Advanced School in Artificial Intelligence

Supervised Learning
Tree Models

Elena Bellodi
elena.bellodi@unife.it

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

Regione Emilia-Romagna

FERRARIAE UNIVERSITAS · 13 91 · EX LABORE FRUCTUS

Università degli Studi di Ferrara

Outline

Tree Models

- Decision Trees
  - classification trees
  - regression trees
- Bagging
- Random Forests

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

Bagging

If we split the data in random different ways, decision trees give different results → **high variance.**

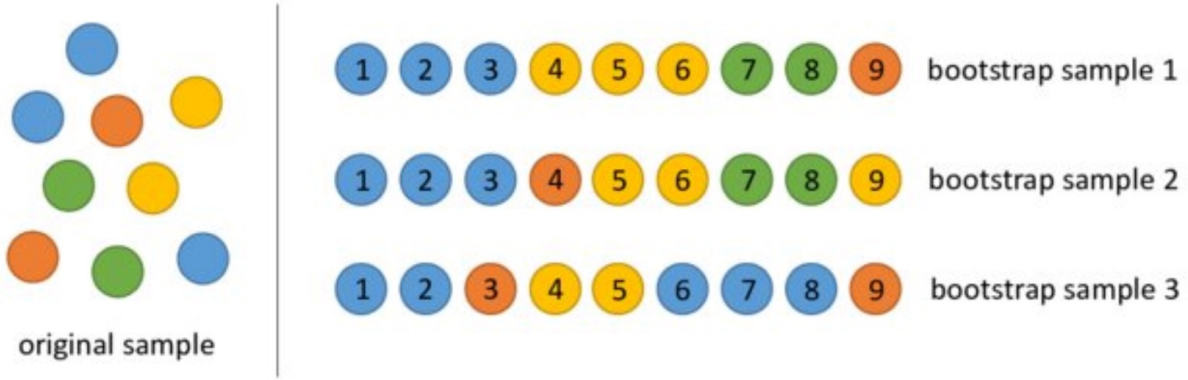
Bagging: **B**ootstrap **agg**regat**ing** is a method that results in low variance of an estimated prediction function.

If we had multiple realizations of the data (or multiple samples) we could calculate the predictions multiple times to produce less uncertain results

