ISO/IEC 27001 Information Security Management CERTIFIED

StoneThree
The Future of Work. Now.™

Exploratory Data Analysis

SACAC – 2025 (Lidia Auret, Tobi Louw)

Decision trees and model interpretation

Decision trees

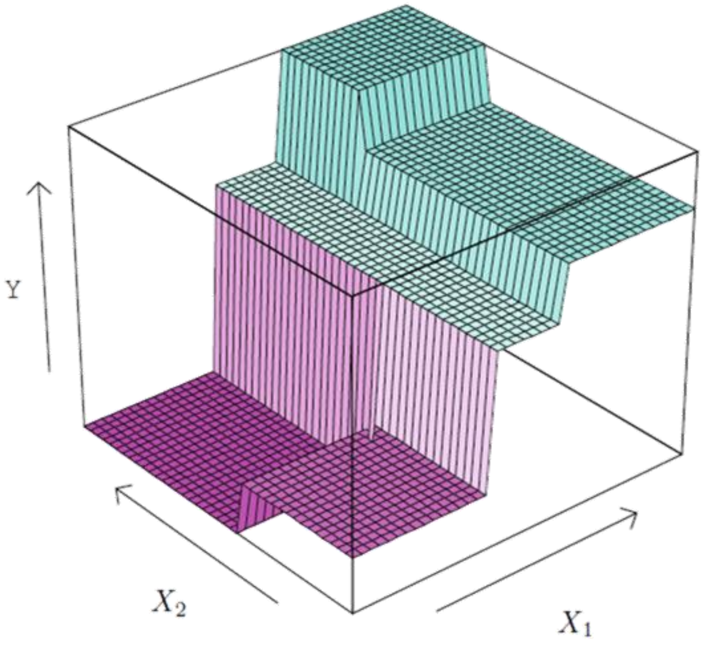
Motivation

- Motivation for nonlinear models
    - Real-world phenomena exhibit nonlinearities (e.g., local optima)
    - Nonlinearity can be non-additive (i.e., cannot be captured by linear-in-parameters regression)

- Motivation for decision trees
    - Conceptually simple, but very flexible to capture nonlinearities
    - Good building block for combining in committees (ensembles) of models

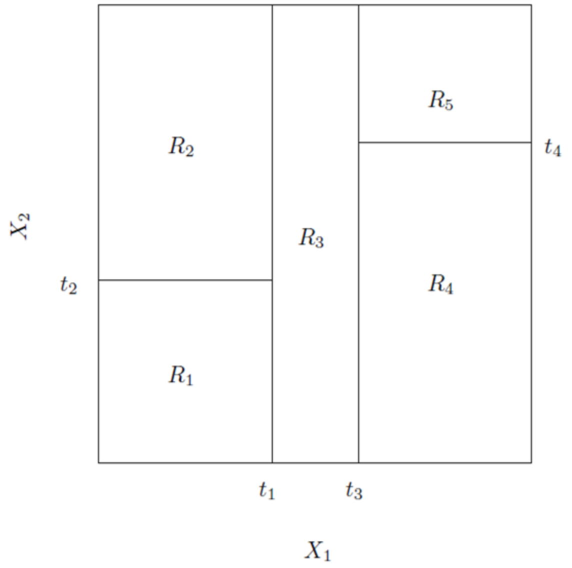
Decision trees: Concepts

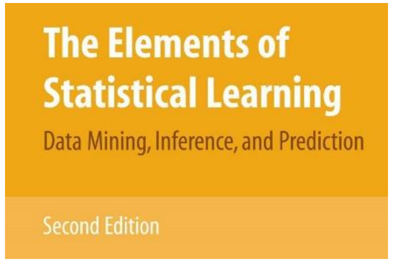
- Input space is divided into subregions
- Simple local models estimated for each subregion
- Division is recursive



E.g., Y as function of X_1 and X_2

- Partition input space into subregions $R_1, \ldots, R_5$
- Build local model for each subregion, e.g., $f(X) = c_1$ in R_1, $f(X) = c_2$ in R_2, etc.

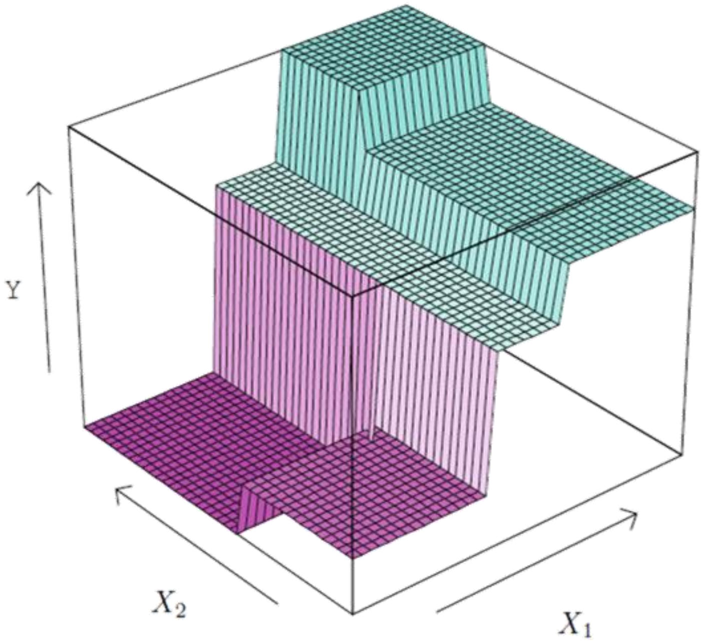




Section 9.2

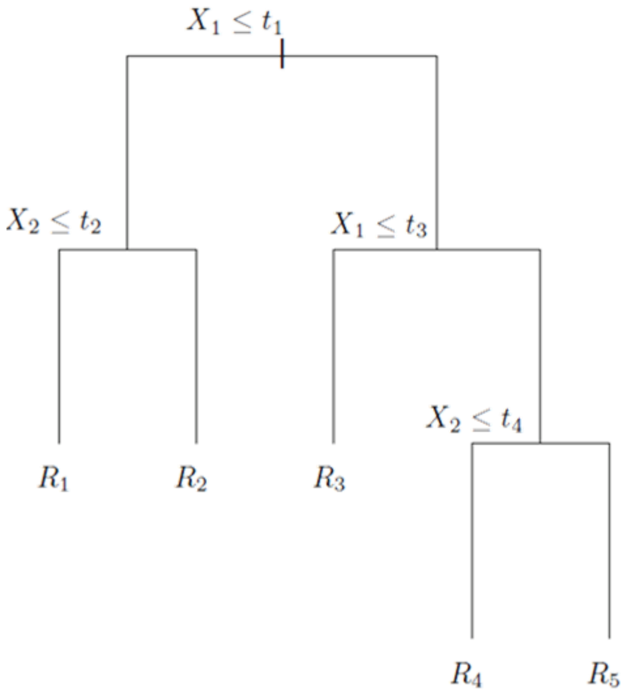
Decision trees: Concepts

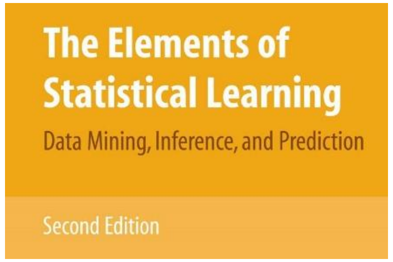
- Division is greedy/short-sighted: only the optimality of the next step, not the entire model, is considered
- Each new locally optimal division has a more homogenous response
- Division expressed as rules
- Model prediction of a new sample done by following division rules



E.g., Y as function of X_1 and X_2

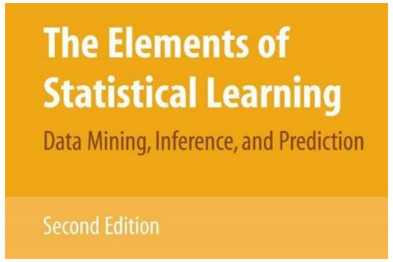
- Subregion membership based on rules, e.g., $x \in R_1$ if $x_1 \le t_1$ and $x_2 \le t_2$
- Model prediction: $f(X) = c_1$ in R_1, $f(X) = c_2$ in R_2, etc.



