INGENIERIA INDUSTRIAL
UNIVERSIDAD DE CHILE

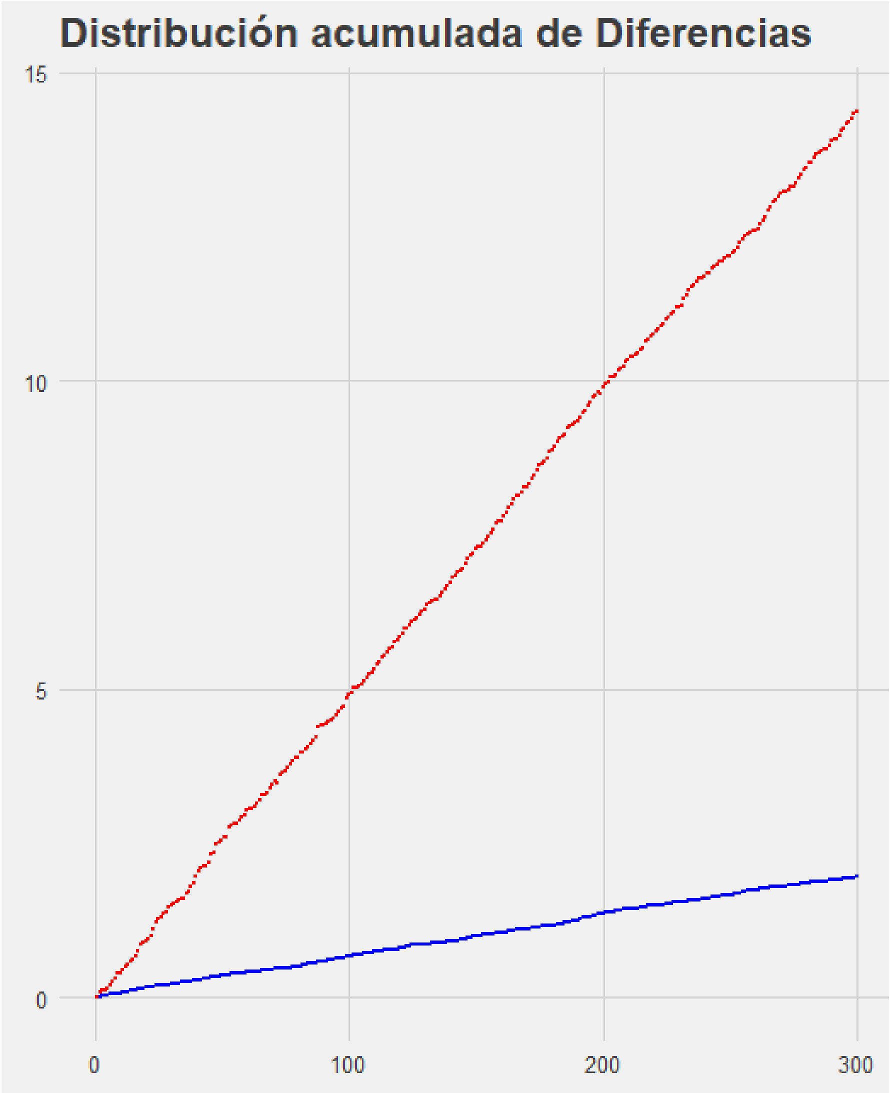
IN4402: Aplicaciones de Probabilidades y Estadística

BAGGING DECISION TREES

ANDRÉS FERNÁNDEZ

NEED FOR BAGGING
INTRODUCTION AND MAIN CONCEPTS

- Trees performs badly:
- Simulation:
  - When splitting 50/50 train and test, the difference between MSE train and test is larger for Trees than Linear Models



