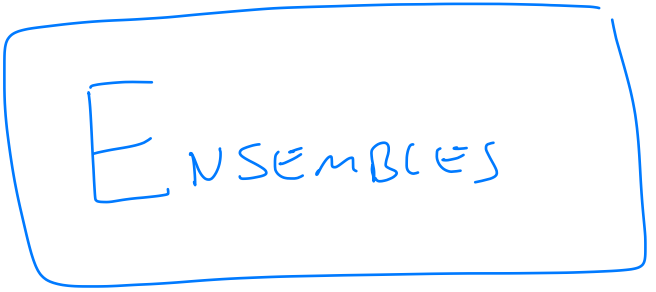
CS 307

Spring 2024

Dalpiaz



MAIN IDEA

- FIT MANY MODELS TO DATA "BASE LEARNER"
- MAKE PREDICTION W/ EACH
- "AVERAGE" PREDICTIONS

WHY?

$X_1, X_2, \ldots, X_n \sim F$

$\mathbb{E}[X_i] = \mu$

$\mathbb{V}[X_i] = \sigma^2$

$\bar{X} = \frac{1}{n} \sum X_i$

$\mathbb{E}[\bar{X}] = \mu$

$\mathbb{V}[\bar{X}] = \sigma^2/n + [\text{COVARIANCE TERM}]$

IF X_i NOT IND

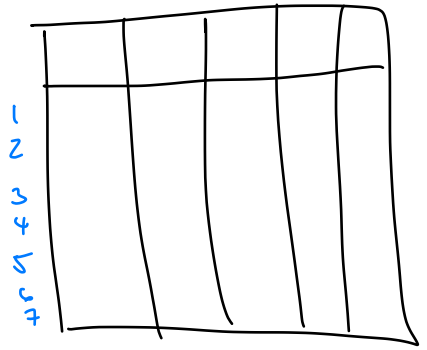
Random Forest

- Fit many deep trees in parallel
- Bootstrap data for each tree
- Random splits
- "Big" filesize
- Slower than single tree to predict
- Feature importance
- Low tuning

Boosting

- Shallow trees in sequence
- Lots of tuning
- Win Kaggle!
  - XGBoost
  - LightGBM
  - CatBoost

BOOTSTRAP



SAMPLE THE ROWS
W / REPLACEMENT

"A BOOTSTRAP RESAMPLE"