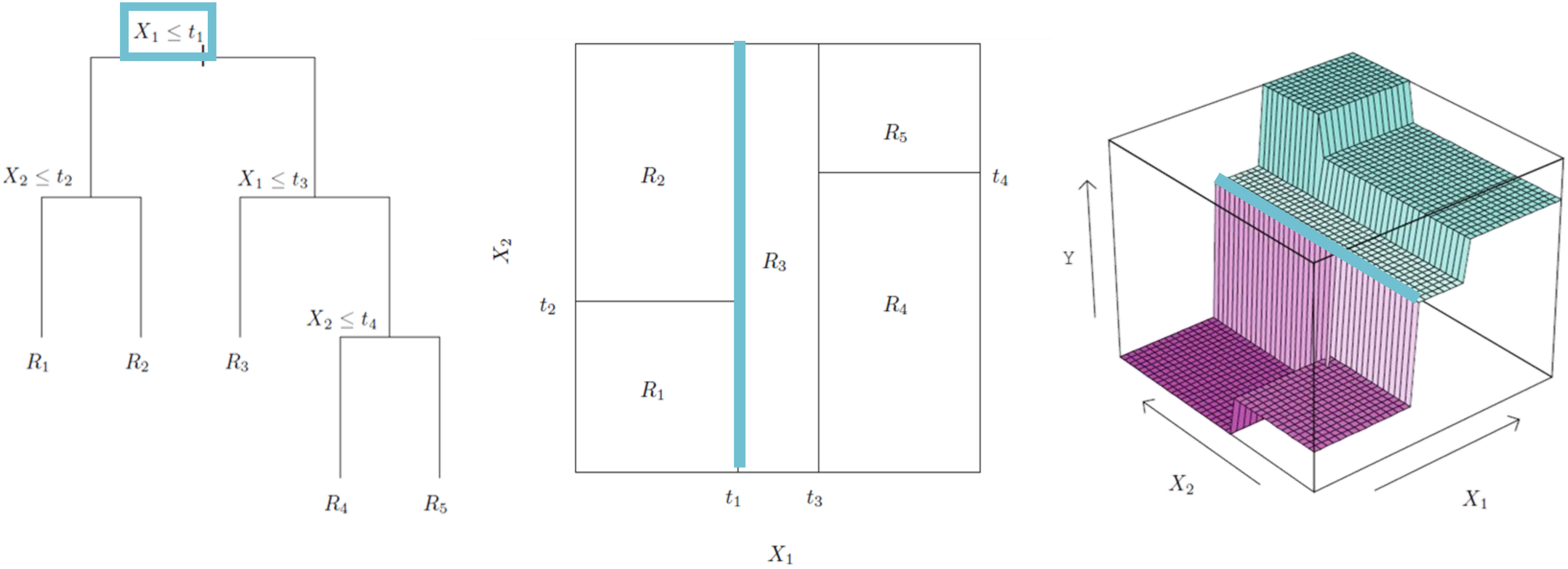


Section 9.2

Decision trees: Concepts

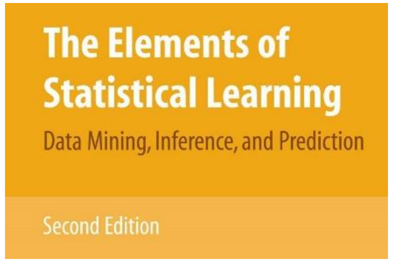


Section 9.2

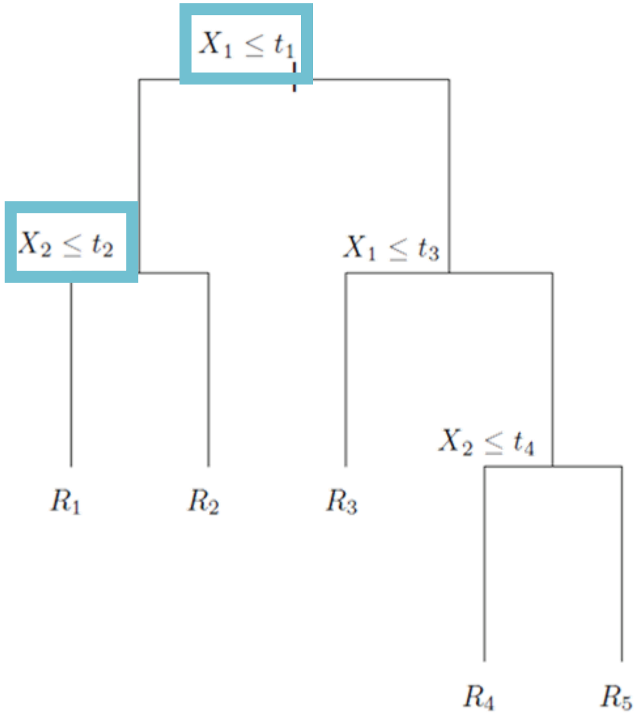


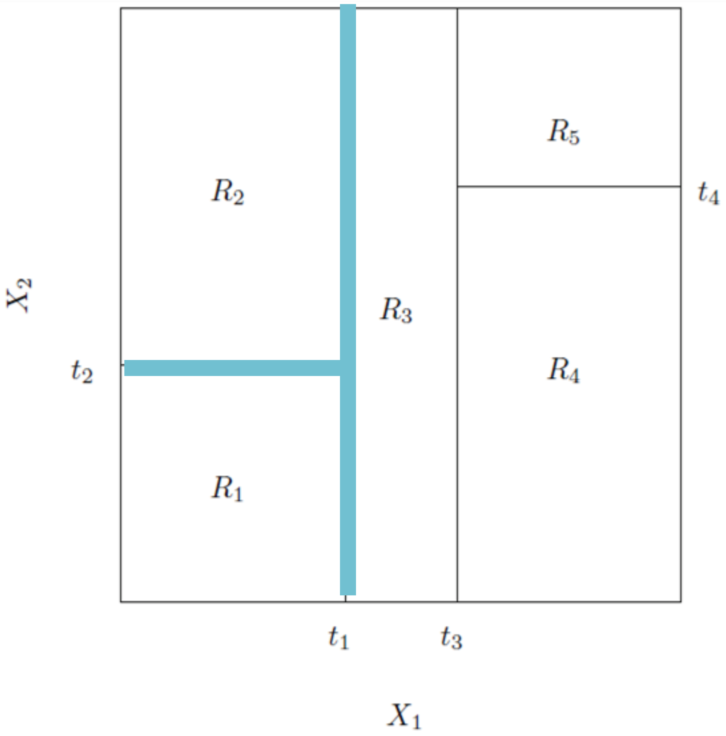
$$f(X) = \sum_m c_m I_m\{(X_1, X_2) \in R_m\}$$

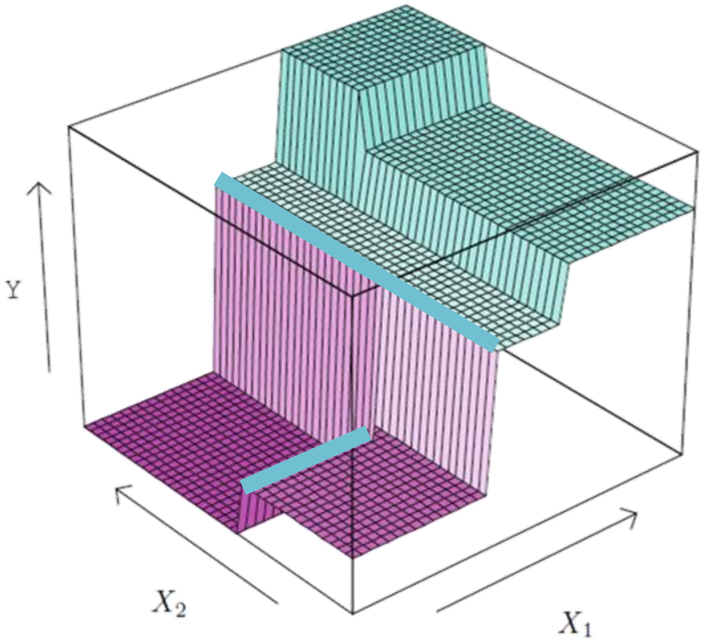
Decision trees: Concepts



Section 9.2

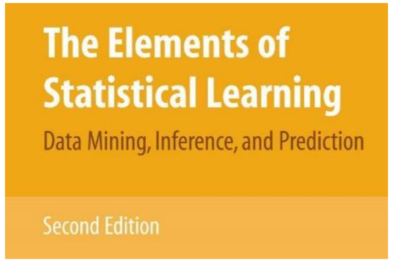




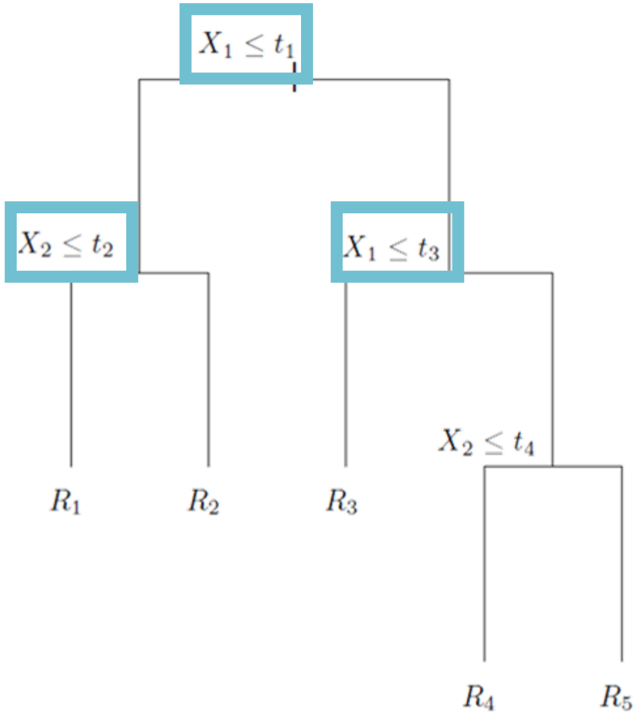


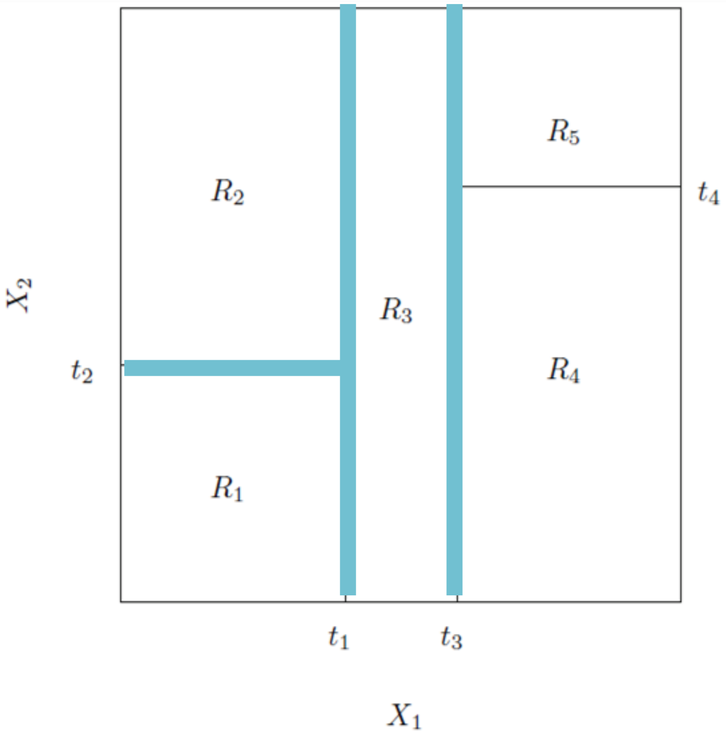
$$f(X) = \sum_m c_m I_m\{(X_1, X_2) \in R_m\}$$

