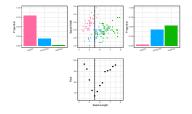
## **Introduction to Machine Learning**

# CART Splitting Criteria for Classification





#### Learning goals

- Understand how to define split criteria via ERM
- Understand how to find splits in regression with L<sub>2</sub> loss

#### **OPTIMAL CONSTANT MODELS**

As losses in classification, we typically use:

- (Multi-class) Brier score  $L(y, \pi) = \sum_{k=1}^{g} (\pi_k o_k(y))^2$ , a.k.a.  $L_2$  loss on probabilities
- (Multi-class) Log loss  $L(y, \pi) = -\sum_{k=1}^{g} o_k(y) \log(\pi_k)$ , as in logistic regression

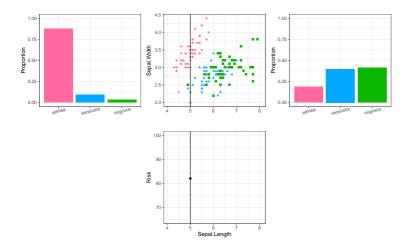
Optimal constant predictions (in a node) for both losses are simply the proportions of the contained classes:

$$c_{\mathcal{N}} = (\hat{\pi}_1^{(\mathcal{N})}, \dots, \hat{\pi}_g^{(\mathcal{N})})$$
 with  $\hat{\pi}_k^{(\mathcal{N})} = \frac{1}{|\mathcal{N}|} \sum_{(\mathbf{x}, y) \in \mathcal{N}} \mathbb{I}(y = k) \quad \forall k \in \{1, \dots, g\}$ 



#### FINDING THE BEST SPLIT

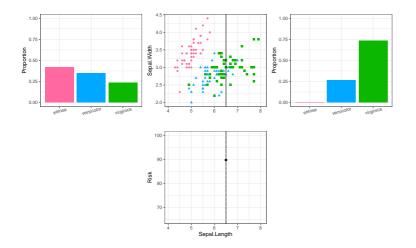
Let's compute the Brier score for all splits, with optimal constant probability vectors in both children





#### FINDING THE BEST SPLIT

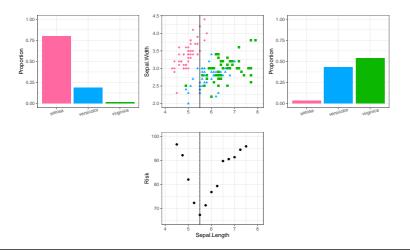
Let's compute the Brier score for all splits, with optimal constant probability vectors in both children





#### FINDING THE BEST SPLIT

The optimal split point typically creates greatest imbalance or purity of label distribution





#### **RISK MINIMIZATION VS. IMPURITY**

- Split crits are sometimes defined in terms of impurity reduction instead of ERM, where a measure of "impurity" is defined per node
- For regression trees, "impurity" is simply defined as variance of y, which is quite obviously L<sub>2</sub> loss
- Brier score is equivalent to Gini impurity

$$I(\mathcal{N}) = \sum_{k=1}^{g} \hat{\pi}_{k}^{(\mathcal{N})} \left(1 - \hat{\pi}_{k}^{(\mathcal{N})}\right)$$

Log loss is equivalent to entropy

$$I(\mathcal{N}) = -\sum_{k=1}^{g} \hat{\pi}_{k}^{(\mathcal{N})} \log \hat{\pi}_{k}^{(\mathcal{N})}$$

 Trees can be understood completely through the lens of ERM, so this new terminology is unnecessary and perhaps confusing



### **SPLITTING WITH MISCLASSIFICATION LOSS**

- Often, we want to minimize the MCE in classification
- Zero-One-Loss is not differentiable, but that is a non-issue in the tree-optimization based on loops
- Brier score and Log loss more sensitive to changes in the node probs, often produce purer nodes, and are still preferred



S	pl	it	1	:

Split 2:

	class 0	class 1
$\mathcal{N}_1$	300	100
$\mathcal{N}_{2}$	100	300

	class 0	class 1
$\mathcal{N}_1$	400	200
$\mathcal{N}_{2}$	0	200

- Both splits are equivalent in MCE
- But: Split 2 results in purer nodes, both Brier score (Gini) and Log loss (Entropy) prefer 2nd split