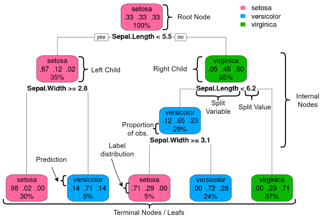
Introduction to Machine Learning

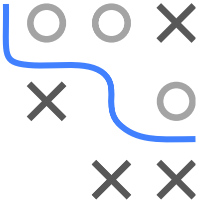
Advanced Risk Minimization
Loss functions and tree splitting (Deep-Dive)



Learning goals

- Tree splitting loss vs impurity:
- Bernoulli loss $\sim$ entropy splitting
- Brier score $\sim$ gini splitting

RISK MINIMIZATION AND IMPURITY

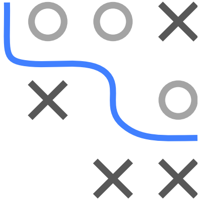


- Tree fitting: Find best way to split parent node $\mathcal{N}_0$ into child nodes $\mathcal{N}_1$ and $\mathcal{N}_2$ such that $\mathcal{N}_1 \cup \mathcal{N}_2 = \mathcal{N}_0$ and $\mathcal{N}_1 \cap \mathcal{N}_2 = \emptyset$
- Two options for evaluating how good a split is: Per node $\mathcal{N}$ compute the following:
  1. Compute impurity $\text{Imp}(\mathcal{N})$ directly from observations in $\mathcal{N}$
  2. Fit optimal constant using loss function, sum up losses for $\mathcal{N}$
- Two common impurity measures are entropy and Gini index where $\pi_k^{(\mathcal{N})}$ are predicted probs for class $k = 1, \ldots, g$ in node $\mathcal{N}$:

$$\text{Imp}^{\text{ent}}(\mathcal{N}) = -\textstyle\sum_{k=1}^{g} \pi_k^{(\mathcal{N})} \log \pi_k^{(\mathcal{N})}$$

$$\text{Imp}^{\text{Gini}}(\mathcal{N}) = \sum_{k=1}^{g} \pi_k^{(\mathcal{N})} \left(1 - \pi_k^{(\mathcal{N})}\right)$$

RISK MINIMIZATION AND IMPURITY

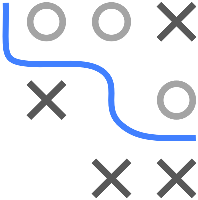


- In the following we will prove that entropy and Gini impurity measures are equivalent to splitting using log loss and Brier score:

$$L(y, \pi(\mathbf{x})) = -\sum_{k=1}^{g} \mathbb{I}[y = k] \log(\pi_k(\mathbf{x})) \quad \text{(log-loss)}$$

$$L(y, \pi(\mathbf{x})) = \sum_{k=1}^{g} (\mathbb{I}[y = k] - \pi_k(\mathbf{x}))^2 \quad \text{(Brier)}$$

BERNOULLI LOSS MIN = ENTROPY SPLITTING



Claim: Entropy as impurity

$$\text{Imp}^{\text{ent}}(\mathcal{N}) = -\sum_{k=1}^{g} \pi_k^{(\mathcal{N})} \log \pi_k^{(\mathcal{N})}$$

is equivalent to mean emp. risk with (multiclass) Bernoulli loss

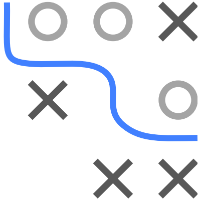
$$\bar{\mathcal{R}}_{\text{emp}}(\mathcal{N}) = \frac{1}{n_{\mathcal{N}}} \sum_{(\mathbf{x},y)\in\mathcal{N}} -\sum_{k=1}^{g} \mathbb{I}[y=k] \log(\pi_k(\mathbf{x}))$$

Proof: Let $\mathcal{N} \subseteq \mathcal{D}$ denote the subset of observations in that node and consider $\bar{\mathcal{R}}_{\text{emp}}(\mathcal{N})$ of node $\mathcal{N}$ with (multiclass) Bernoulli loss

$\Rightarrow$ Optimal constant per node $\pi_k^{(\mathcal{N})} = \frac{1}{n_{\mathcal{N}}} \sum\limits_{(\mathbf{x},y)\in\mathcal{N}} \mathbb{I}[y=k] = \frac{n_{\mathcal{N},k}}{n_{\mathcal{N}}}$

where $n_{\mathcal{N},k}$ is the number of class k observations in node $\mathcal{N}$

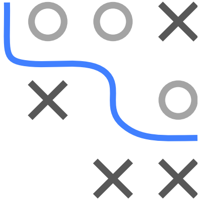
RISK MINIMIZATION AND IMPURITY



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

Avg. Bernoulli-risk of node $\mathcal{N}$ is equal to $\text{Imp}^{\text{ent}}(\mathcal{N})$