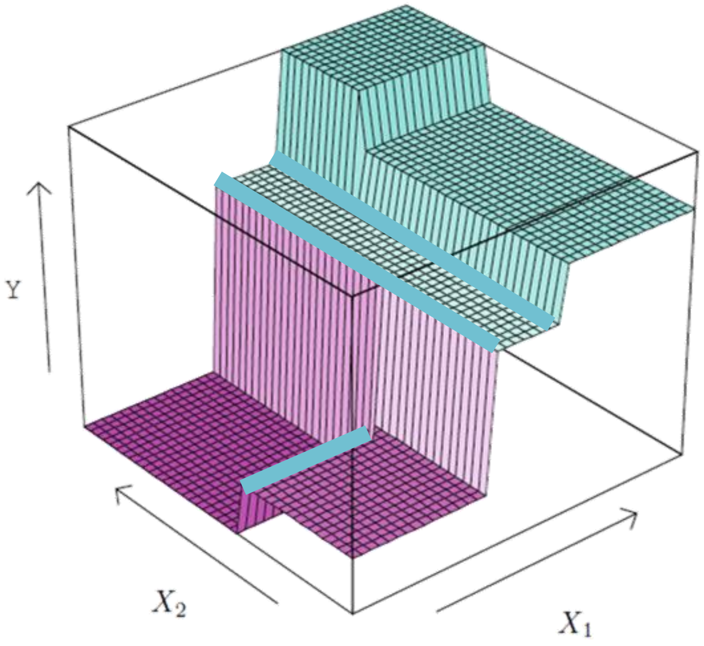
Decision trees: Concepts



Section 9.2

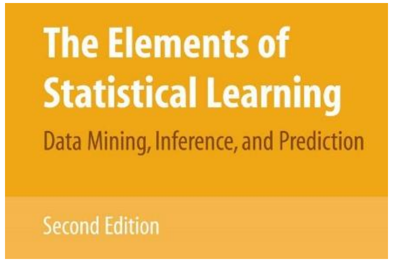




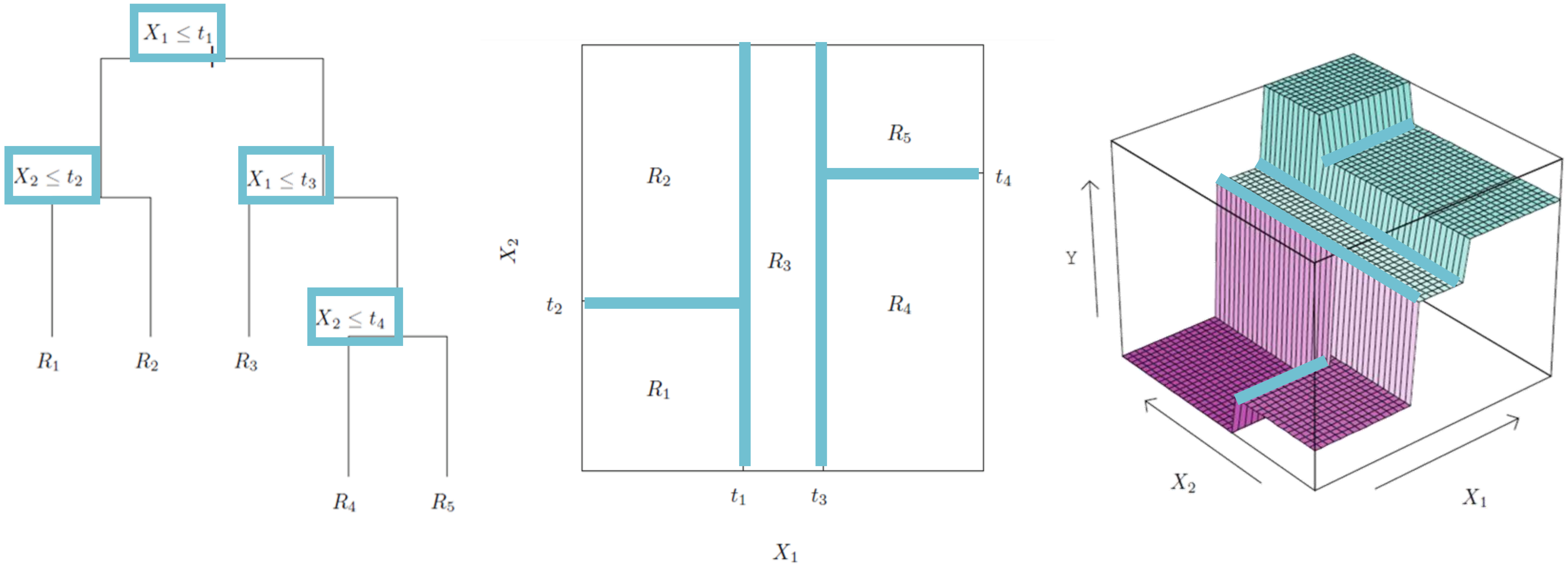


$$f(X) = \sum_m c_m I_m\{(X_1, X_2) \in R_m\}$$

Decision trees: Concepts

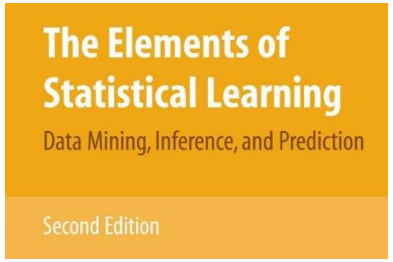


Section 9.2

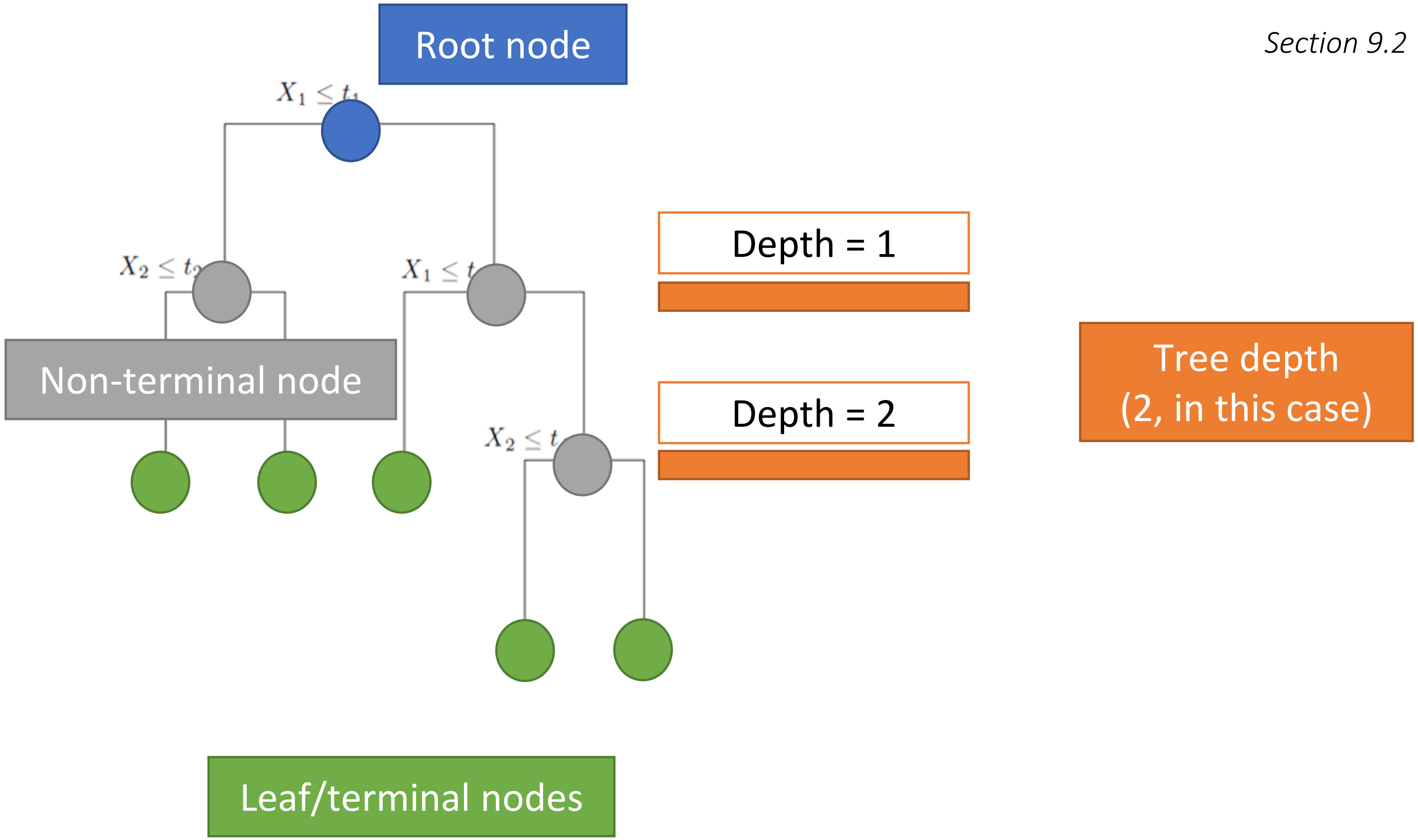


$$f(X) = \sum_m c_m I_m\{(X_1, X_2) \in R_m\}$$

Decision trees: Concepts



Section 9.2

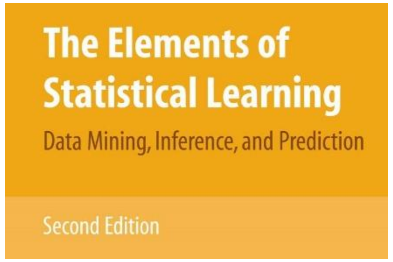


Decision trees: Regression

- What is the local model c_m?
  - Average of y_i in R_m
  - $\hat{c}_m = ave(y_i | x_i \in R_m)$
  - This minimizes sum of squares in R_m