TAKING TREES OUT OF A BAG

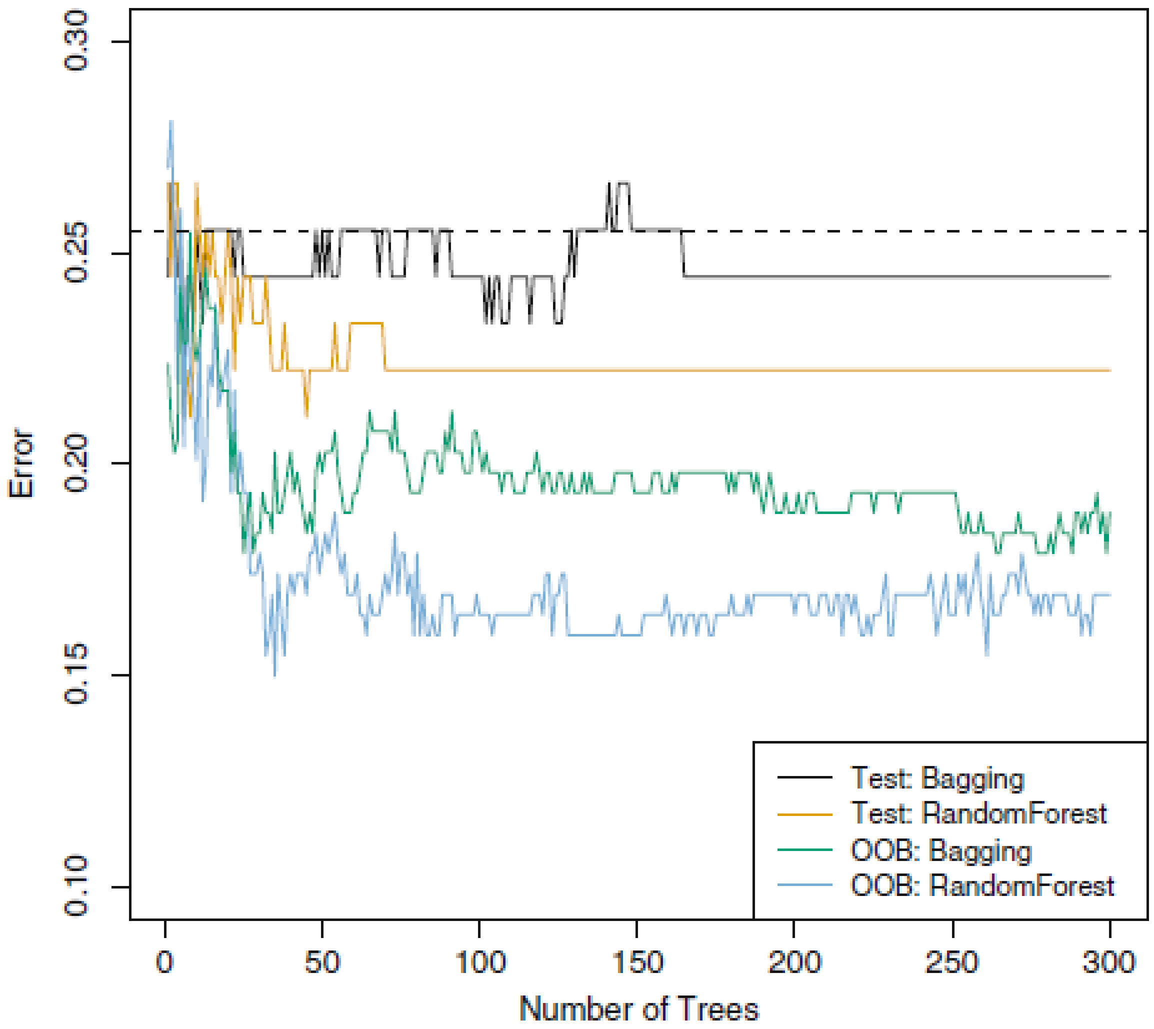
INTRODUCTION AND MAIN CONCEPTS

- "Averaging reduces variance": we average the result of many trees in **bagging**
- In the *train sample* we take out of the bag $m < N_{train}$ for B times

- In classification problems we use "majority vote" instead of "average"