BRIER SCORE MINIMIZATION = GINI SPLITTING



Claim: Using Gini as impurity

$$\text{Imp}^{\text{Gini}}(\mathcal{N}) = \sum_{k=1}^{g} \pi_k^{(\mathcal{N})}(1 - \pi_k^{(\mathcal{N})})$$

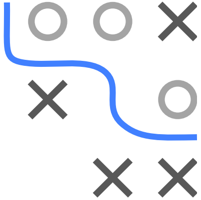
is equivalent to avg. emp. risk using Brier score

$$\bar{\mathcal{R}}_{\text{emp}}(\mathcal{N}) = \frac{1}{n_{\mathcal{N}}} \sum_{(\mathbf{x},y)\in\mathcal{N}} L\left(y, \pi(\mathbf{x})\right) = \frac{1}{n_{\mathcal{N}}} \sum_{(\mathbf{x},y)\in\mathcal{N}} \sum_{k=1}^{g} (\mathbb{I}[y = k] - \pi_k(\mathbf{x}))^2$$

Proof: Avg. empirical risk $\bar{\mathcal{R}}_{\text{emp}}(\mathcal{N})$ of node $\mathcal{N}$ using (multiclass) Brier score has optimal constant per node:

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

BRIER SCORE MINIMIZATION = GINI SPLITTING

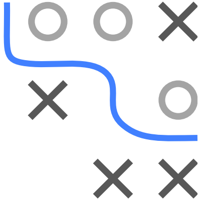


Inserting the optimal constant, the risk simplifies to

$$
\begin{aligned}
\bar{\mathcal{R}}_{\text{emp}}(\mathcal{N}) =& \frac{1}{n_{\mathcal{N}}} \sum_{(\mathbf{x},y)\in\mathcal{N}} \sum_{k=1}^{g} (\mathbb{I}[y = k] - \pi_k^{(\mathcal{N})})^2 = \frac{1}{n_{\mathcal{N}}} \sum_{k=1}^{g} \sum_{(\mathbf{x},y)\in\mathcal{N}} (\mathbb{I}[y = k] - \frac{n_{\mathcal{N},k}}{n_{\mathcal{N}}})^2 \\
=& \frac{1}{n_{\mathcal{N}}} \sum_{k=1}^{g} (\sum_{(\mathbf{x},y)\in\mathcal{N}:\ y=k} (1 - \frac{n_{\mathcal{N},k}}{n_{\mathcal{N}}})^2 + \sum_{(\mathbf{x},y)\in\mathcal{N}:\ y\neq k} (0 - \frac{n_{\mathcal{N},k}}{n_{\mathcal{N}}})^2) \\
=& \frac{1}{n_{\mathcal{N}}} \sum_{k=1}^{g} n_{\mathcal{N},k}(1 - \frac{n_{\mathcal{N},k}}{n_{\mathcal{N}}})^2 + (n_{\mathcal{N}} - n_{\mathcal{N},k})(\frac{n_{\mathcal{N},k}}{n_{\mathcal{N}}})^2
\end{aligned}
$$

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

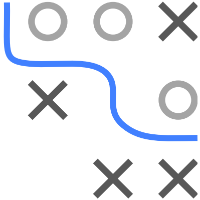


We further simplify the expression to

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

Avg. Brier-risk $\bar{\mathcal{R}}_{\text{emp}}(\mathcal{N})$ of the node is equal to its gini-impurity $\text{Imp}^{\text{Gini}}(\mathcal{N})$

WEIGHTING FOR RISK AND IMPURITY



- The emp. risk of a *split* is given by sum of per-node risks ($n_0 = n_1 + n_2$ are number of obs in nodes):

$$\begin{aligned}\mathcal{R}_{\text{emp}}(\text{split}) = \mathcal{R}_{\text{emp}}(\mathcal{N}_1) + \mathcal{R}_{\text{emp}}(\mathcal{N}_2) &= n_1\bar{\mathcal{R}}_{\text{emp}}(\mathcal{N}_1) + n_2\bar{\mathcal{R}}_{\text{emp}}(\mathcal{N}_2) \\ &= n_1\text{Imp}(\mathcal{N}_1) + n_2\text{Imp}(\mathcal{N}_2) \\ &= n_0\left(\frac{n_1}{n_0}\text{Imp}(\mathcal{N}_1) + \frac{n_2}{n_0}\text{Imp}(\mathcal{N}_2)\right)\end{aligned}$$

- As you can see above: if working with average risk, we need to reweight in the addition (as the averages are computed on subsets of potentially unequal sizes)
- Average risks are used in impurity formulas,we so simply have to adhere to that slight modification in split finding with them
- The impurity of a split is defined as a weighted average

$$\text{Imp}(\text{split}) = \frac{n_1}{n_0}\text{Imp}(\mathcal{N}_1) + \frac{n_2}{n_0}\text{Imp}(\mathcal{N}_2)$$