- How to determine each new division?
  - Greedy algorithm:
    - Consider every possible split s on every possible variable j
    - Minimize the sum of squares for the pair of subregions produced by (j, s)

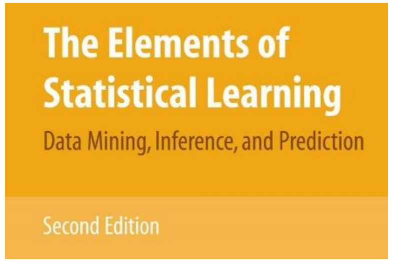
$$\min_{j,s}\left[\sum_{x_i \in R_1(j,s)} (y_i - \hat{c}_1)^2 + \sum_{x_i \in R_2(j,s)} (y_i - \hat{c}_2)^2\right]$$



Section 9.2

Decision trees: Regression

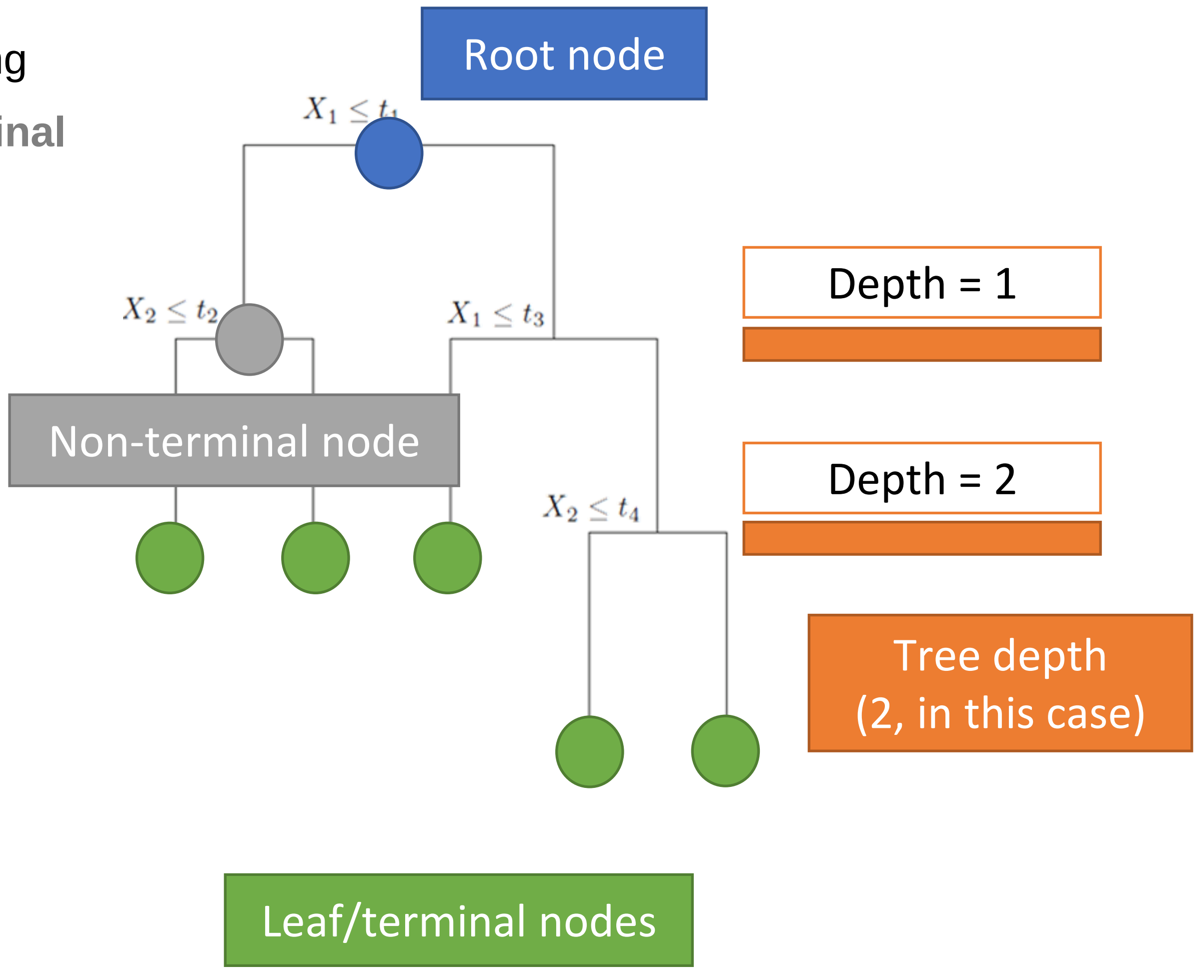
- When to stop splitting?
  - A large tree (many recursive subdivisions) may overfit the data
  - A small tree may not capture sufficient structure in the data
  - Tree size is a tuning parameter, determining model complexity
  - Tree size limited by:
    - Depth of tree
    - Minimum samples required to consider a division (split)
    - Minimum samples in a subregion (leaf node)



Section 9.2

Decision trees: Hyperparameters

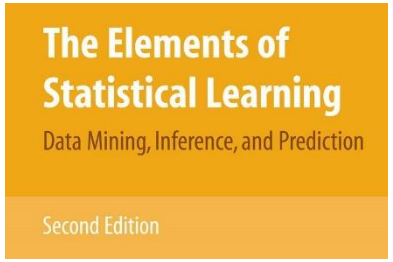
- Maximum **depth** of tree
- Minimum samples in **non-terminal node** to allow splitting
- Minimum required splitting criteria increase in **non-terminal node** to allow splitting
- Minimum samples allowed in a **leaf node**
- Maximum number of **leaf nodes**



Ensembles of trees

Ensemble methods: Concept

- Premise:
  - Combination of predictions of population of models can improve overall prediction *especially if models are uncorrelated*
- Approach:
  - Generate population of uncorrelated models
  - Combine model predictions from population



Chapter 15

Ensemble methods: Random forest

- Random forest: Generate population of uncorrelated models
  - Introducing data set variation

Bagging (bootstrap aggregation)

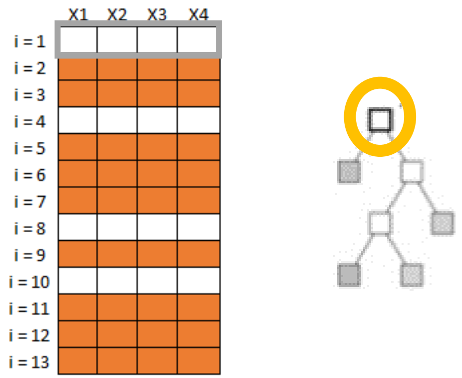
Different bootstrap sample (with replacement) of data for each tree, same original number of samples

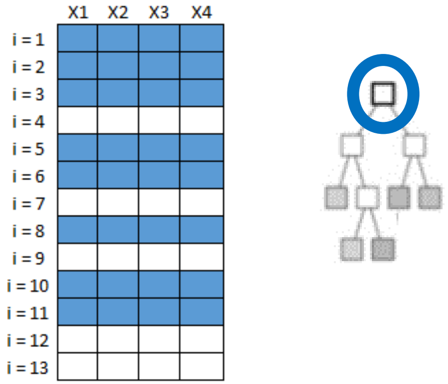
Training data set for tree 1

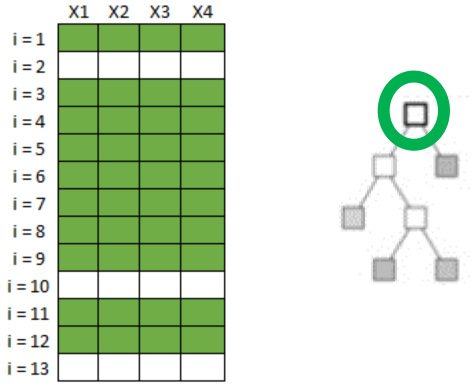
Training data set for tree 2

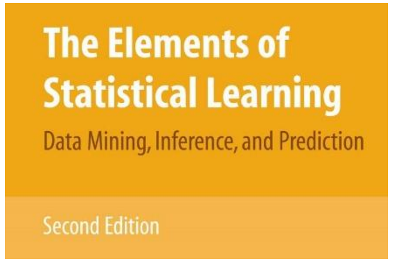
Training data set for tree 3

Out-of-bag sample for tree 1









Chapter 15

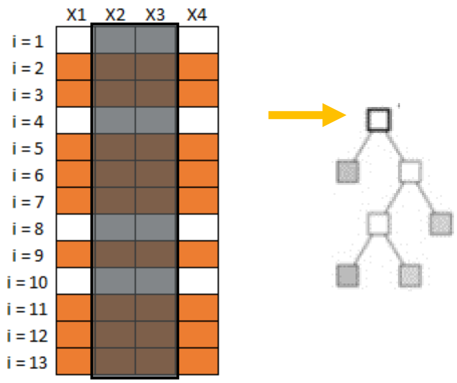
Ensemble methods: Random forest

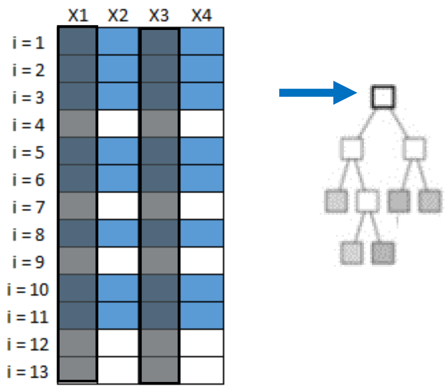
- Random forest: Generate population of uncorrelated models
  - Introducing data set variation

