PS5841

Data Science in Finance & Insurance

Decision Tree

Yubo Wang

Autumn 2022

COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

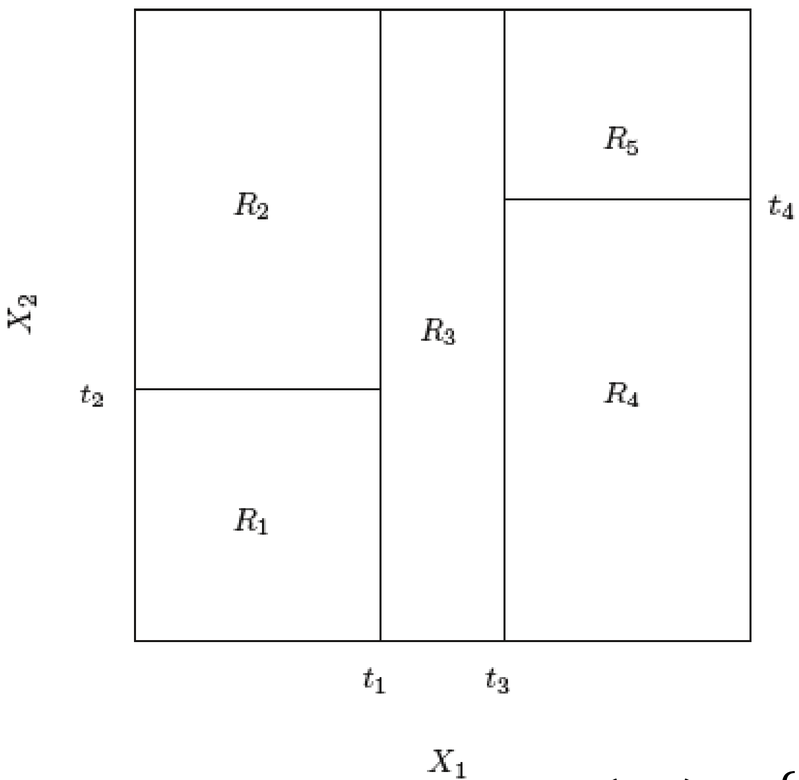
Decision Trees

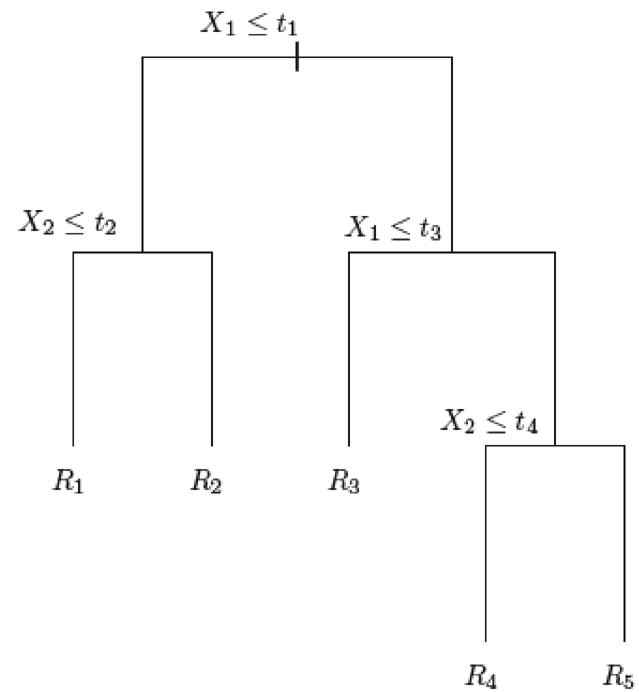
- Prediction via stratification of the feature space
  - Divide the predictor space into high-dimensional rectangles that minimizes “loss” via recursive binary splitting
  - Prediction based on the mean (for regression) or the most commonly occurring class (for classification) of training responses in the same terminal node

Recursive Binary Splitting

- Top-Down
  - Start from the top of the tree
- Greedy
  - The best split for a particular node is made at that particular step only, rather than taking into account of future steps
- Each split involves a cut-point s which splits a predictor X_j into two partitions

Example





$$R_-(j,s) = \{X \mid X_j \le s\}, R_+(j,s) = \{X \mid X_j > s\}$$

Find the values of j (feature) and s (cut point) that minimize "loss"

$$\sum_{i:\boldsymbol{x}_i \in R_-(j,s)} loss(y_i, \hat{y}_{R_-}) + \sum_{i:\boldsymbol{x}_i \in R_+(j,s)} loss(y_i, \hat{y}_{R_+})$$

Split Criteria ("loss")

- Regression
  - RSS

$$\sum_{j=1}^{J}\sum_{i\in R_j}\left(y_i - \hat{y}_{R_j}\right)^2$$

- Classification
  - Gini index

$$G_{R_m} = \sum_{k=1}^{K} \hat{p}_{R_m C_k}\left(1 - \hat{p}_{R_m C_k}\right)$$

  - Entropy

$$D_{R_m} = -\sum_{k=1}^{K} \hat{p}_{R_m C_k} \log \hat{p}_{R_m C_k}$$

$\hat{p}_{R_m C_k}$ for Classification

- The proportion of training observations in the m-th region R_m that are from the k-th class C_k

$$\hat{p}_{mk} = \hat{p}_{R_m C_k} = \frac{n_{R_m C_k}}{n_{R_m}}$$

Gini Index

- For the m-th region R_m

$$G_{R_m} = \sum_{k=1}^{K} \hat{p}_{R_m C_k}\left(1 - \hat{p}_{R_m C_k}\right)$$

  - a measure of variance across the K classes for observations in that region
  - G_{R_m} will take on a small value if the m-th node is pure, containing predominantly observations from a single class
- Overall Gini Index