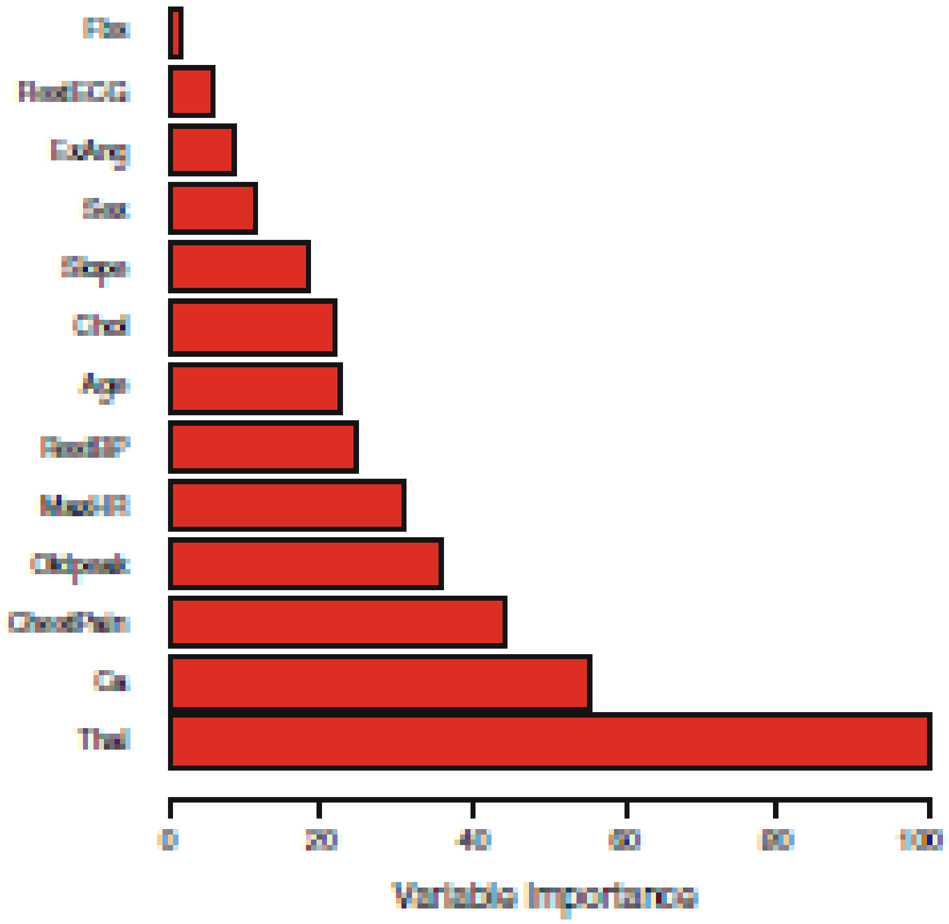
BAGGING AND OUT OF THE BAG ERROR
INTRODUCTION AND MAIN CONCEPTS

- How do we *measure* the performance?
- Out of the Bag (OOB) Error: the ones left out of the bag are used to test
  - Each observation will be left ~1/3 of the times out of the bag.
  - For every observation we can average all the predictions
  - It's an approximate cross-validation error
- Higher number of tres bagged does not overfit



VARIABLE IMPORTANCE
INTRODUCTION AND MAIN CONCEPTS

- Trees are easily interpreted: but what about the average of many of them?
- We loose interpretability with bagging performance.
  - (regression) how much each variable decreases *RSS* in average
  - (classification) hoy much each variable decreases *impurity* in average



- In Summary:
  - Bagging is a method that repeats B times the following:
    - Takes a subsample of the training sample
    - Applies Decision Trees to the subsample
  - We average errors for the observations Out Of the Bag (OOB)
  - We average error for the predicted trained observations

  - We can sort the variables according to their “importance” in building the trees on average

