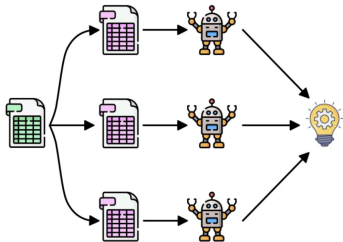


Introduction to Machine Learning

Random Forest Bagging Ensembles

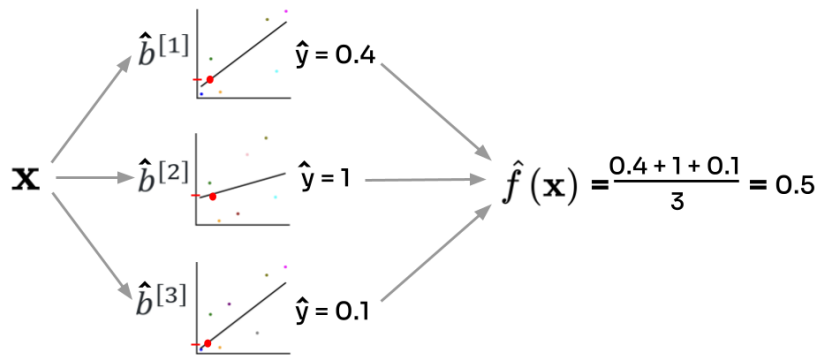


Learning goals

- Understand idea of bagging
- Be able to explain the connection between bagging and bootstrap
- Understand why bagging improves predictive performance

PREDICTING WITH A BAGGED ENSEMBLE

Average predictions of M fitted models for ensemble:
(here: regression)



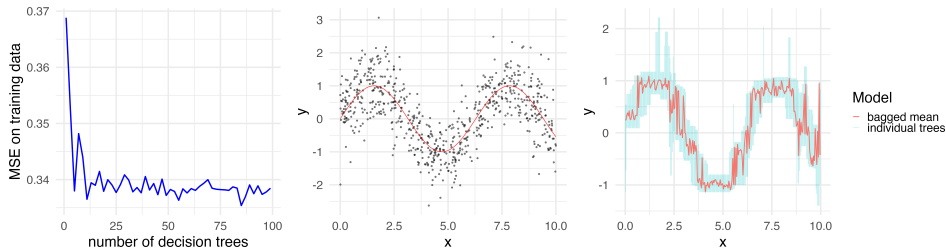
BAGGING PSEUDO CODE

Bagging algorithm: Training

- 1: **Input:** Obs. \mathbf{x} , trained BLs $\hat{b}^{[m]}$ (as scores $\hat{f}^{[m]}$, hard labels $\hat{h}^{[m]}$ or probs $\hat{\pi}^{[m]}$)
- 2: Aggregate/Average predictions

WHY/WHEN DOES BAGGING HELP?

- Bagging reduces the variability of predictions by averaging the outcomes from multiple BL models
- It is particularly effective when the errors of a BL are mainly due to (random) variability rather than systematic issues



- Increasing **nr. of BLs** improves performance, up to a point, optimal ensemble size depends on inducer and data distribution

MINI BENCHMARK

Bagged ensembles with 100 BLs each on spam:

Bagging seems especially helpful for less stable learners like CART

