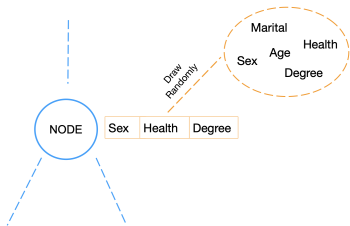


Introduction to Machine Learning

Random Forest In a Nutshell

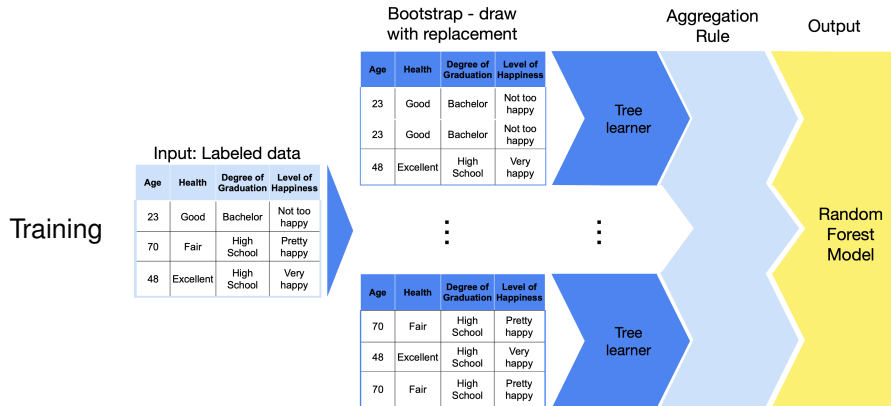


Learning goals

- Understand basic concept of random forest
- Know basic aggregation rules
- Understand concept of feature importance

LEARNING AND PREDICTION WITH RF

- Stabilizes tree learner by bagging (bootstrap aggregation)
- Randomizes tree learner and combines models into one meta model
- Can be adapted to learning task, i.e., classification or regression



LEARNING AND PREDICTION WITH RF

Prediction

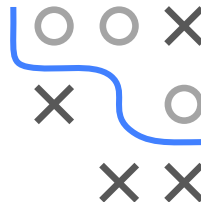
Input: Unlabeled data

Age	Health	Degree of Graduation	Level of Happiness
41	Fair	Bachelor	?
35	Good	Bachelor	?
22	Fair	High School	?

Random
Forest
Model

Prediction

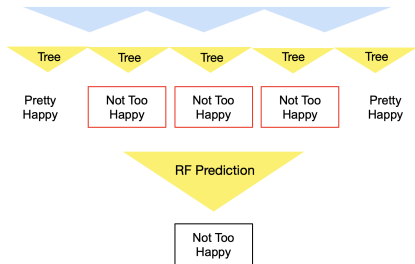
Level of Happiness
Not too happy
Pretty happy
Not too happy



AGGREGATION RULES FOR DIFFERENT TASKS

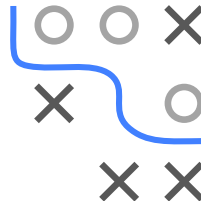
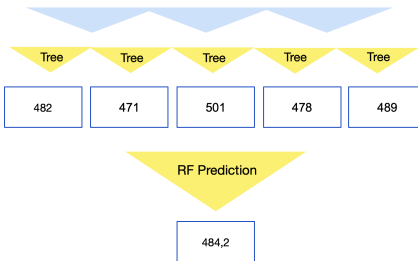
Classification Task - Majority Vote

Age	Health	Degree of Graduation	Level of Happiness
41	Fair	Bachelor	?



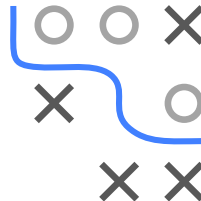
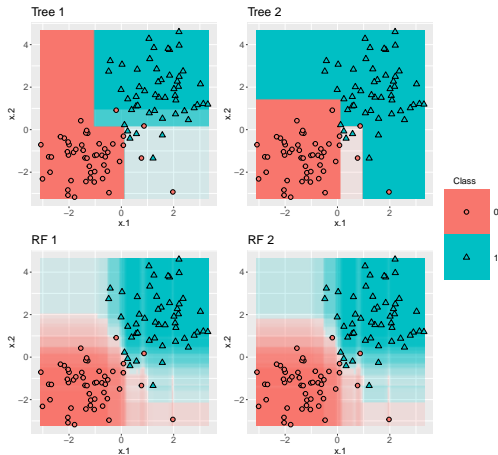
Regression Task - Averaging

Rating	Income	Credit Limit	Credit Card Balance
107	32.318	4351	?