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

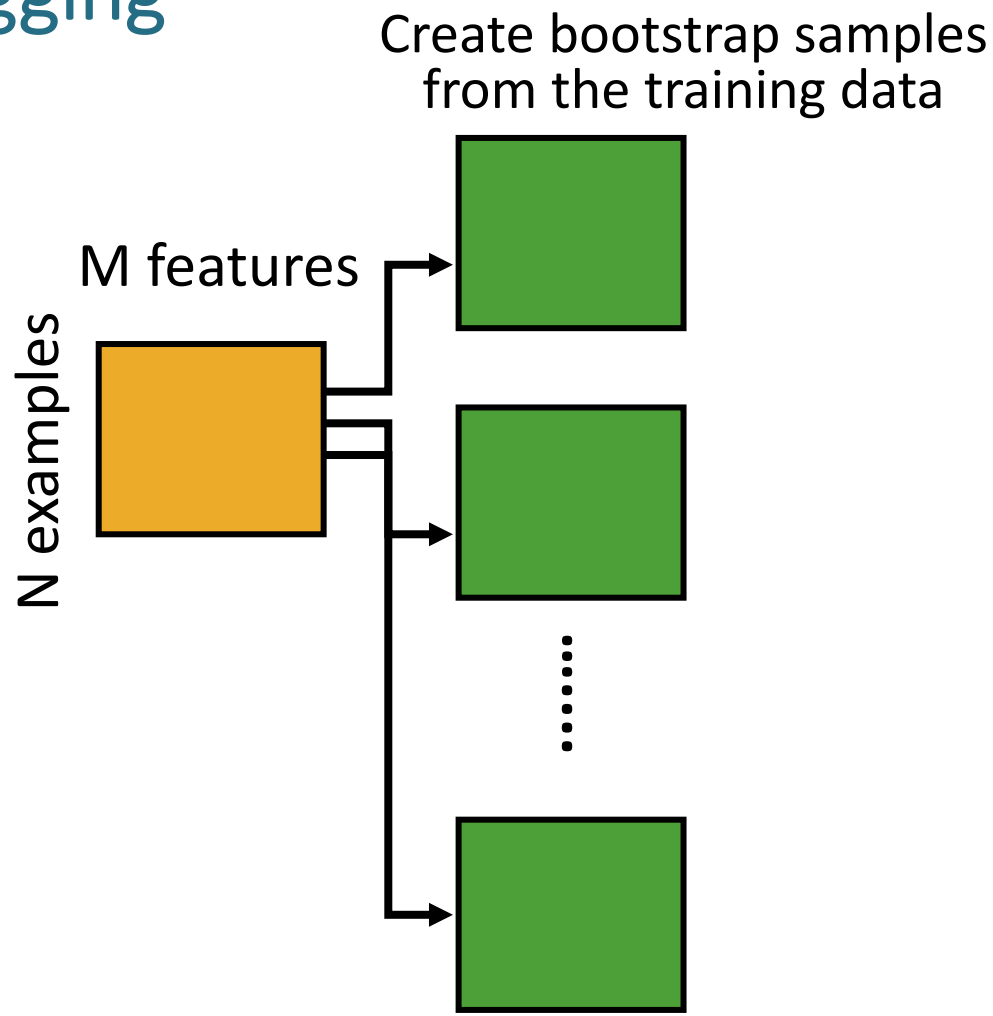
Bootstrap

- The basic idea:
- randomly draw datasets *with replacement* from the
- training data, each sample *the same size as the original training set*



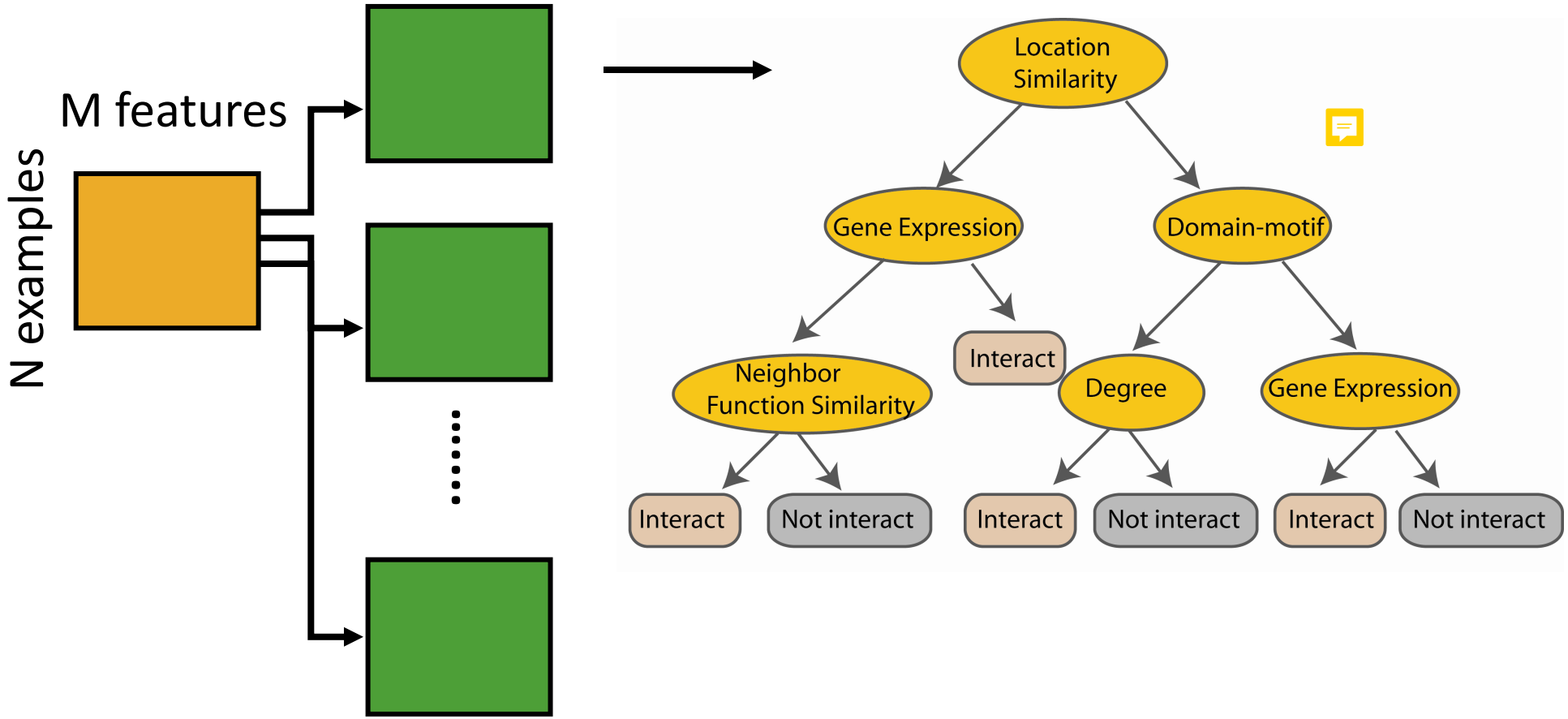
- Each bootstrap sample is likely to leave out about a 1/3 of the data points
- We call them out-of-bag samples (OOB)

