Introduction to Machine Learning

CART: Computational Aspects of Finding Splits

compstat-lmu.github.io/lecture_i2ml

MONOTONE FEATURE TRANSFORMATIONS

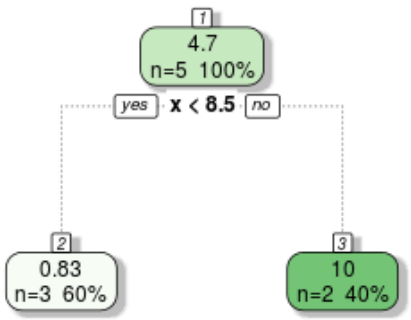
Monotone transformations of one or several features will neither change the value of the splitting criterion nor the structure of the tree, only the numerical value of the split point.

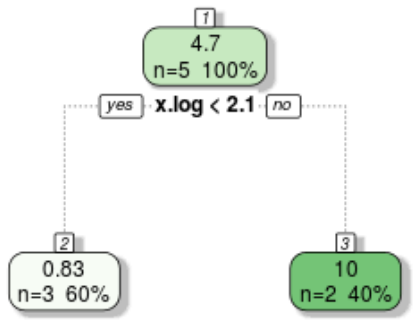
Original data

x	1	2	7.0	10	20
y	1	1	0.5	10	11

Data with log-transformed x

log(x)	0	0.7	1.9	2.3	3
y	1	1.0	0.5	10.0	11





CART: NOMINAL FEATURES

- A split on a nominal feature partitions the feature levels:

$$x_j \in \{a, c, e\} \leftarrow \mathcal{N} \rightarrow x_j \in \{b, d\}$$

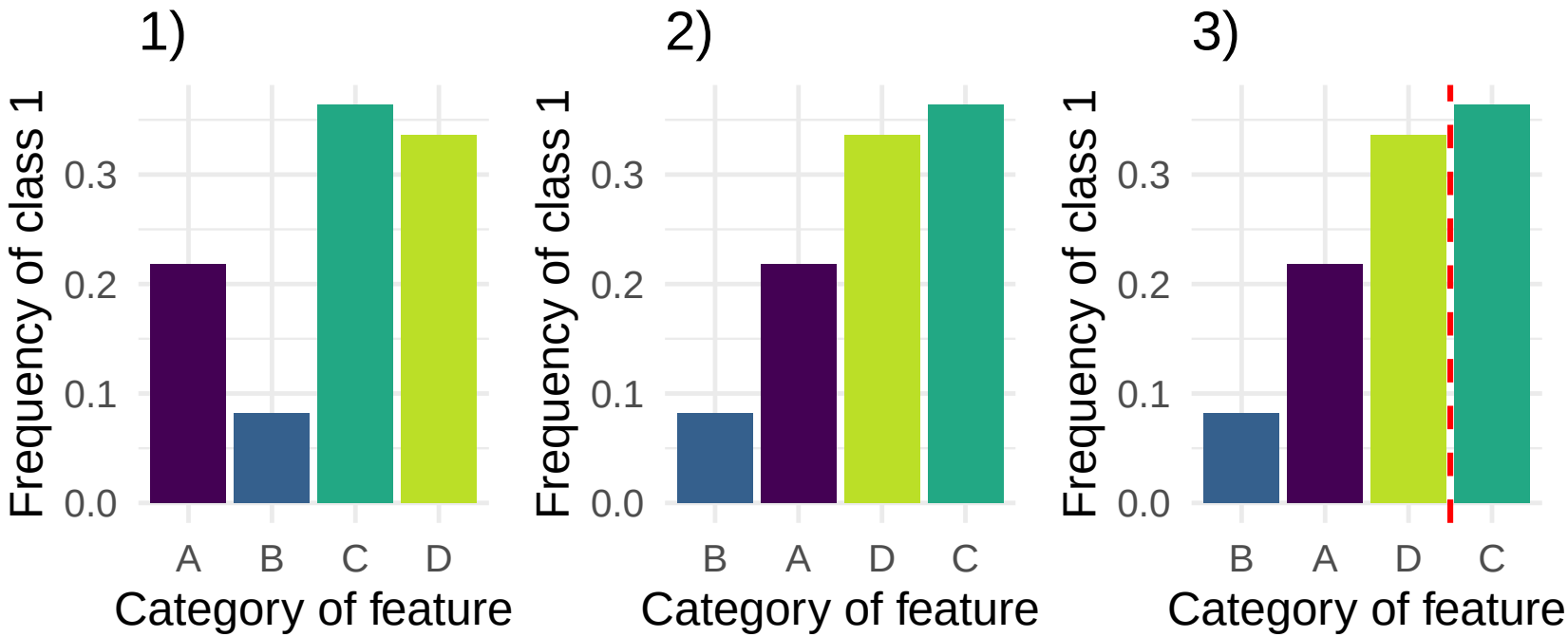
- For a feature with m levels, there are about 2^m different possible partitions of the m values into two groups ($2^{m-1} - 1$ because of symmetry and empty groups).
- Searching over all these becomes prohibitive for larger values of m.
- For regression with squared loss and binary classification, we can define clever shortcuts.

