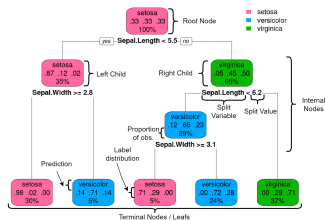
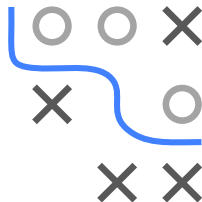


Introduction to Machine Learning

Advanced Risk Minimization Loss functions and tree splitting



Learning goals

- Know how tree splitting is 'nothing new' and related to loss functions
- Brier score minimization corresponds to gini splitting
- Bernoulli loss minimization corresponds to entropy splitting

BERNOULLI LOSS MIN = ENTROPY SPLITTING

For an introduction on trees and splitting criteria we refer our **I2ML** lecture (Chapter 6, [Bischi et al. 2022](#))

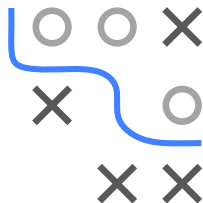
When fitting a tree we minimize the risk within each node \mathcal{N} by risk minimization and predict the optimal constant. Another common approach is to minimize the average node impurity $\text{Imp}(\mathcal{N})$.

Claim: Entropy splitting $\text{Imp}(\mathcal{N}) = -\sum_{k=1}^g \pi_k^{(\mathcal{N})} \log \pi_k^{(\mathcal{N})}$ is equivalent to minimize risk measured by the Bernoulli loss.

Note that $\pi_k^{(\mathcal{N})} := \frac{1}{n_{\mathcal{N}}} \sum_{(\mathbf{x}, y) \in \mathcal{N}} [y = k]$.

Proof: To prove this we show that the risk related to a subset of observations $\mathcal{N} \subseteq \mathcal{D}$ fulfills $\mathcal{R}(\mathcal{N}) = n_{\mathcal{N}} \text{Imp}(\mathcal{N})$, where $\mathcal{R}(\mathcal{N})$ is calculated w.r.t. the (multiclass) Bernoulli loss

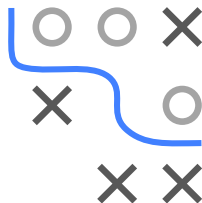
$$L(y, \pi(\mathbf{x})) = -\sum_{k=1}^g [y = k] \log(\pi_k(\mathbf{x})).$$



BERNOULLI LOSS MIN = ENTROPY SPLITTING / 2

$$\begin{aligned}\mathcal{R}(\mathcal{N}) &= \sum_{(\mathbf{x}, y) \in \mathcal{N}} \left(- \sum_{k=1}^g [y = k] \log \pi_k(\mathbf{x}) \right) \\ &\stackrel{(*)}{=} - \sum_{k=1}^g \sum_{(\mathbf{x}, y) \in \mathcal{N}} [y = k] \log \pi_k^{(\mathcal{N})} \\ &= - \sum_{k=1}^g \log \pi_k^{(\mathcal{N})} \underbrace{\sum_{(\mathbf{x}, y) \in \mathcal{N}} [y = k]}_{n_{\mathcal{N}} \cdot \pi_k^{(\mathcal{N})}} \\ &= -n_{\mathcal{N}} \sum_{k=1}^g \pi_k^{(\mathcal{N})} \log \pi_k^{(\mathcal{N})} = n_{\mathcal{N}} \text{Imp}(\mathcal{N}),\end{aligned}$$

where in $(*)$ the optimal constant per node $\pi_k^{(\mathcal{N})} = \frac{1}{n_{\mathcal{N}}} \sum_{(\mathbf{x}, y) \in \mathcal{N}} [y = k]$ was plugged in.



BRIER SCORE MINIMIZATION = GINI SPLITTING

When fitting a tree we minimize the risk within each node \mathcal{N} by risk minimization and predict the optimal constant. Another approach that is common in literature is to minimize the average node impurity $\text{Imp}(\mathcal{N})$.

Claim: Gini splitting $\text{Imp}(\mathcal{N}) = \sum_{k=1}^g \pi_k^{(\mathcal{N})} (1 - \pi_k^{(\mathcal{N})})$ is equivalent to the Brier score minimization.

