Bagging + Random Forests

ways to improve existing tree-based models

bagging works for any type of model

regression tree ⟶ low bias + high variance

decrease variance = better model

bagging ⟶ average good regression trees, variance ↓

RF ⟶ average "weak" regression trees, bias ↑ variance ↓↓

Bagging

take B bootstrap samples of the data

↳ fit a good regression tree for each bootstrap sample

↳ no pruning

↳ final model an average of these B separate trees

↳ each tree a "base learner"

↳ no impact on bias, reduce variance

$$\hat{f}(x) = \frac{1}{B}\sum_{b=1}^{B} \hat{f}^{*b}(x)$$

← new

bootstrap sample will be "similar" = similar splits on each tree

Random Forest

- almost identical to bagging
- make each tree a bit more random
- for every split on every tree:
  - ↳ take a sample of $m \leq p$ predictors
  - ↳ use the best split from the m sampled
- "weak learners"
- each tree will now have some bias
- less correlation between the trees ⟶ variance ↓↓

variable importance

- decrease in RSS
- decrease in accuracy { If you remove the information in a predictor, how much worse does your model get?