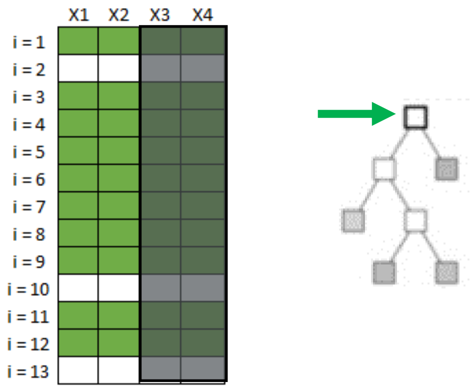
Random split selection

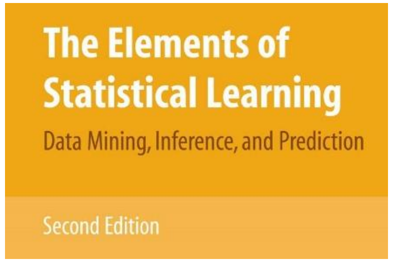
At each split opportunity, only a random subset of the variables are available for consideration

Tree 1, split 1 options	Tree 2, split 1 options	Tree 3, split 1 options





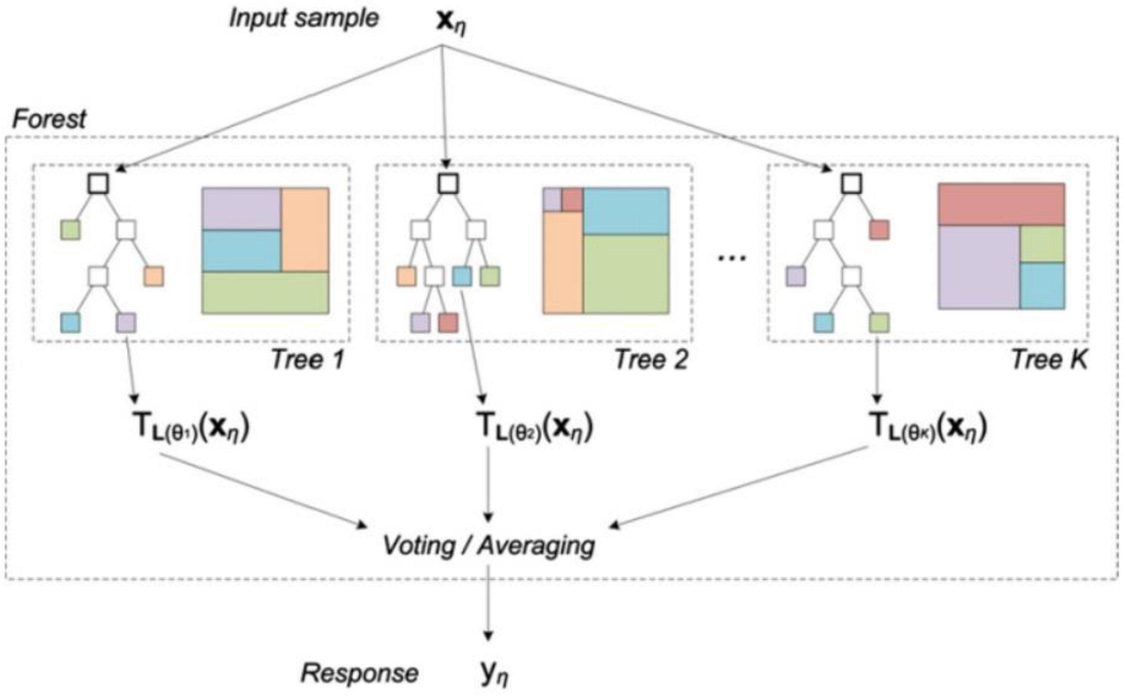


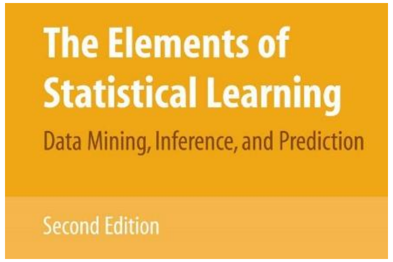


Chapter 15

Ensemble methods: Random forest

- Random forest: Combine model predictions from population
  - Regression: Average of predictions of all trees
  - Classification: Majority vote of predictions of all trees

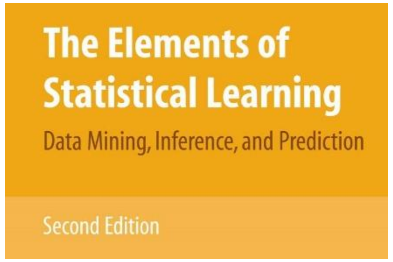




Chapter 15

Ensemble methods: Random forest

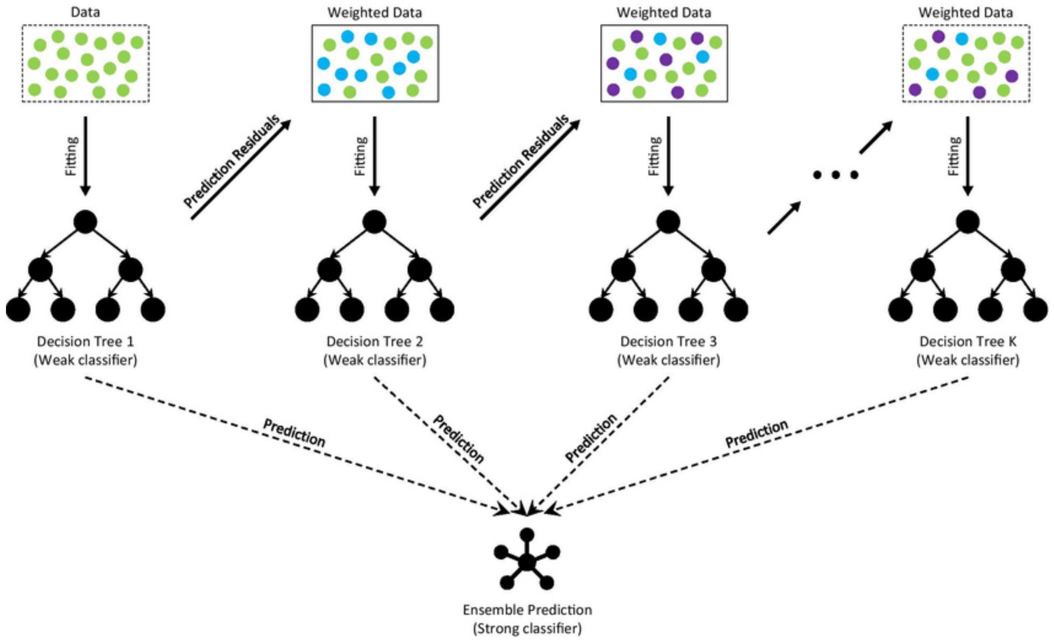
- Random forest: Hyperparameters
  - Number of trees
  - Anything from 10 to 1000
- Number of variables available for selection per split
  - Example guideline: $\sqrt{m}$ where m is number of variables in data set
- Note: Individual tree depth typically not limited



Chapter 15

Ensemble methods: Boosted trees

- Boosting
  - Base models fitted sequentially, not independently
  - Each subsequent model aims to improve on errors made by previous result
  - Preferential that base model underfits / has high bias E.g., low depth decision tree (also decreases computational cost)



Gradient boosting

Each iteration:
- Change **target** of estimator (residual error made by previous estimator), with shrinkage parameter/learning rate

Overall prediction:
- **Sum** of estimations

XGBoost: Boosting with "tricks" and heuristics

Deng, Haowen & Zhou, Youyou & Wang, Lin & Zhang, Cheng. (2021). Ensemble learning for the early prediction of neonatal jaundice with genetic features. BMC Medical Informatics and Decision Making.

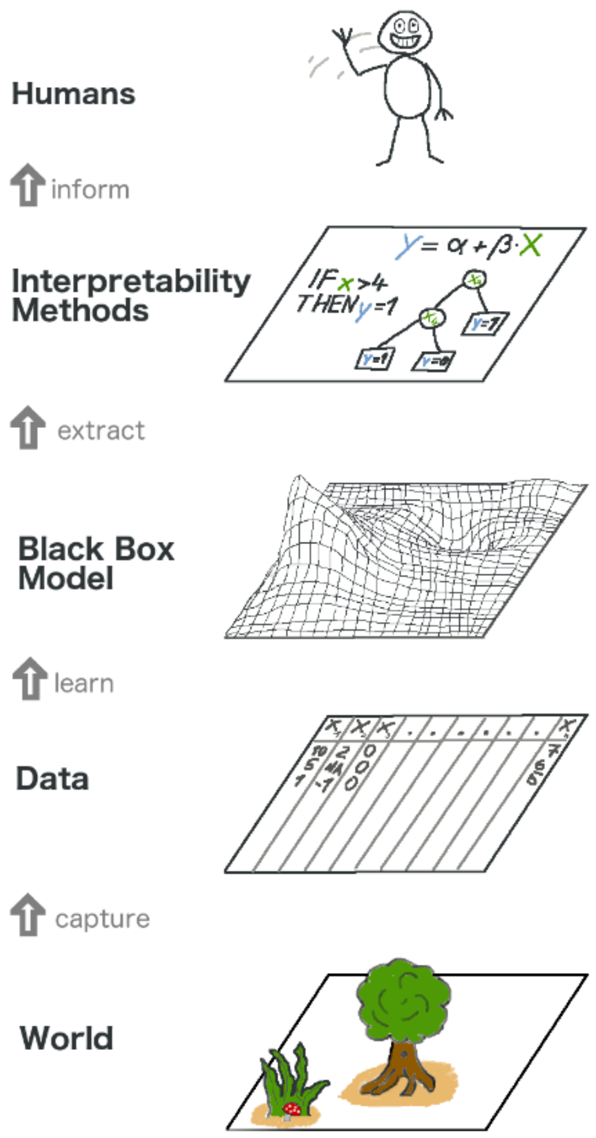
Machine Learning in Python for Process Systems Engineering