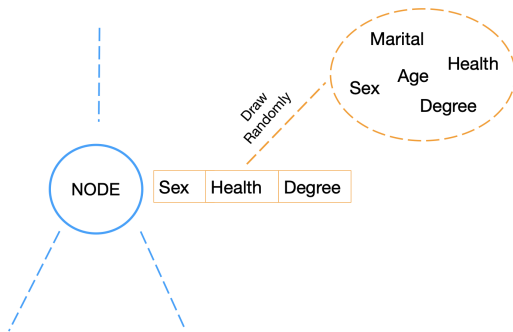
PERFORMANCE OF RF

- RF performs well for classification tasks:
 - Two different trees → Quite different decision regions
 - Two different RFs → Similar decision regions



PERFORMANCE OF RF

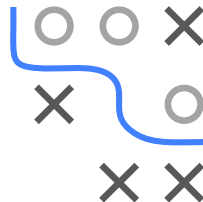
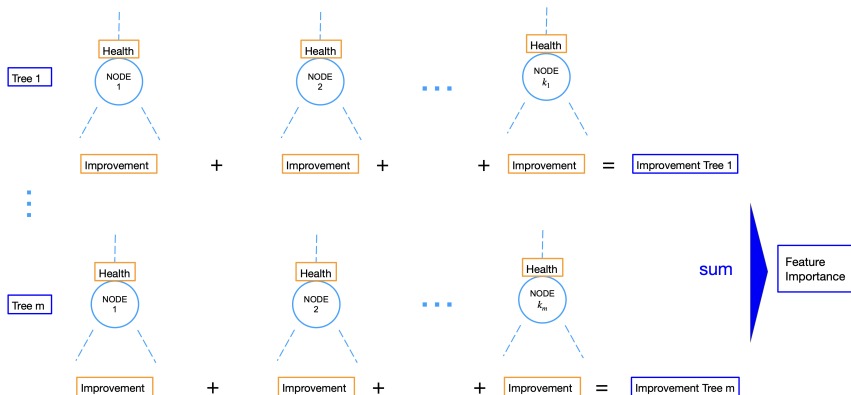
- Trees should be decorrelated, i.e., make mistakes in different directions
- Avoid correlation by
 - Bootstrap sampling
 - Randomized splits. In each node of each tree, consider different features for splitting:



FEATURE IMPORTANCE

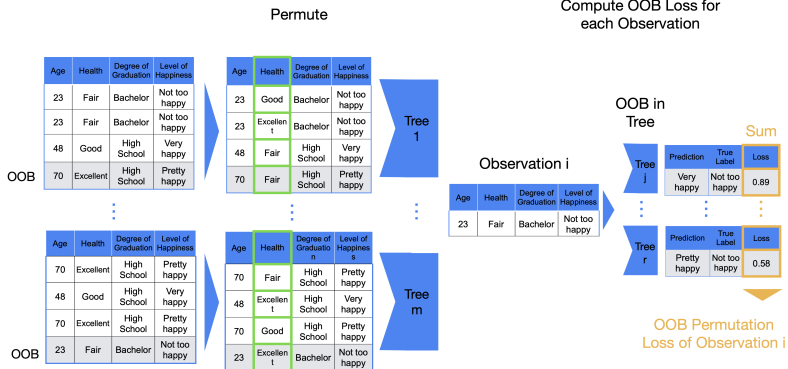
Several options, e.g., measure contribution of feature to model:

- Measure based on improvement in splitting criterion
- E.g. Feature importance of 'Health', search all nodes with 'Health' as splitting variable:



FEATURE IMPORTANCE

- Measure based on OOB Loss



FEATURE IMPORTANCE