Bagging



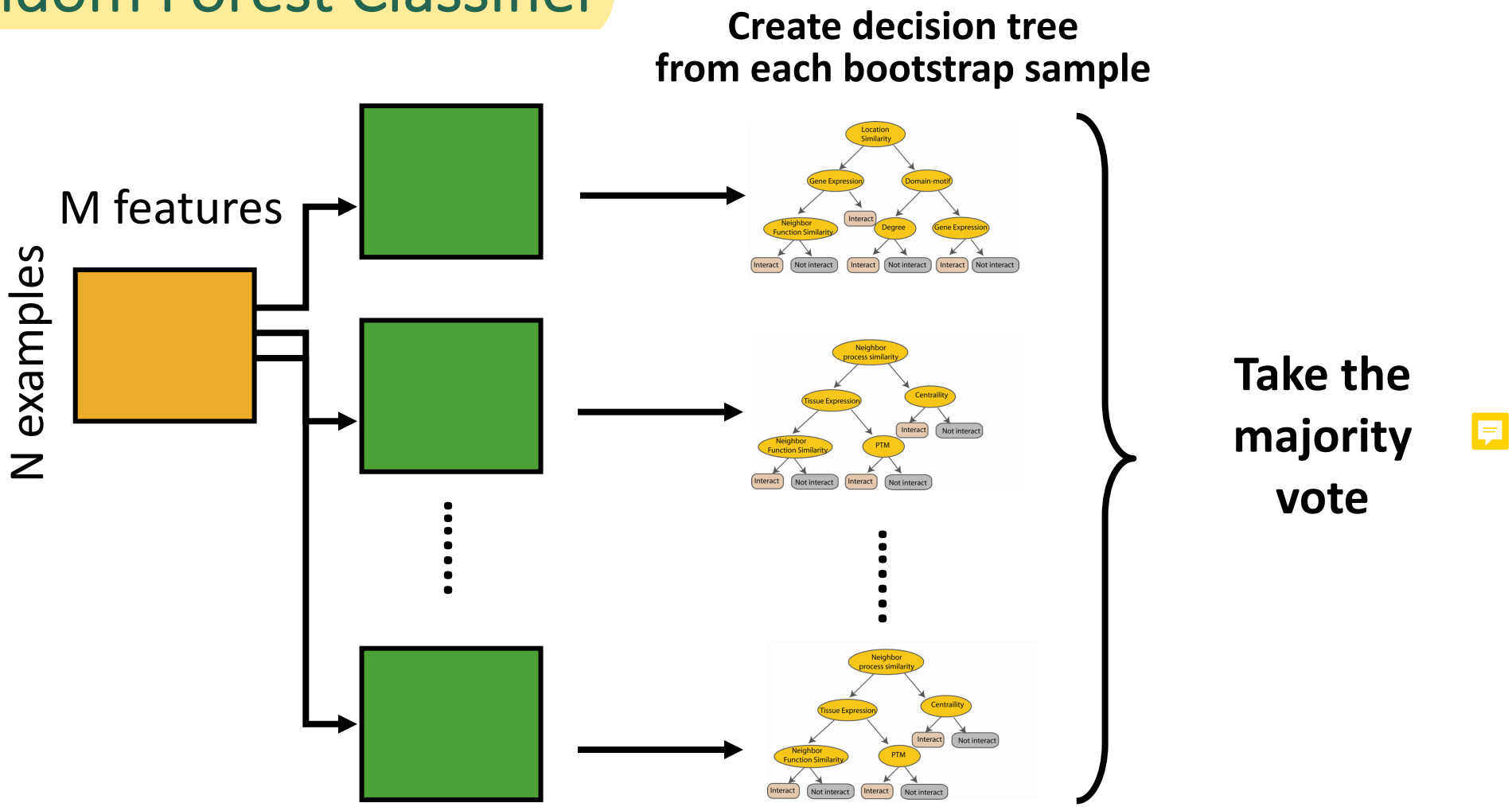
Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