Chapter 9

Model interpretation

Machine learning: Interpretation

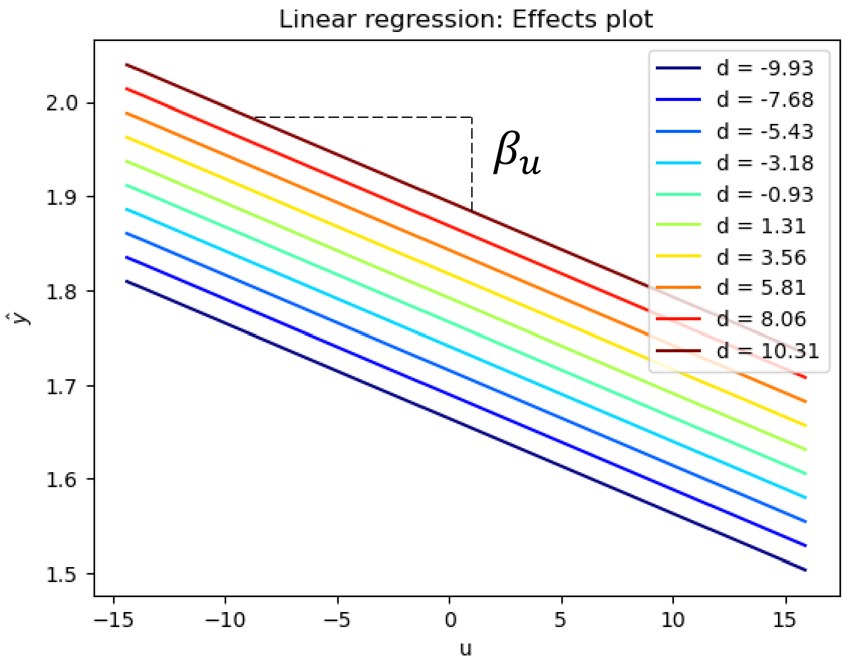
- Interpretability = the degree to which a human can understand the cause of a decision
- Importance of model interpretability:
  - Increases scientific knowledge of world
  - Increases social acceptance of model
  - Allows error-finding and auditing of model



Molnar (2023) Interpretable Machine Learning: A Guide for Making Black Box Models Explainable.
https://christophm.github.io/interpretable-ml-book/

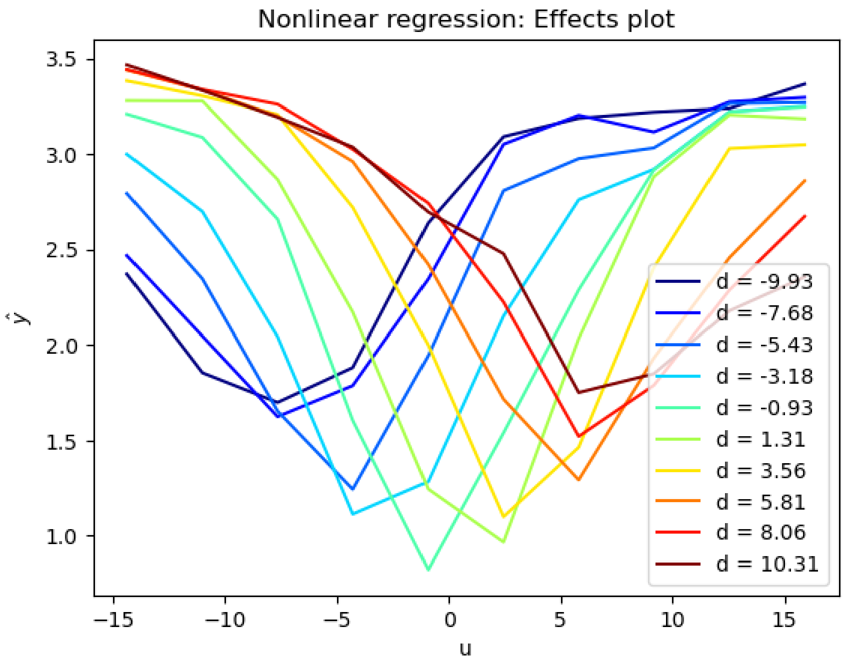
Machine learning: Interpretation

- Interpretability = the degree to which a human can understand the cause of a decision
- Linear models are typically interpretable:
  - $\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2$
  - Constant, visible effect (β_i) of input variable x_i on prediction $\hat{y}$
  - Effect typically independent of other variables



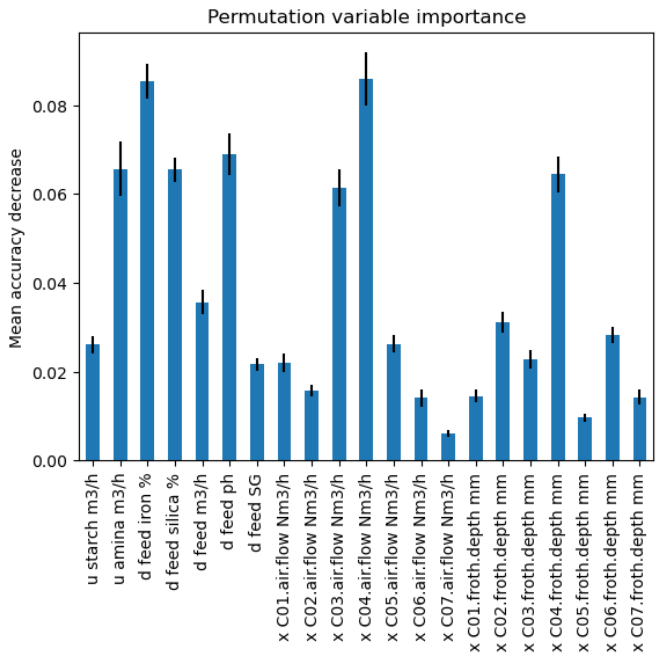
Machine learning: Interpretation

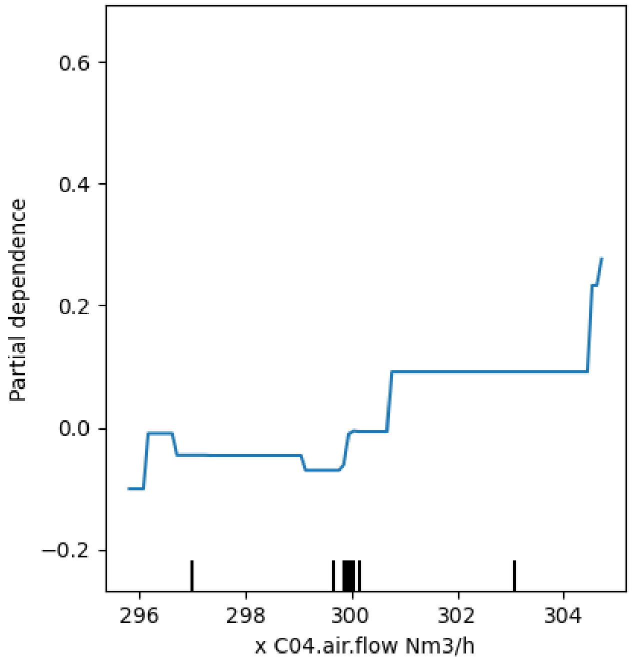
- Interpretability = the degree to which a human can understand the cause of a decision
- Machine learning models are typically “black boxes”:
  - Complex global structures (neural networks) and/or very local-based predictions (K-nn)
  - Difficult to trace effect that the change in an individual variable has on the final model prediction
  - Variable effect of input variable x_i on prediction $\hat{y}$, dependent on other input variables



Global methods for interpretation

- Global interpretation method = describe **average** behaviour of a machine learning model
- Provides expected values based on the distribution of the data
- Useful to understand the general mechanisms in the data or model
- Examples:
  - Permutation feature importance
  - Partial dependence plot





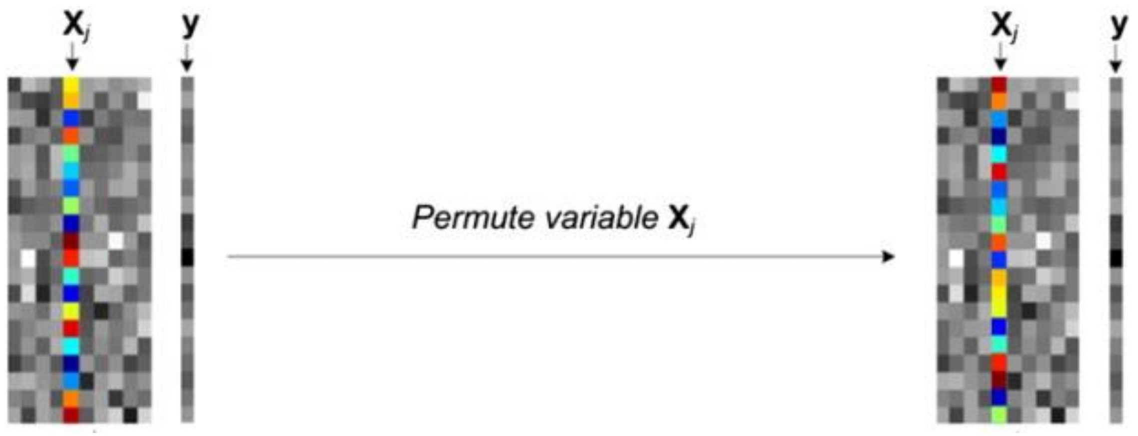
Global methods for interpretation

- **Permutation feature importance**
- Measures the increase in prediction error of the model, after a feature's value has been *permuted*