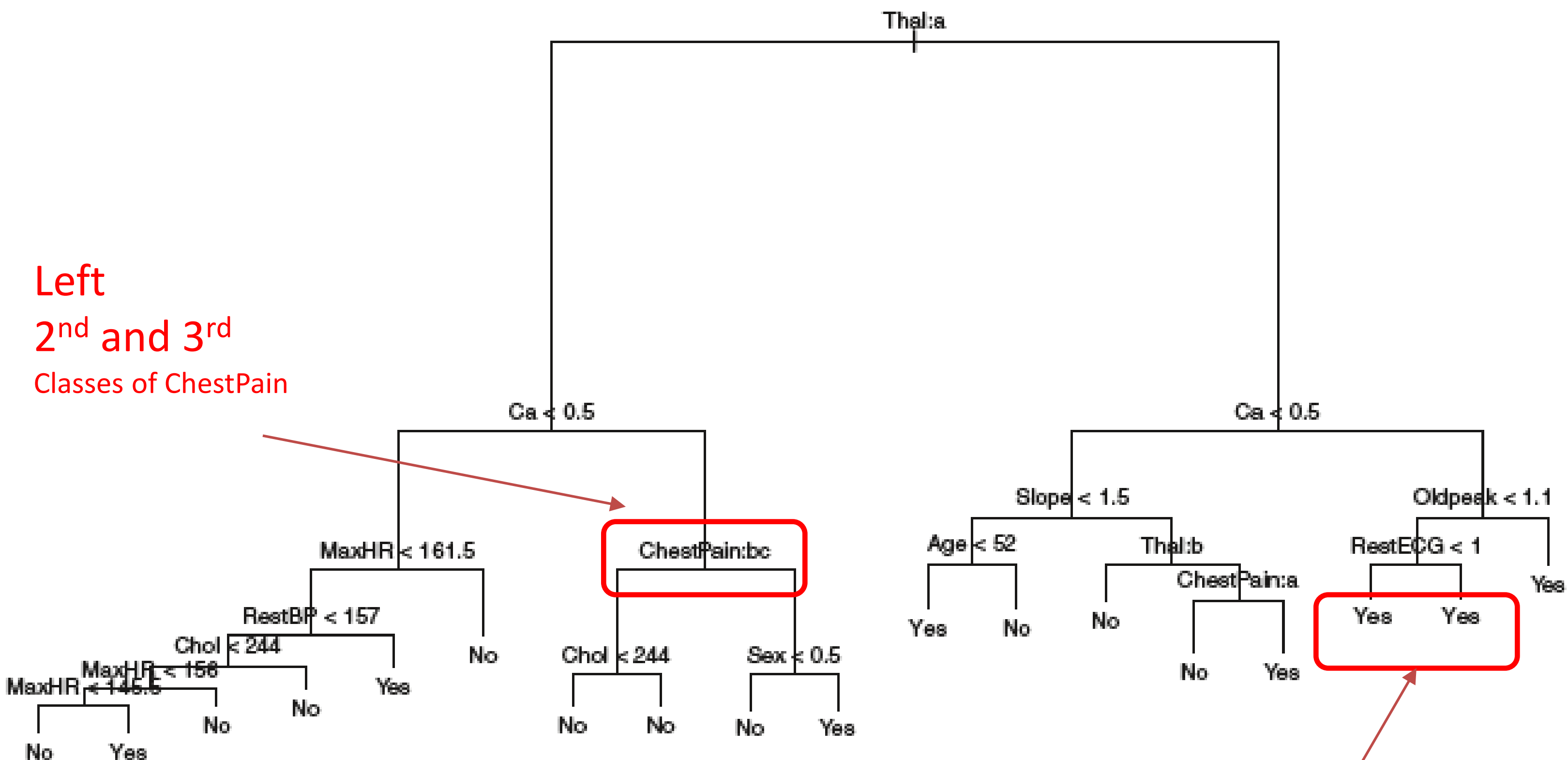
$$G = \sum_{m=1}^{M} \frac{n_{R_m}}{N} G_{R_m}$$

  - pooled variance involving regional variances

Example: Binary Split on Gini Index

- 2-class responses and 2-D features X_1 and X_2
- Find the optimal split for predictor X_1
  - Find $s_{X_1}^*$ that minimizes G as $G^{X_1}(s_{X_1}^*)$
- Find the optimal split for predictor X_2
  - Find $s_{X_2}^*$ that minimizes G as $G^{X_2}(s_{X_2}^*)$
- If $G^{X_1}\left(s_{X_1}^*\right) < G^{X_2}(s_{X_2}^*)$, the current step splits X_1, otherwise, the current step splits X_2

Split, Node Purity



Left
2^{nd} and 3^{rd}
Classes of ChestPain

Split for node purity
Left $\hat{p}_{mk} = 0.64$
Right $\hat{p}_{mk} = 1.00$

- Node purity – the degree to which a node contains predominantly observations from a single class

Entropy

- For the m-th region R_m

$$D_{R_m} = -\sum_{k=1}^{K} \hat{p}_{R_m c_k} \log \hat{p}_{R_m c_k}$$

  – D_{R_m} will take on a small value if the m-th node is pure, containing predominantly observations from a single class

- Overall entropy
  – is the sum of entropy over all regions
- The Gini index and the entropy are quite similar numerically

That was