Random Forest Classifier



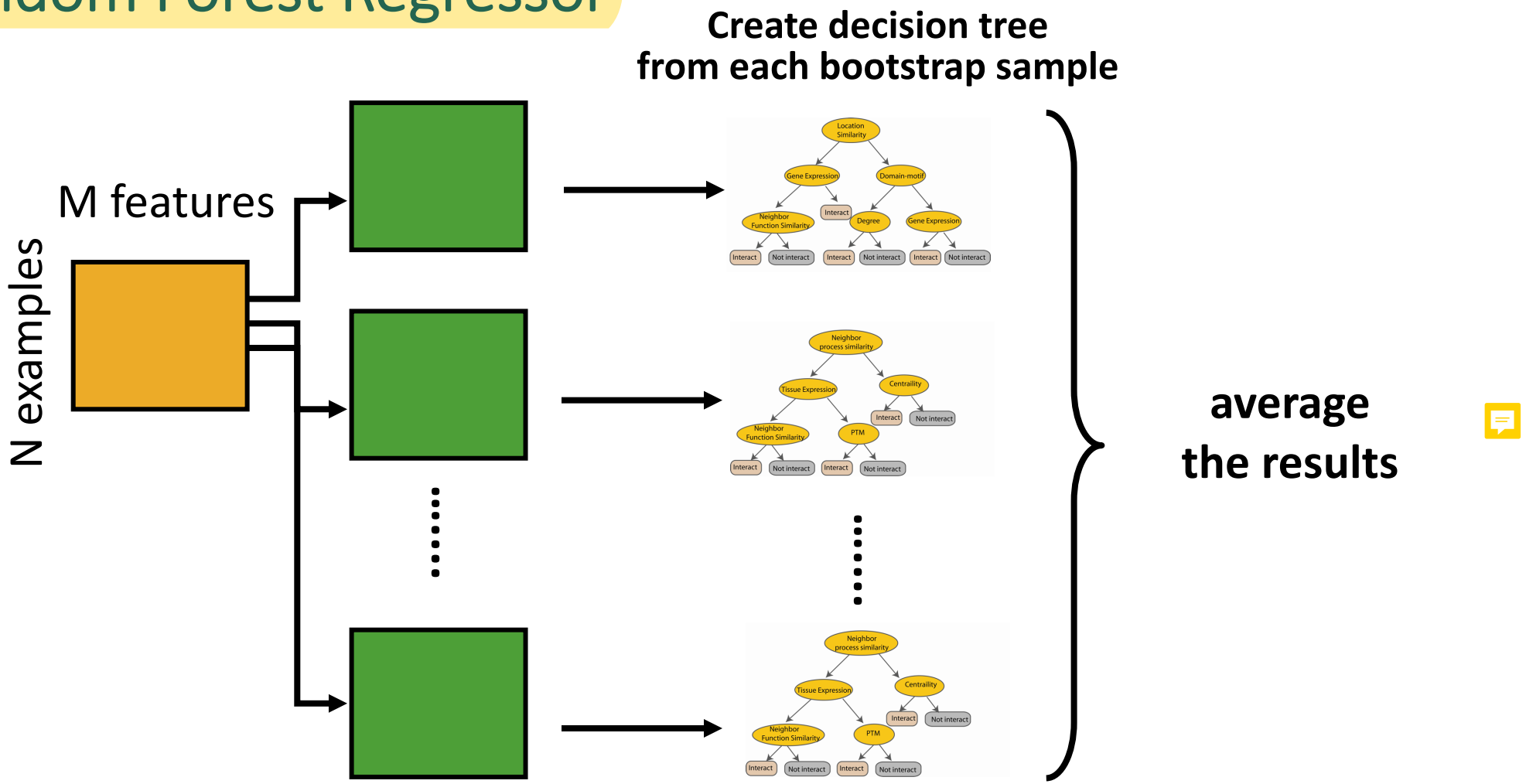
Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