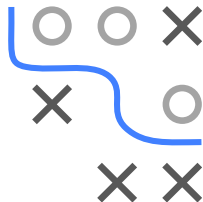
Note that $\pi_k^{(\mathcal{N})} := \frac{1}{n_{\mathcal{N}}} \sum_{(\mathbf{x}, y) \in \mathcal{N}} [y = k]$

Proof: We show that the risk related to a subset of observations $\mathcal{N} \subseteq \mathcal{D}$ fulfills

$$\mathcal{R}(\mathcal{N}) = n_{\mathcal{N}} \text{Imp}(\mathcal{N}),$$

where Imp is the Gini impurity and $\mathcal{R}(\mathcal{N})$ is calculated w.r.t. the (multiclass) Brier score

$$L(y, \pi(\mathbf{x})) = \sum_{k=1}^g ([y = k] - \pi_k(\mathbf{x}))^2.$$



BRIER SCORE MINIMIZATION = GINI SPLITTING / 2

$$\mathcal{R}(\mathcal{N}) = \sum_{(\mathbf{x}, y) \in \mathcal{N}} \sum_{k=1}^g ([y = k] - \pi_k(\mathbf{x}))^2 = \sum_{k=1}^g \sum_{(\mathbf{x}, y) \in \mathcal{N}} \left([y = k] - \frac{n_{\mathcal{N}, k}}{n_{\mathcal{N}}} \right)^2,$$

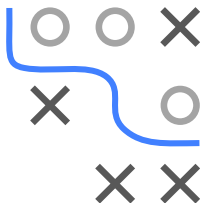
by plugging in the optimal constant prediction w.r.t. the Brier score ($n_{\mathcal{N}, k}$ is defined as the number of class k observations in node \mathcal{N}):

$$\hat{\pi}_k(\mathbf{x}) = \pi_k^{(\mathcal{N})} = \frac{1}{n_{\mathcal{N}}} \sum_{(\mathbf{x}, y) \in \mathcal{N}} [y = k] = \frac{n_{\mathcal{N}, k}}{n_{\mathcal{N}}}.$$

We split the inner sum and further simplify the expression

$$\begin{aligned} &= \sum_{k=1}^g \left(\sum_{(\mathbf{x}, y) \in \mathcal{N}: y=k} \left(1 - \frac{n_{\mathcal{N}, k}}{n_{\mathcal{N}}} \right)^2 + \sum_{(\mathbf{x}, y) \in \mathcal{N}: y \neq k} \left(0 - \frac{n_{\mathcal{N}, k}}{n_{\mathcal{N}}} \right)^2 \right) \\ &= \sum_{k=1}^g n_{\mathcal{N}, k} \left(1 - \frac{n_{\mathcal{N}, k}}{n_{\mathcal{N}}} \right)^2 + (n_{\mathcal{N}} - n_{\mathcal{N}, k}) \left(\frac{n_{\mathcal{N}, k}}{n_{\mathcal{N}}} \right)^2, \end{aligned}$$

since for $n_{\mathcal{N}, k}$ observations the condition $y = k$ is met, and for the remaining $(n_{\mathcal{N}} - n_{\mathcal{N}, k})$ observations it is not.



BRIER SCORE MINIMIZATION = GINI SPLITTING / 3

We further simplify the expression to

$$\begin{aligned}\mathcal{R}(\mathcal{N}) &= \sum_{k=1}^g n_{\mathcal{N},k} \left(\frac{n_{\mathcal{N}} - n_{\mathcal{N},k}}{n_{\mathcal{N}}} \right)^2 + (n_{\mathcal{N}} - n_{\mathcal{N},k}) \left(\frac{n_{\mathcal{N},k}}{n_{\mathcal{N}}} \right)^2 \\ &= \sum_{k=1}^g \frac{n_{\mathcal{N},k}}{n_{\mathcal{N}}} \frac{n_{\mathcal{N}} - n_{\mathcal{N},k}}{n_{\mathcal{N}}} (n_{\mathcal{N}} - n_{\mathcal{N},k} + n_{\mathcal{N},k}) \\ &= n_{\mathcal{N}} \sum_{k=1}^g \pi_k^{(\mathcal{N})} \cdot (1 - \pi_k^{(\mathcal{N})}) = n_{\mathcal{N}} \text{Imp}(\mathcal{N}).\end{aligned}$$