CART: NOMINAL FEATURES

For $0 - 1$ responses, in each node:

1. Calculate the proportion of 1-outcomes for each category of the feature in $\mathcal{N}$.
2. Sort the categories according to these proportions.
3. The feature can then be treated as if it was ordinal, so we only have to investigate at most $m - 1$ splits.

CART: NOMINAL FEATURES



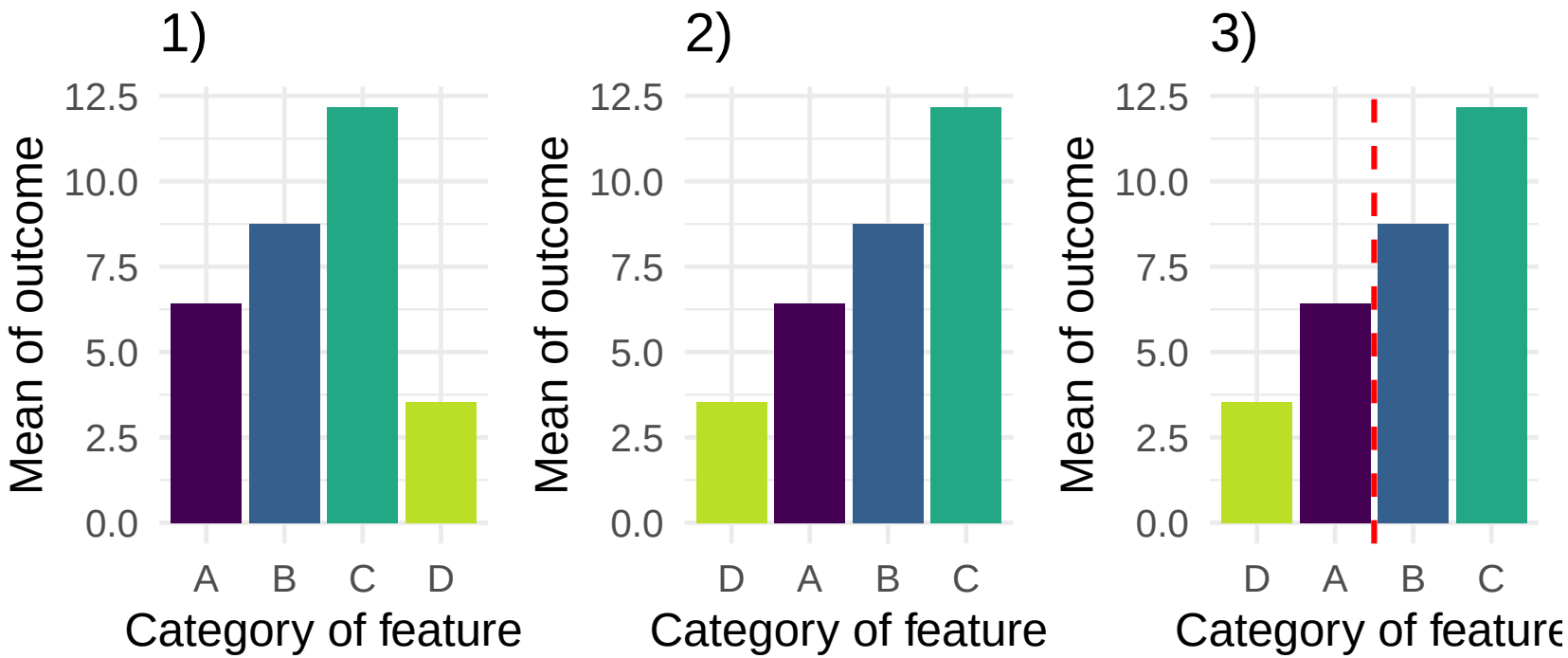
CART: NOMINAL FEATURES

- This procedure finds the optimal split.
- This result also holds for regression trees (with squared error loss) if the levels of the feature are ordered by increasing mean of the target
- The proofs are not trivial and can be found here:
  - for 0-1 responses:
    - Breiman, 1984, Classification and Regression Trees.
    - Ripley, 1996, Pattern Recognition and Neural Networks.
  - for continuous responses:
    - Fisher, 1958, On grouping for maximum homogeneity.
- Such simplifications are not known for multiclass problems.

CART: NOMINAL FEATURES

For continuous responses, in each node:

1. Calculate the mean of the outcome in each category
2. Sort the categories by increasing mean of the outcome



CART: MISSING FEATURE VALUES

- When splits are evaluated, only observations for which the used feature is not missing are used. (This can actually bias splits towards using features with lots of missing values.)
- CART often uses the so-called **surrogate split** principle to automatically deal with missing values in features used for splits during prediction.
- Surrogate splits are created during training. They define replacement splitting rules using a different feature that result in almost the same child nodes as the original split.
- When observations are passed down the tree (in training or prediction), and the feature value used in a split is missing, we use a "surrogate split" instead to decide to which branch of the tree the data should be assigned.