INGENIERIA INDUSTRIAL
UNIVERSIDAD DE CHILE

IN4402: Aplicaciones de Probabilidades y Estadística

RANDOM FORESTS

ANDRÉS FERNÁNDEZ

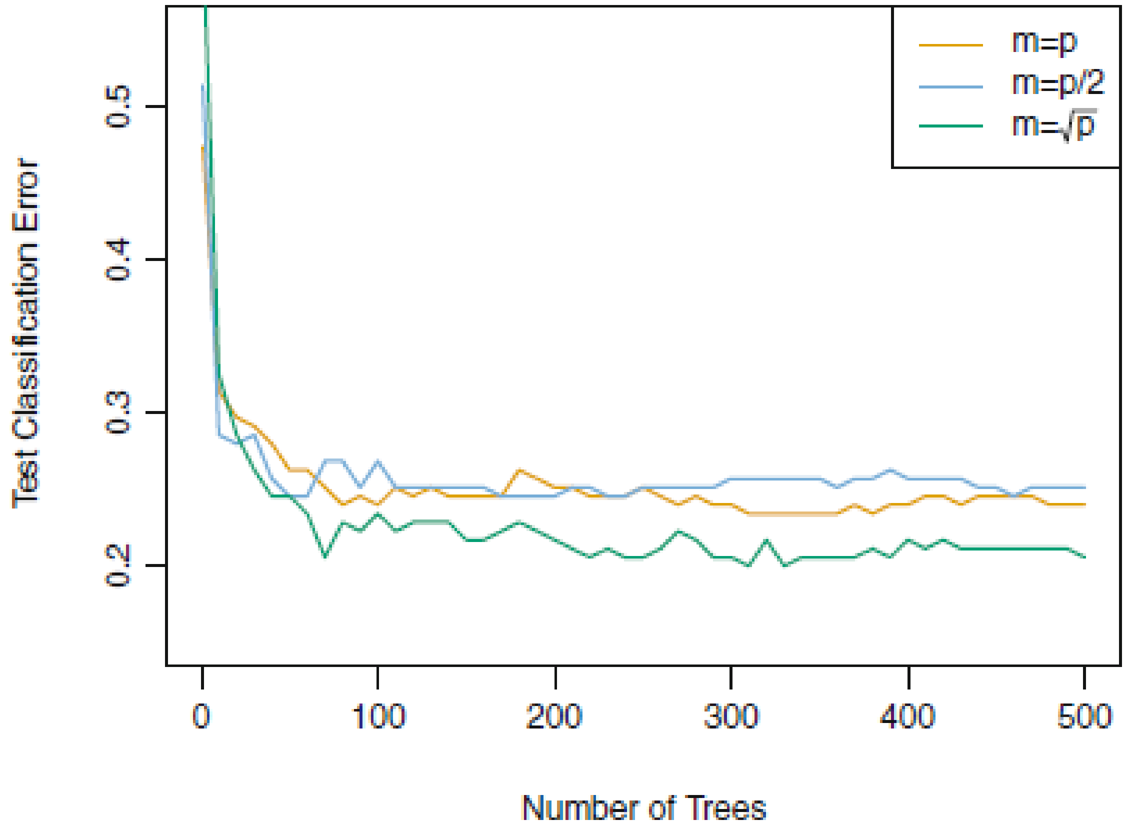
BAGGING DECISION TREES

INTRODUCTION AND MAIN CONCEPTS

- Bagging many trees might not change anything:
    - If there's an important predictor in will always be the *root*

- Then let's also sample the number of predictors we choose: **random forests**

- Random Forests *decorrelate* trees by restrincting the number of predictors:
  - How much? $m \approx \sqrt{p}$
- Since trees are independent the averaging is more robust to whatever randomness occurs, the opposite of trees wich are highly dependent on the sample they are used for.



- Finally, one could argue that particular trees estimated are more informative tan other ones:
  - We use a weighting function when averaging trees: **generalized random forests.**

- Because trees (and forests) estimate a conditional results function (how much of Y has the group that X=x)
  - It can be use to estimate **conditional treatment effects**
  - **Effect heterogeneity**

- Many policy applications: **causal trees and random forests**

INGENIERIA INDUSTRIAL
UNIVERSIDAD DE CHILE

IN4402: Aplicaciones de Probabilidades y Estadística

CAUSAL TREES AND CAUSAL RANDOM FORESTS

ANDRÉS FERNÁNDEZ

CAUSAL TREES

INTRODUCTION AND MAIN CONCEPTS

- A *regular* decisión tree (DT) or classification and regression trees (CART) predicts a results for a certain group of subjects **given a set of values of X**

$$f(X = x) = \sum_{m=1}^{p} c_m \cdot 1_{(x \in R_m)}$$

- In a way, the decision tree acts like a *matching procedure:* conditional on covariables, within a terminal leaf the observations are *very similar*
- Between leaves, the characteristics are *different.*
- If we use Y as the result of a treatment, and within each leaf we compare treated and untreated observations, we could estimate an **ATE conditional on observables: CATE**

