MACHINE LEARNING FOR HEALTH

Classical Machine Learning: From Labels to Discovery

Regression, trees, ensembles, SVMs, clustering, and computational phenotyping

TECHNICAL HANDOUT · SESSION 06

Professor Flávio Luiz Seixas
Graduate Program in Computing

Academic teaching material based exclusively on the PDFs available in the project

Table of Contents

Introduction

Classical algorithms offer different inductive biases for representing relationships between inputs and outputs or discovering structure without labels. Choosing among them should not be an abstract performance contest: it depends on the task, dimensionality, sample size, stability, and need for interpretation. [1]

This session compares logistic regression, trees, ensembles, support vector machines, clustering, and dimensionality reduction, highlighting inputs, outputs, assumptions, and interpretive pitfalls. [1]

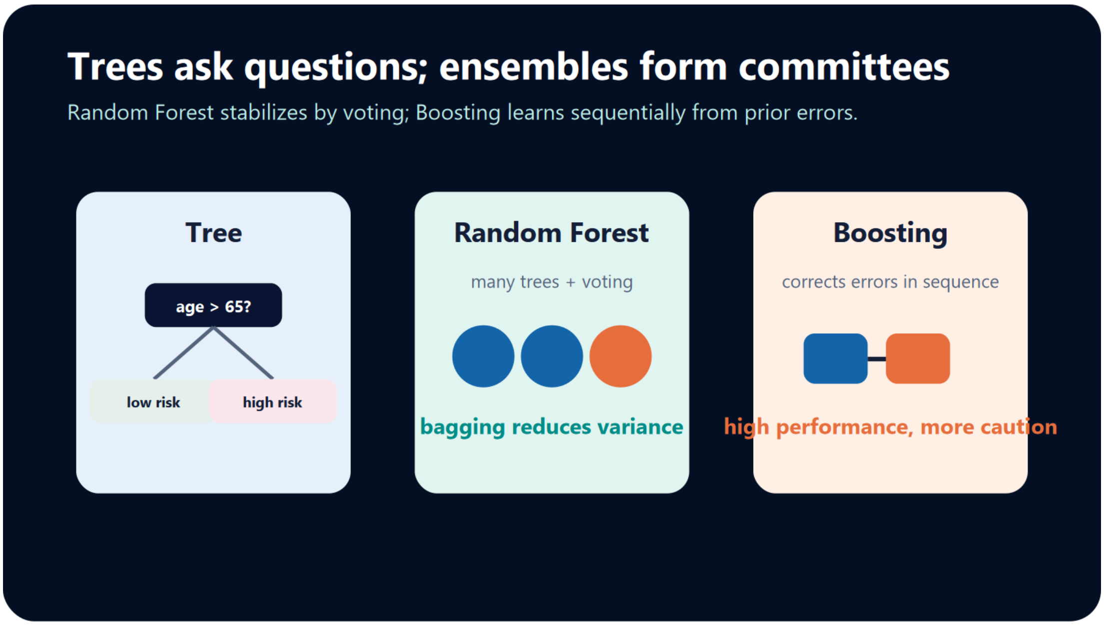

Figure 1 — Trees and ensembles. Source: original illustration based on the course material and [1] and [2].

1. Supervised and unsupervised learning

In supervised learning, examples include predictors and a known output, and the model learns a function for new cases. In unsupervised learning, there is no target label and the algorithm seeks organization in the data. This distinction changes the type of claim that can be made: a cluster suggests structure, but it does not demonstrate that the groups are naturally occurring diseases [1].

KEY CONCEPT
In supervised learning, examples include predictors and a known output, and the model learns a function for new cases.

2. Logistic regression and margins

Logistic regression models the relationship between predictors and the probability of a binary class. Coefficients can be transformed into odds ratios, but this association should not automatically be interpreted as a causal effect. Regularization can control complexity and instability in datasets with many variables [1].

Support vector machines seek a boundary with a margin between classes; kernel functions allow nonlinear boundaries in a transformed space. Variable scaling, kernel choice, and penalty parameters influence the result and its generalization [2].

3. Trees and ensembles

Trees divide the predictor space through successive rules and produce a class or value at the leaves. Deep trees may fit noise. Random forests combine trees constructed with resampling and subsets of variables, reducing variance through aggregation [2].

Boosting builds models sequentially, with later stages focusing on previous errors. The resulting flexibility increases the need for regularization, validation, and explanations appropriate to clinical use [1].

4. Clustering, visualization, and phenotypes

Clustering depends on the representation, distance, and criterion used to form groups. Changes in scale or variable selection can change the solution. Reductions such as t-SNE and UMAP are useful for visualizing neighborhoods, but two-dimensional geometry should not be taken as proof of clinical separation [1].

Computational phenotyping uses patterns in clinical data to characterize groups or states. Interpretation requires external validation, stability, plausibility, and a relationship to decisions or outcomes; naming a cluster does not substitute for evidence of validity [3].

Session summary

Supervised and unsupervised learning: in supervised learning, examples include predictors and a known output, and the model learns a function for new cases.
Logistic regression and margins: logistic regression models the relationship between predictors and the probability of a binary class.
Trees and ensembles: trees divide the predictor space through successive rules and produce a class or value at the leaves.
Clustering, visualization, and phenotypes: clustering depends on the representation, distance, and criterion used to form groups.

References

[1] Subramanian, D.; Cohen, T. A. Machine Learning Systems. In: Intelligent Systems in Medicine and Health. Springer, 2022. External link

Local file: pdfs/3-Intelligent Systems/VI-Machine Learning Systems.pdf · PDF pages used: 17, 25, 31, 37, 44, 70, 71.

[2] Agrawal, R.; Chatterjee, J. M.; Jena, A. K. (eds.). Machine Learning for Healthcare: Handling and Managing Data. CRC Press, 2020.

Local file: pdfs/1-Machine Learning/Machine Learning for Healthcare _ Handling and Managing Data -- edited by Rashmi Agrawal, Jyotir Moy Chatterjee, Abhishek -- 1, PT, 2020 -- CRC Press, -- isbn13 9780367352332 -- 6284b1970cc3abcc85ea2ecea9ac7116 --.pdf · PDF pages used: 26, 27, 28, 30.

[3] Bellazzi, R.; Dagliati, A.; Nicora, G. Predicting Medical Outcomes. In: Intelligent Systems in Medicine and Health. Springer, 2022. External link

Local file: pdfs/3-Intelligent Systems/XI-Predicting Medical Outcomes.pdf · PDF pages used: 9, 17, 24, 25.