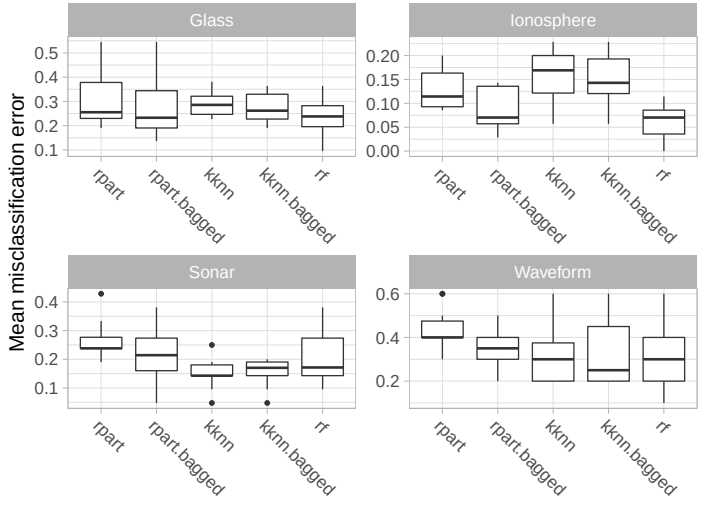
Introduction to Machine Learning

Random Forest: Benchmarking Trees, Forests, and Bagging K-NN



Learning goals

- Understand for which kind of learners bagging can improve predictive power

BENCHMARK: RANDOM FOREST VS. (BAGGED) CART VS. (BAGGED) K-NN

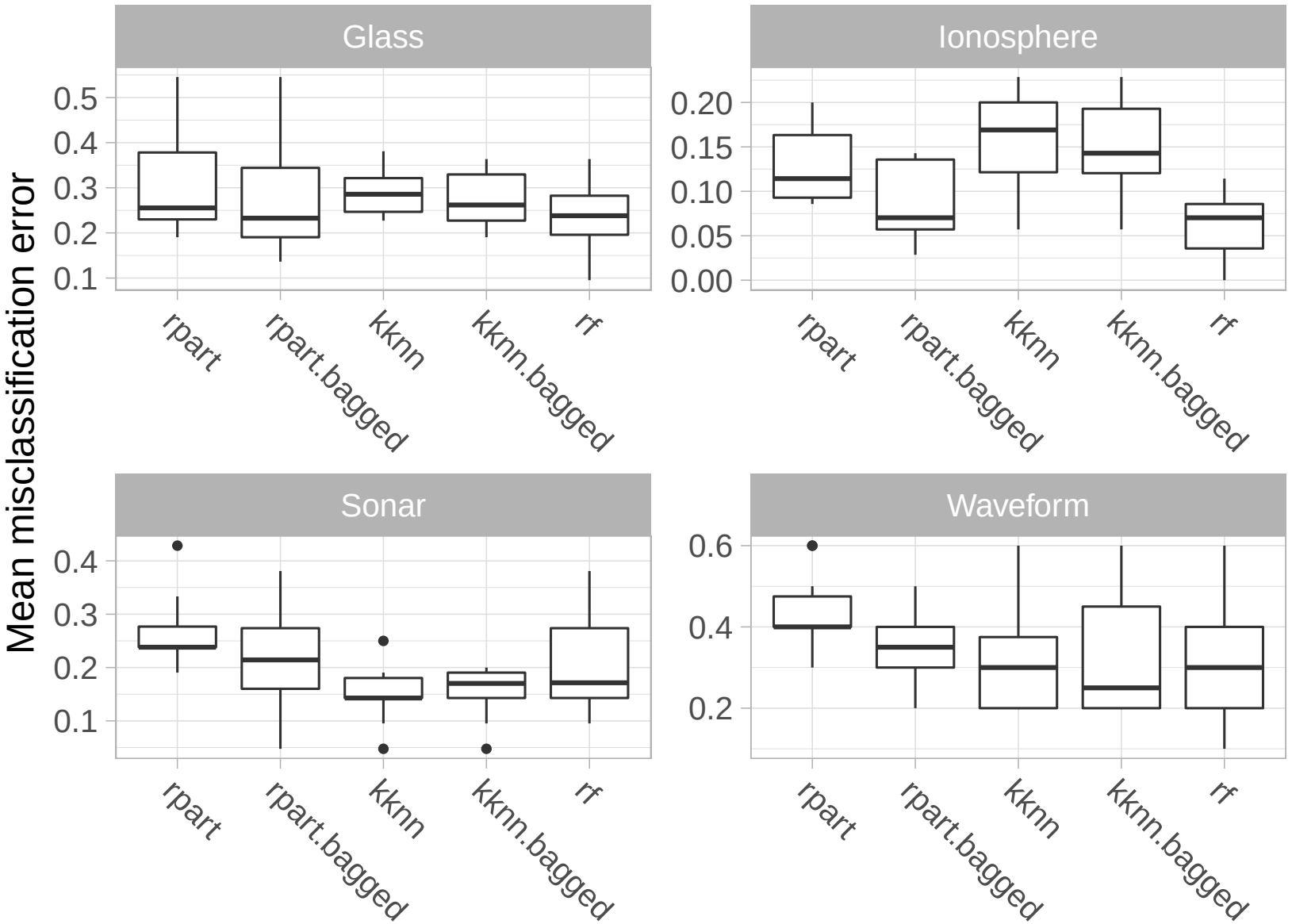
- Goal: Compare performance of random forest against (bagged) stable and (bagged) unstable methods
- Algorithms:
  - classification tree (CART, implemented in `rpart`, `max.depth`: 30, `min.split`: 20, `cp`: 0.01)
  - bagged classification tree using 50 bagging iterations (`bagged.rpart`)
  - k-nearest neighbors (k-NN, implemented in `kknn`, $k = 7$)
  - bagged k-nearest neighbors using 50 bagging iterations (`bagged.knn`)
  - random forest with 50 trees (implemented in `randomForest`)
- Method to evaluate performance: 10-fold cross-validation
- Performance measure: mean misclassification error on test sets

BENCHMARK: RANDOM FOREST VS. (BAGGED) CART VS. (BAGGED) K-NN

- Datasets from **mlbench**:

Name	Kind of data	n	p	Task
Glass	Glass identification data	214	10	Predict the type of glass (6 levels) on the basis of the chemical analysis of the glasses represented by the 10 features
Ionosphere	Radar data	351	35	Predict whether the radar returns show evidence of some type of structure in the ionosphere (“good”) or not (“bad”)
Sonar	Sonar data	208	61	Discriminate between sonar signals bounced off a metal cylinder (“M”) and those bounced off a cylindrical rock (“R”)
Waveform	Artificial data	100	21	Simulated 3-class problem which is considered to be a difficult pattern recognition problem. Each class is generated by the waveform generator.

BENCHMARK: RANDOM FOREST VS. (BAGGED) CART VS. (BAGGED) K-NN



BENCHMARK: RANDOM FOREST VS. (BAGGED) CART VS. (BAGGED) K-NN

Bagging k-NN does not improve performance because:

- k-NN is stable w.r.t. perturbations
- In a 2-class problem, nearest-neighbor-based classification only changes under bagging if both
  - the nearest neighbor in the learning set is **not** in at least half of the bootstrap samples, but the probability that any given observation is in the bootstrap sample is 63%, which is greater than 50%,
  - and, simultaneously, the *new* nearest neighbor(s) all have a different label than the missing nearest neighbor in those bootstrap samples, which is unlikely for most regions of $\mathcal{X} \times \mathcal{Y}$.