- Heterogeneity in treatment effect

$$CATE \equiv \tau_i(x) = E(Y_i(1) - Y_i(0) \;|X = x)$$

- We assume
  - Unconfoundness $Y_i(1), Y_i(0) \perp T_i |\, X_i$
  - Overlap or common support $0 < Pr(T_i = 1 \,|X_i = x) < 1, \forall x$
- But minimizing RSS is **not** a good approach for CATE estimation
  - It produce **not consistent** estimations

- New splitting criterion: we need a term to address heterogeneity
  - We want treatment heterogeneity to be **maximum between leaves**
  - We want **balance** between treated and untreated observations

- Athey & Imbens (2016) proove that this can be achieved with a certain estimator called

Expected Mean Square Error for Treatment Effects ($\boldsymbol{EMSE_{\tau}}$)

- Maintain **balance** between treated and untreated observations
- Maximizes **accuracy** of the treatment estimation in each leaf
- CATE this way:
  - Can be **estimated** via Generalized Random Forests (GRF)
  - Has **asymptotic** behaviour so CI can be computed

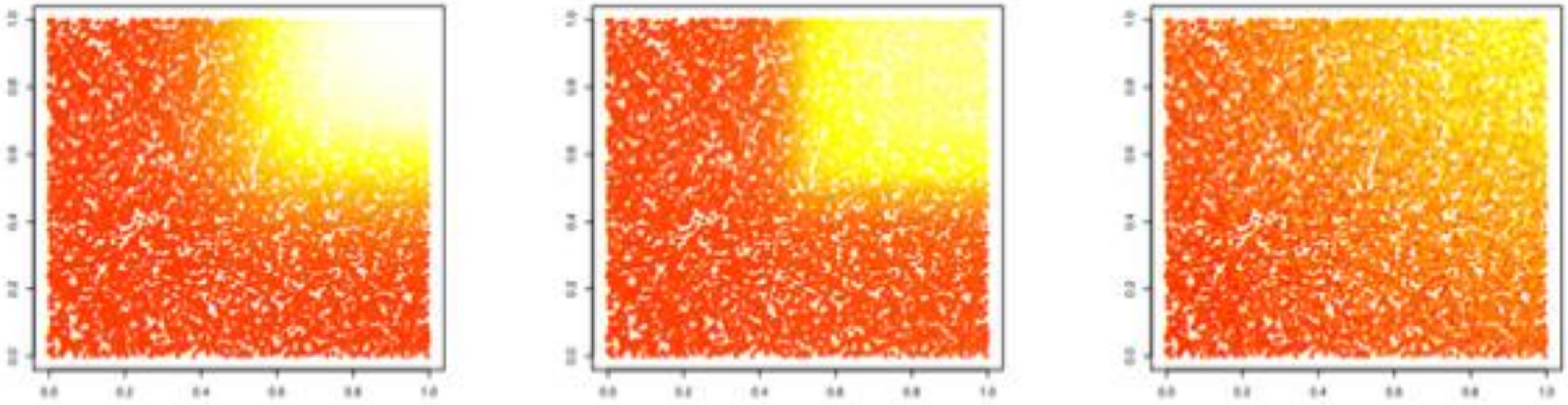
CAUSAL RANDOM FORESTS

- Because trees are unstable, we use random forest of causal trees
  - **Causal Random Forests**

- But, if we use the data to **build** the forest that maximized heterogeneity, and then also to **estimate** the CATE, then there should be bias.

- We take an **honest** approach and split the sample in **splitting/estimate** samples
  - Very much likely train/test approach
  - We use first sample to build the tree and the second one to estimate
  - We will use **honest causal random forests**

- Honest Causal Random Forests improves perfomance on estimation
  - For example, against the common K-Nearest Neighbour procedure



True effect $\tau(x)$ | Causal forest | k^*-NN

Source: https://www.causalflows.com/causal-tree-learning/

- In summary:
  - When there is need for a treatment estimation, we can use Trees to estimate a Conditional Average Treatment Effect (CATE)
    - Because leaves provide a good similarity in conditional covariables

  - We modify the splitting criterion to maximize **heterogeneity**
    - Decision Trees produce **biased** and **not consistent** estimators
    - We use $\boldsymbol{EMSE_\tau}$ as the criterion to maximize heterogeneity and balance
    - We estimate many **causal** trees to produce a **causal** random forest
  - We use splitting and estimate samples to produce an **honest** result
  - We estimate an ATE that is a function of covariables