Original data $\boldsymbol{X}, \boldsymbol{y}$

Permuted data $\boldsymbol{X}_{perm}, \boldsymbol{y}$

X_j	y
1	5
2	9
3	13
4	17
5	21
6	25



X_j	y
4	5
1	9
6	13
2	17
5	21
3	25

Model predictions $f(\boldsymbol{X})$

Model error/loss e

Model predictions $f(\boldsymbol{X}_{perm})$

Model error/loss e_{perm}

Feature importance $FI_j = e_{perm} - e$

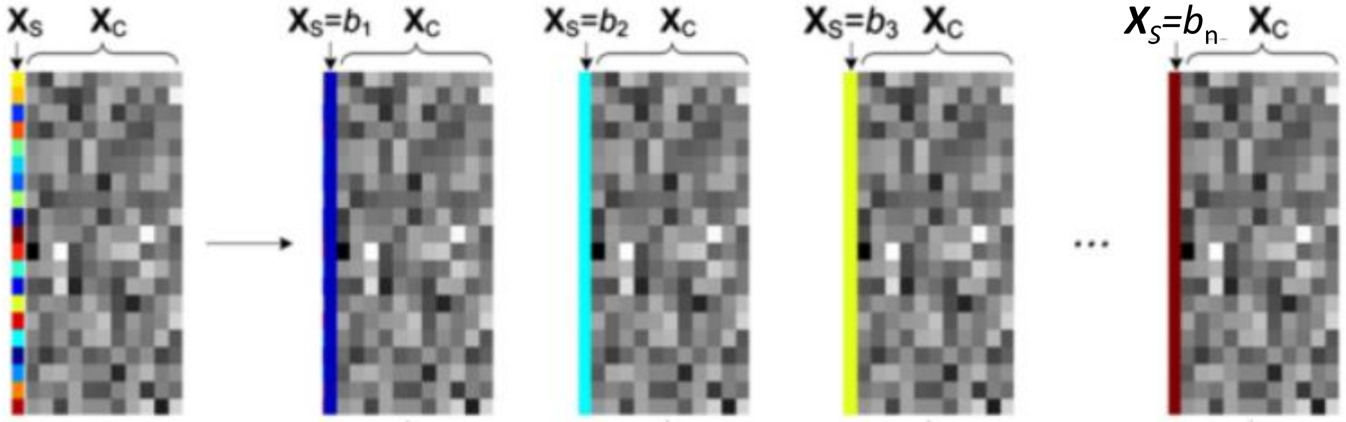
Global methods for interpretation

- **Permutation feature importance**
- Measures the increase in prediction error of the model, after a feature's value has been *permuted*
- Pseudo-algorithm:
  - **Input:** Trained model f, data $\{\boldsymbol{x}_i, y_i\}_{i=1}^{n}$ as feature matrix $\boldsymbol{X}$ and response vector $\boldsymbol{y}$, error/loss measure $L(\boldsymbol{y}, f(\boldsymbol{X}))$
  - **Result:** Feature importance FI_j for each feature (variable) j
  - **Steps:**
    1. Estimate the original model error $e_{orig} = L(\boldsymbol{y}, f(\boldsymbol{X}))$
    2. For each feature (variable) $j \in \{1, \dots, m\}$ do:
       a) Generate feature matrix $\boldsymbol{X}_{perm}$ by permuting feature j in the data $\boldsymbol{X}$ (this breaks the association between feature j and true response $\boldsymbol{y}$)
       b) Estimate loss/error $e_{perm} = L\left(\boldsymbol{y}, f\left(\boldsymbol{X}_{perm}\right)\right)$
       c) Calculate permutation feature importance $FI_j = e_{perm} - e_{orig}$

Global methods for interpretation

- **Partial dependence plot**
- Shows the *marginal* effect of one or two features have on the predicted outcome of a model
  - *Marginal:* Averaging over the effect of other features
- Feature(s) of interest are $\boldsymbol{X}_S$, rest of features in model are $\boldsymbol{X}_C$

Original data $\boldsymbol{X}_S, \boldsymbol{X}_C,$	Original data for $\boldsymbol{X}_C$; with $\boldsymbol{X}_S = b_1$	Original data for $\boldsymbol{X}_C$; with $\boldsymbol{X}_S = b_2$	Original data for $\boldsymbol{X}_C$; with $\boldsymbol{X}_S = b_3$	Original data for $\boldsymbol{X}_C$; with $\boldsymbol{X}_S = b_n$



$$\bar{f}_S(\boldsymbol{X}_s = b_1) = \frac{1}{n}\sum_i^n f(b_1, \boldsymbol{X}_{i,C})$$

$$\bar{f}_S(\boldsymbol{X}_s = b_2) = \frac{1}{n}\sum_i^n f(b_2, \boldsymbol{X}_{i,C})$$

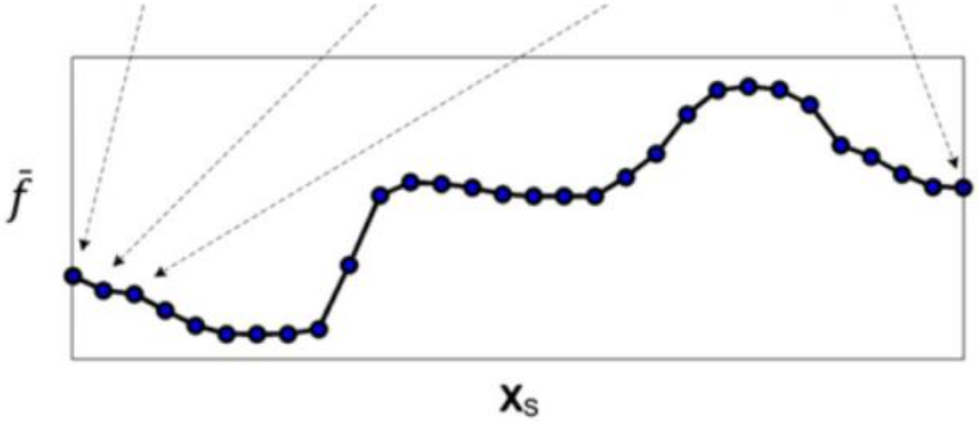
$$\bar{f}_S(\boldsymbol{X}_s = b_3) = \frac{1}{n}\sum_i^n f(b_3, \boldsymbol{X}_{i,C})$$

$$\bar{f}_S(\boldsymbol{X}_s = b_n) = \frac{1}{n}\sum_i^n f(b_n, \boldsymbol{X}_{i,C})$$

Global methods for interpretation

- **Partial dependence plot**
- Shows the *marginal* effect of one or two features have on the predicted outcome of a model
  - *Marginal:* Averaging over the effect of other features
- Feature(s) of interest are $\boldsymbol{X}_S$, rest of features in model are $\boldsymbol{X}_C$

Original data $\boldsymbol{X}_S, \boldsymbol{X}_C,$	Original data for $\boldsymbol{X}_C$; with $\boldsymbol{X}_S = b_1$	Original data for $\boldsymbol{X}_C$; with $\boldsymbol{X}_S = b_2$	Original data for $\boldsymbol{X}_C$; with $\boldsymbol{X}_S = b_3$	Original data for $\boldsymbol{X}_C$; with $\boldsymbol{X}_S = b_n$



$\bar{f}_S(\boldsymbol{X}_s = b_1) = \frac{1}{n}\sum_i^n f(b_1, \boldsymbol{X}_{i,C})$

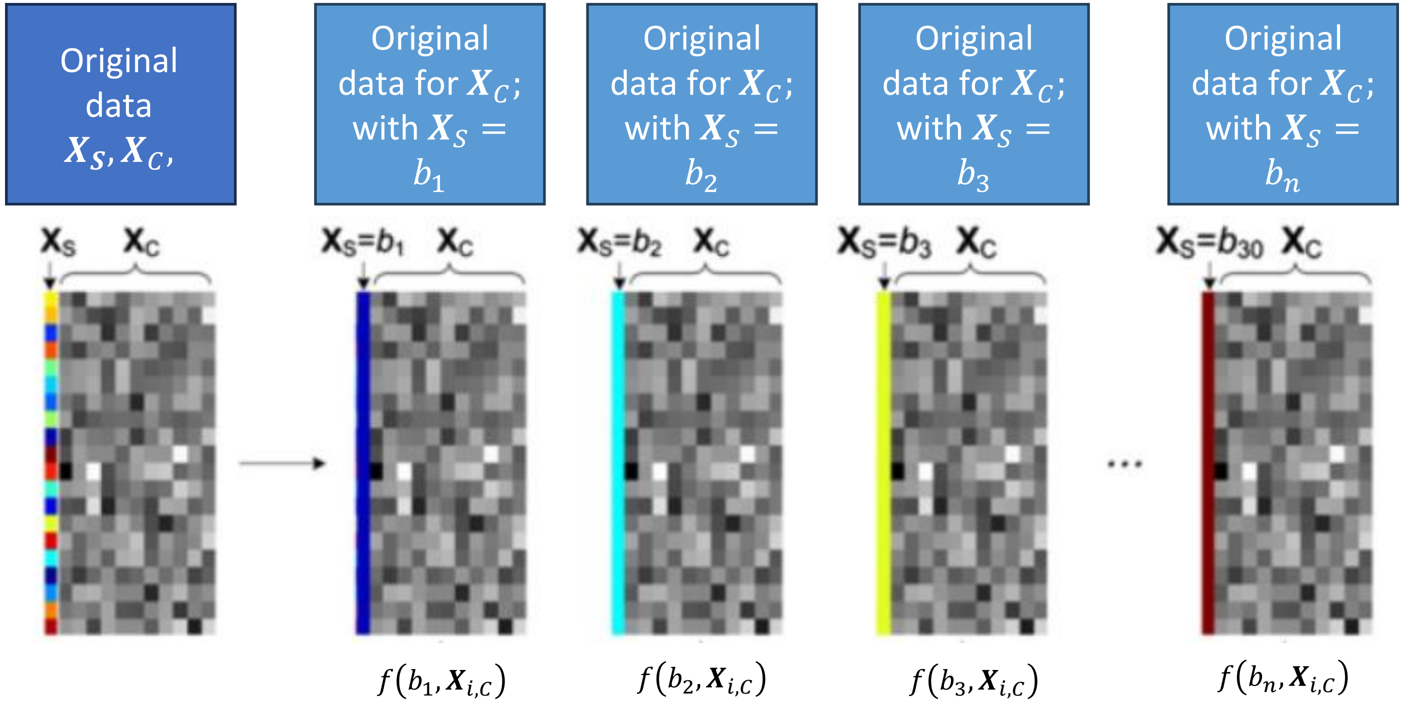
$\bar{f}_S(\boldsymbol{X}_s = b_2) = \frac{1}{n}\sum_i^n f(b_2, \boldsymbol{X}_{i,C})$

