Supervised, Unsupervised and Reinforcement Learning in Finance

Week 1: Supervised Learning

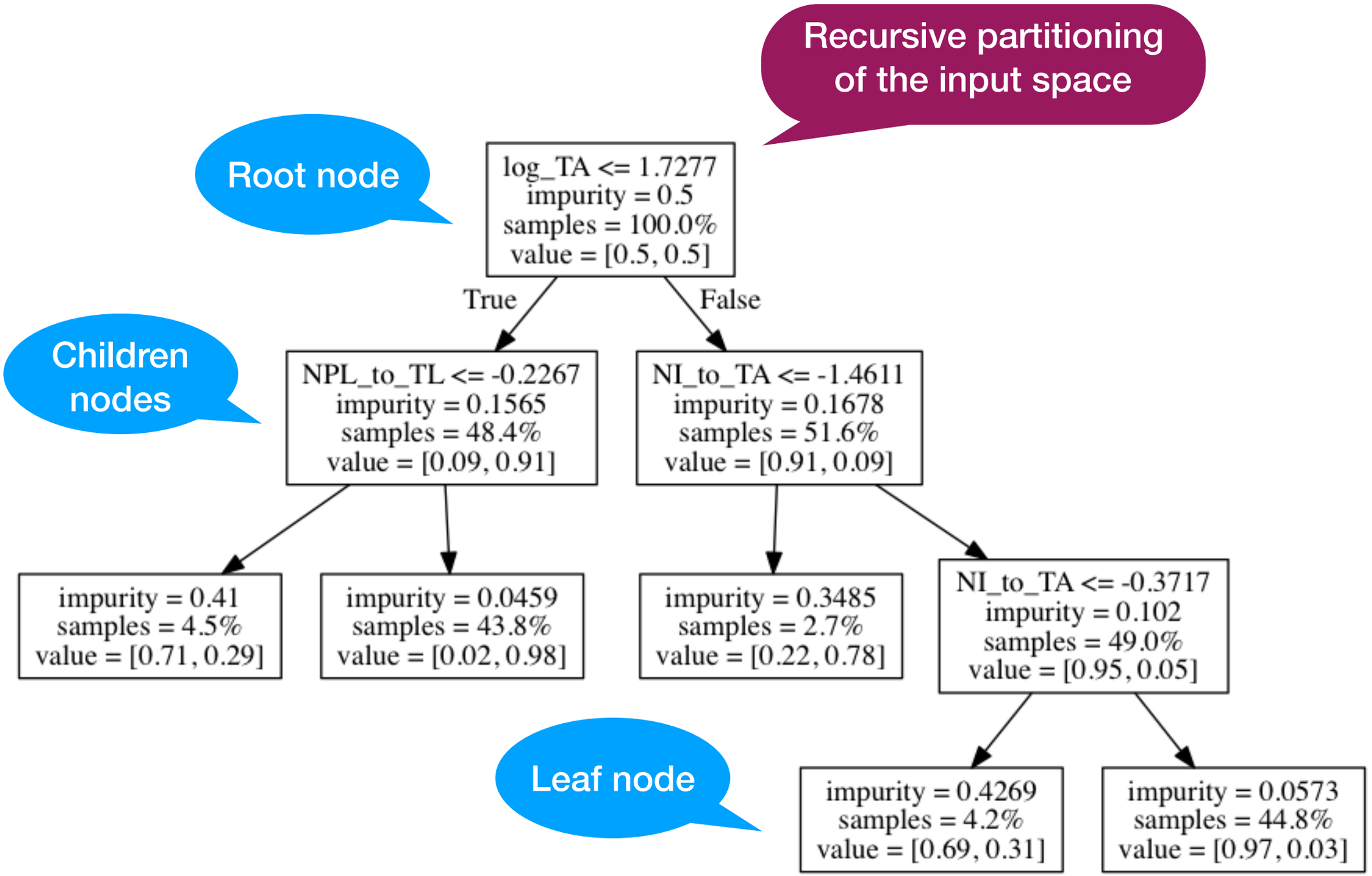
Tree methods: CART Trees

Igor Halperin

NYU Tandon School of Engineering, 2017

CART for bank analysis

CART = Classification and Regression Tree

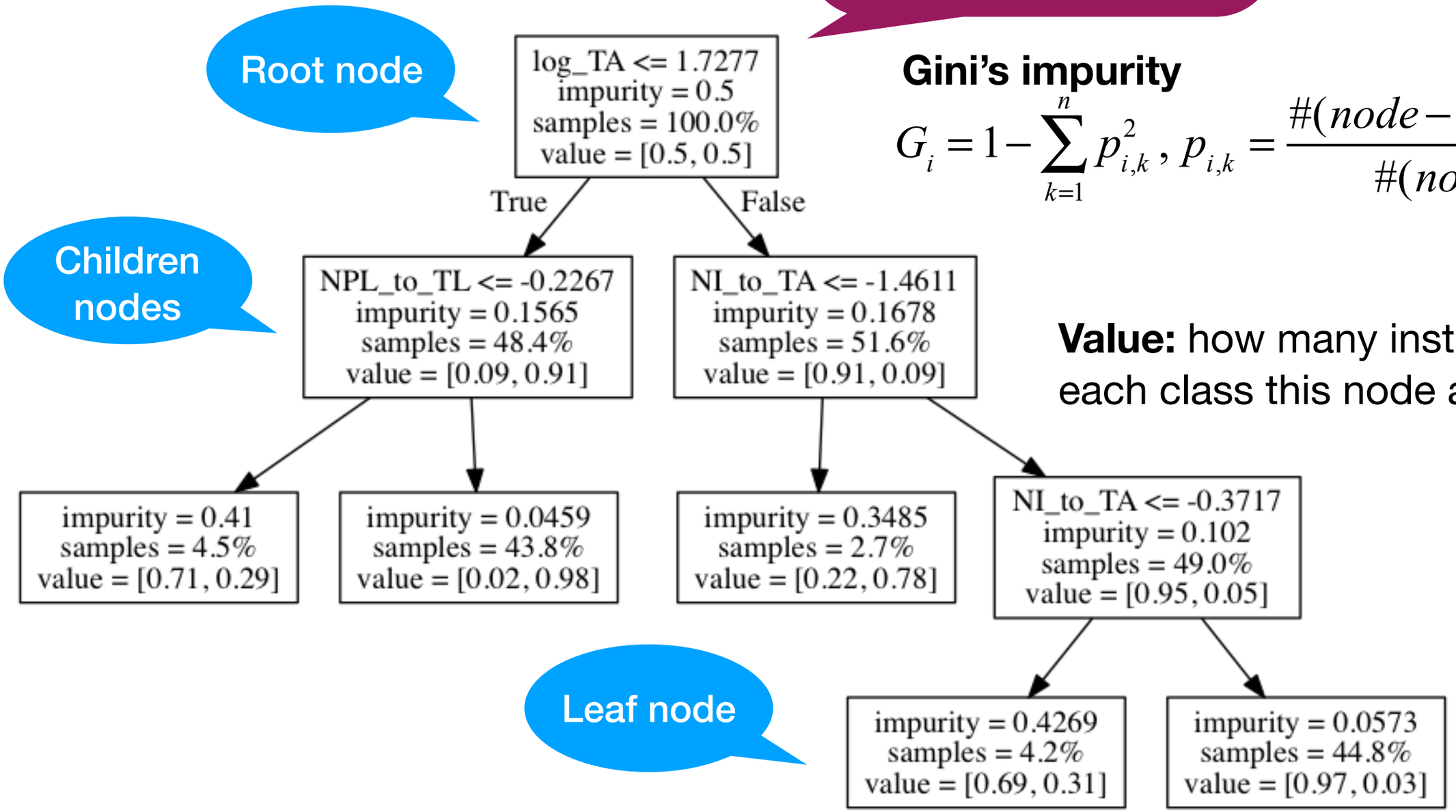


CART for bank analysis

CART = Classification and Regression Tree

Recursive partitioning of the input space

Root node



Children nodes

Leaf node

Gini's impurity

$$G_i = 1 - \sum_{k=1}^{n} p_{i,k}^2 \,,\; p_{i,k} = \frac{\#(node - i, class - k)}{\#(node - i)}$$

Value: how many instances of each class this node applies to

The CART training algorithm

Greedy algorithm to grow a tree starting from a root note:
The cost function

$$J(k,t_k) = \frac{m_{left}}{m} G_{left} + \frac{m_{right}}{m} G_{right}$$

$$where\ G_{left,right} = impurity\ of\ the\ left\ /\ right\ subset$$

At each step, choose feature k and threshold t_k to minimize the cost function

Can use other measures instead of Gini impurities $G_{left,right}$:

Entropy: $H_i = -\sum_{k=1}^{n} p_{i,k} \log p_{i,k}$ (Entropy is zero if there is only one class in a node, so behaves similar to Gini)

Entropy and Gini often, but not always, produce similar trees.

Complexity control for trees

Trees can easily overfit, need regularization:

- Constrain the maximum depth of the tree (max_depth)
- Min_samples_split (minimal number of samples in a node to be considered for a split)
- Min_samples_leaf (min number of samples to be in each leaf node)
- Min_impurity_split (a node will split if its impurity is above the threshold, otherwise it is a leaf)

2. These hyper-parameters can be tuned using a validation set, or by cross-validation.

Trees: pros and cons

Pros:

- Simple to interpret
- Require almost no pre-processing
- Can handle missing data
- Insensitive to monotone transformations of inputs
- Perform automatic variable selection
- Scale up well to large datasets

Cons:

- Lower accuracy than other model (due to the greedy tree construction)
- Potential instability under small variation of input data (the same origin) (Trees are high variance estimators)

Control question

Select all correct answers

1. Gini impurity is defined as $G_i = 1 - \sum^{n} p_{i,k}^2$
   where $p_{i,k}$ is a fraction of instances of class k among all instances at node i.

2.Gini impurity is defined as $G_i = -\sum^{n} p_{i,k} \log p_{i,k}$

3. Hyper-parameters of a CART tree are optimized by minimizing the CART cost function on the training set.

4. Trees are simple to interpret, and they perform an automatic feature selection

5. Trees are typically high variance estimations

Correct answers: 1, 4, 5.