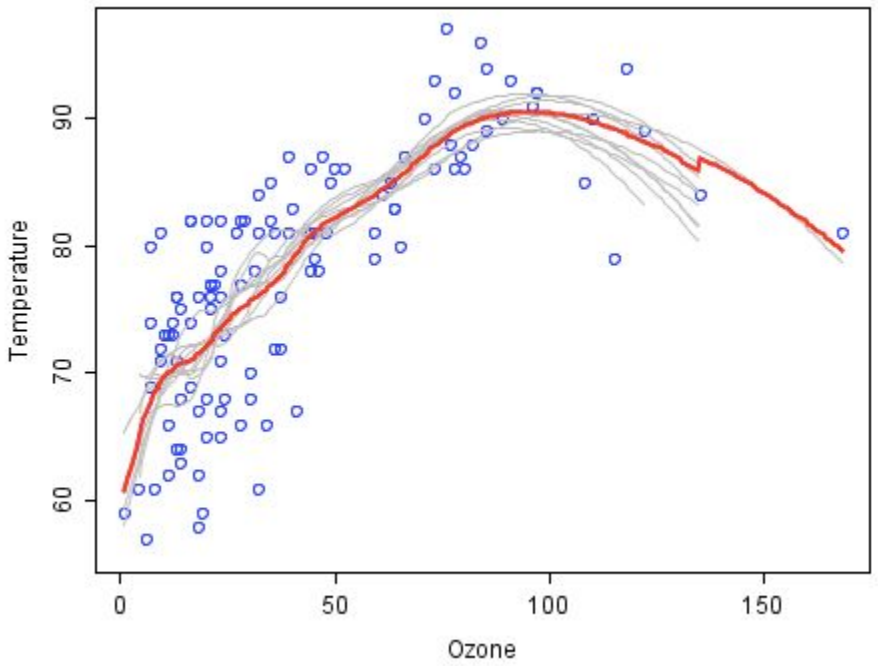
  => Random sampling with replacement

  Example:
  - Dataset: 100 000 samples
  - Random forest:
    - 1000 trees on 10 random samples

      => trees are uncorrelated
    - 1000 trees on the entire dataset - 1 sample

      => trees are correlated

Random forests



Why random forests?

Advantages	Disadvantages
Generally applicable to structured data	Time-consuming to construct for large datasets
Flexible and high accuracy	Not as easy to interpret as a single decision tree
Easy to visualize separate decision trees	
Able to handle both numerical and categorical data	
Able to handle multi class classification