$\bar{f}_S(\boldsymbol{X}_s = b_3) = \frac{1}{n}\sum_i^n f(b_3, \boldsymbol{X}_{i,C})$

$\bar{f}_S(\boldsymbol{X}_s = b_n) = \frac{1}{n}\sum_i^n f(b_n, \boldsymbol{X}_{i,C})$

Local methods for interpretation

- **Individual conditional expectation (ICE) plot**
- Shows the *individual* effect one or two features have on the predicted outcome of a model
- Feature(s) of interest are $\boldsymbol{X}_S$, rest of features in model are $\boldsymbol{X}_C$



Local methods for interpretation

- **Individual conditional expectation (ICE) plot**
- Shows the *individual* effect one or two features have on the predicted outcome of a model
- Feature(s) of interest are $\boldsymbol{X}_S$, rest of features in model are $\boldsymbol{X}_C$

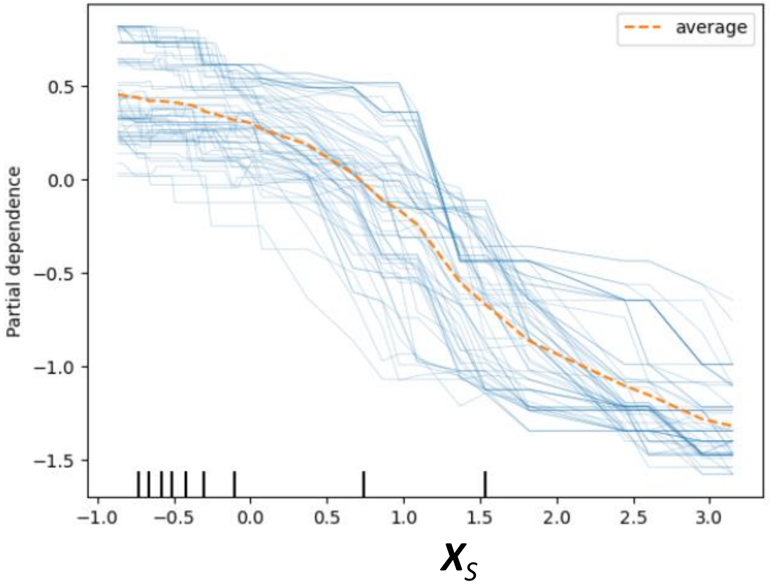
Original data $\boldsymbol{X}_S, \boldsymbol{X}_C, \boldsymbol{y}$

Original data for $\boldsymbol{X}_C, \boldsymbol{y}$; with $\boldsymbol{X}_S = b_1$

Original data for $\boldsymbol{X}_C, \boldsymbol{y}$; with $\boldsymbol{X}_S = b_2$

Original data for $\boldsymbol{X}_C, \boldsymbol{y}$; with $\boldsymbol{X}_S = b_3$

Original data for $\boldsymbol{X}_C, \boldsymbol{y}$; with $\boldsymbol{X}_S = b_n$

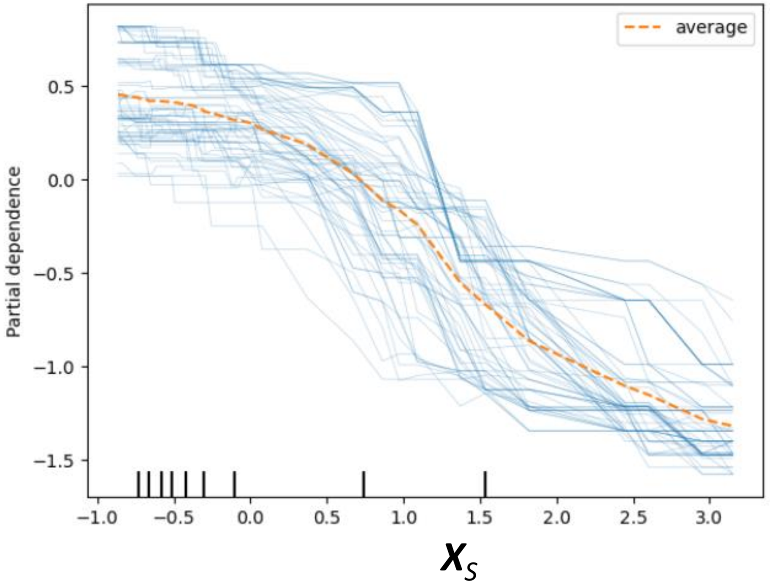


n lines =
*n observations in **X** data*

Local methods for interpretation

- **Individual conditional expectation (ICE) plot**
- Shows the *individual* effect one or two features have on the predicted outcome of a model
- Feature(s) of interest are $\boldsymbol{X_S}$, rest of features in model are $\boldsymbol{X_C}$

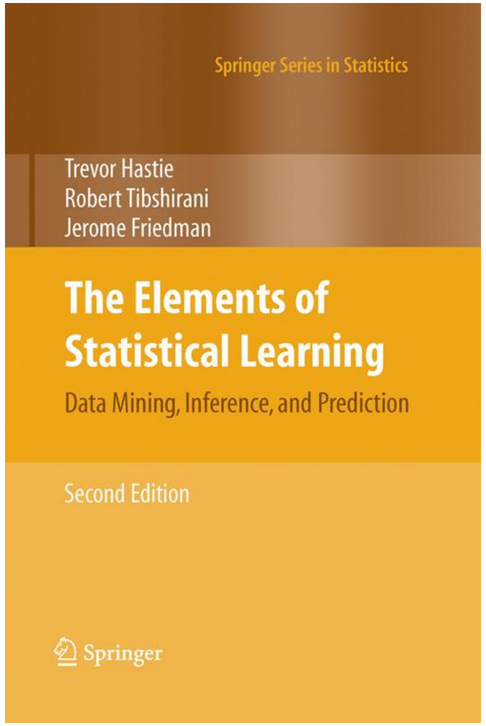
Original data $\boldsymbol{X_S}, \boldsymbol{X_C}, \boldsymbol{y}$	Original data for $\boldsymbol{X_C}, \boldsymbol{y}$; with $\boldsymbol{X_S} = b_1$	Original data for $\boldsymbol{X_C}, \boldsymbol{y}$; with $\boldsymbol{X_S} = b_2$	Original data for $\boldsymbol{X_C}, \boldsymbol{y}$; with $\boldsymbol{X_S} = b_3$	Original data for $\boldsymbol{X_C}, \boldsymbol{y}$; with $\boldsymbol{X_S} = b_n$

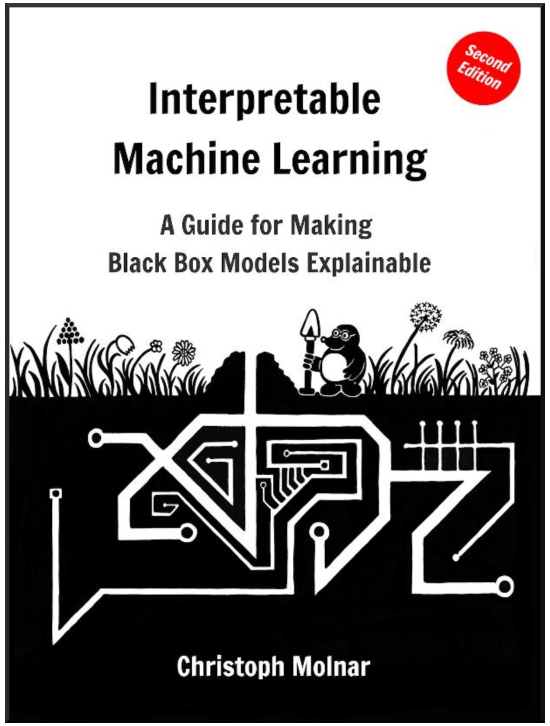


n lines =
*n observations in **X** data*

References

- [Hastie et al. (2009) The Elements of Statistical Learning]
- [Molnar (2023) Interpretable Machine Learning: A Guide for Making Black Box Models Explainable.]





StoneThree

The Future of Work. Now.™

+27 21 851 3123

info@stonethree.com

24 Gardner Williams Avenue

Paardevlei Somerset West

South Africa 7130

www.stonethree.com

ISO/IEC 27001 Information Security Management CERTIFIED