Random Forest Classifier



Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

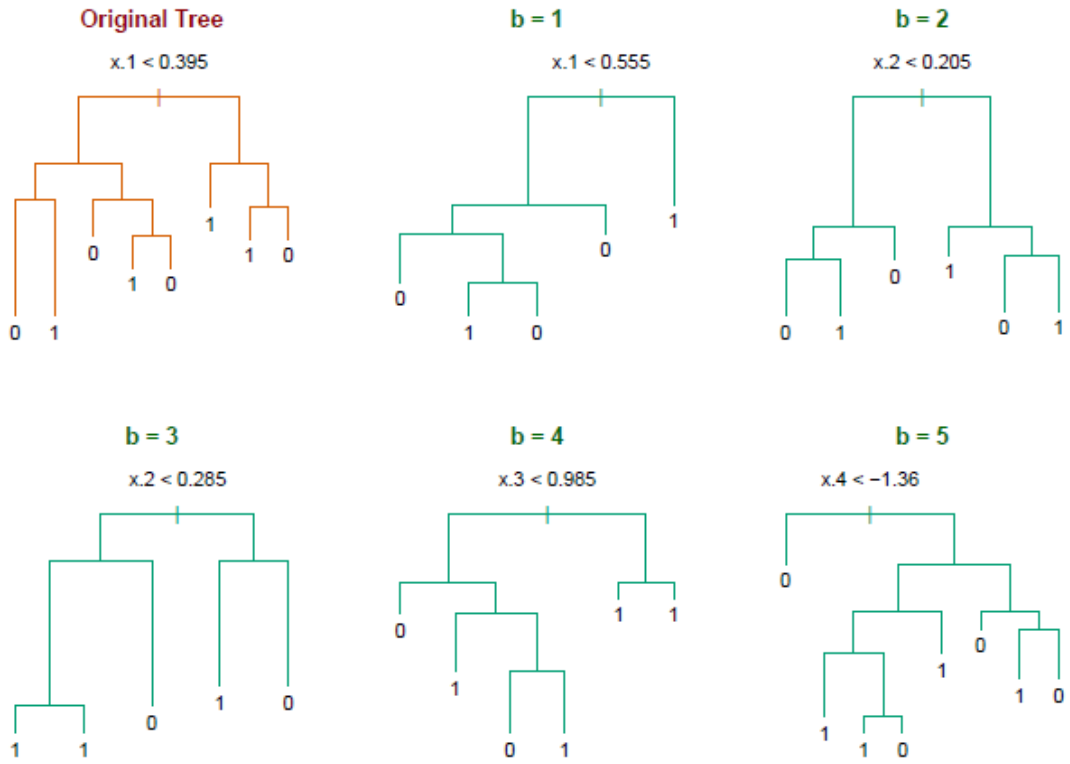
Random Forest Regressor



Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

Random Forest Classifier

- A random forest is a combination of models, called **model ensemble**
- Notice the bootstrap trees are different than the original tree



Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

Random Forest issues

Suppose that there is one very strong predictor in the data set, along with a number of other moderately strong predictors.

Then all bagged trees will select the strong predictor at the top of the tree and therefore all trees will look similar.

How do we avoid this?

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

Random Forest issues

- Differently from the regular decision trees, in random forest if there are F input features, a number f ≪ F is specified such that at each node, f features are selected at random out of the F (‘random feature selection’)

- The best split on these f is used to split the node

- The value of f is maintained fixed during the forest growing

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

Random Forest issues

- The inventors make the following recommendations:
  - For classification, the default value for *f* is *√F* and the minimum node size is 1.
  - For regression, the default value for f is *F/3* and the minimum node size is 5.

In practice the best values for these parameters will depend on the problem, and they should be treated as tuning parameters (hyperparameters)

The robustness of the classifier arises from bootstrapping of the training data and the random selection of features

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

Random Forest Issues

When the number of attributes is large, but the fraction of relevant attributes is small, random forests are likely to perform poorly

Why?

Because:

At each split the chance can be small that the relevant attributes will be selected

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

Can RF overfit?

- Random forests do not overfit the data as the number of trees increases

- Each tree is grown to the largest extent possible and there is no pruning

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

Bagging

Reduces overfitting (variance)

Normally uses one type of classifier

Decision trees are popular

Random forests often outperform other methods

Easy to parallelize

It runs efficiently on large databases

It can handle thousands of input features

It gives estimates of what features are important in the classification

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

Variable Importance Measures

- Bagging results in improved accuracy over prediction using a single tree
- Unfortunately, difficult to interpret the resulting model.
- Bagging improves prediction accuracy at the expense of interpretability.

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

Out-of-Bag Error Estimation

- Each example left out in the construction of the k-th tree is passed down through the k-th tree to get a classification ('vote').
- We can predict the class for the *i-th* example using each of the trees in which that observation was OOB and do this for *n* examples
- These votes are aggregated to obtain the majority class j: the proportion of times that j is not equal to the true class of each example averaged over all examples is the "oob error estimate"

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

Random Forests

- Web site of the “inventor”: https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm
- Breiman, L.: Random forests. Machine Learning 45(1), 5–32 (2001)

- Random forests are popular and are included in every ML/STAT package

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021