Lecture 7: Impurity Measures for Decision Trees

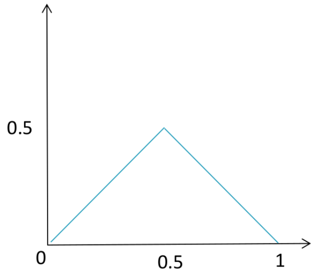
Madhavan Mukund

https://www.cmi.ac.in/~madhavan

Data Mining and Machine Learning
August–December 2020

Misclassification rate

- Goal: partition with uniform category — pure leaf
- Impure node — best prediction is majority value
- Minority ratio is misclassification rate
- Heuristic: reduce impurity as much as possible
- For each attribute, compute weighted average misclassification rate of children
- Choose the minimum



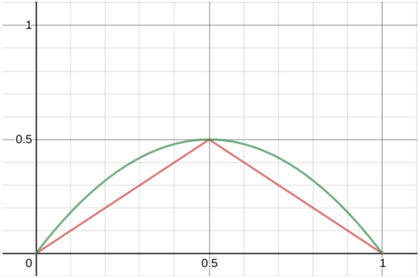
$c = 1$

Misclassification rate is linear

- $c \in \{0, 1\}$
- x-axis: fraction of inputs with $c = 1$

A better impurity function

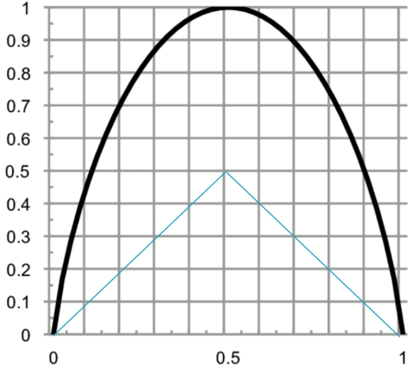
- Misclassification rate is linear
- Impurity measure that increases more sharply performs better, empirically
- Entropy — [Quinlan]
- Gini index — [Breiman]



$c = 1$

Entropy

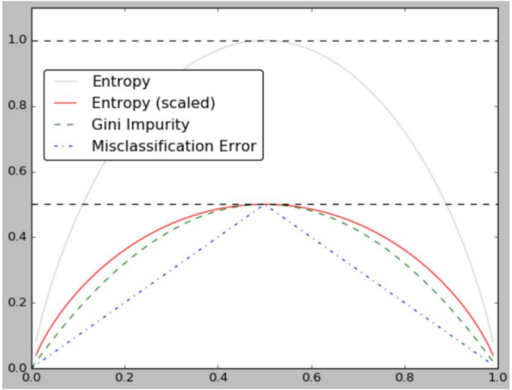
- Information theoretic measure of randomness
- Minimum number of bits to transmit a message — [Shannon]
- n data items
  - n_0 with $c = 0$, $p_0 = n_0/n$
  - n_1 with $c = 1$, $p_1 = n_1/n$
- Entropy
  $E = -(p_0 \log_2 p_0 + p_1 \log_2 p_1)$
- Minimum when $p_0 = 1, p_1 = 0$ or vice versa — note, declare $0 \log_2 0$ to be 0
- Maximum when $p_0 = p_1 = 0.5$



$c = 1$

Gini Index

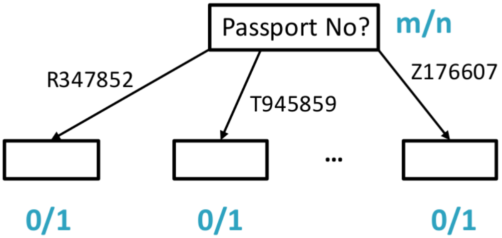
- Measure of unequal distribution of wealth
- Economics — [Corrado Gini]
- As before, n data items
  - n_0 with $c = 0$, $p_0 = n_0/n$
  - n_1 with $c = 1$, $p_1 = n_1/n$
- Gini Index $G = 1 - (p_0^2 + p_1^2)$
- $G = 0$ when $p_0 = 0$, $p_1 = 0$ or v.v.
  $G = 0.5$ when $p_0 = p_1 = 0.5$
- Entropy curve is slightly steeper, but Gini index is easier to compute
- Decision tree libraries usually use Gini index



$c = 1$

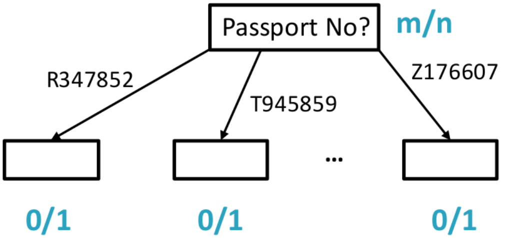
Information gain

- Greedy strategy: choose attribute to maximize reduction in impurity — maximize information gain
- Suppose an attribute is a unique identifier
  - Roll number, passport number, Aadhaar ...
- Querying this attribute produces partitions of size 1
  - Each partition guaranteed to be pure
  - New impurity is zero
- Maximum possible impurity reduction, but useless!



Information gain

- Tree building algorithm blindly picks attribute that maximizes information gain
- Need a correction to penalize attributes with highly scattered attributes
- Extend the notion of impurity to attributes



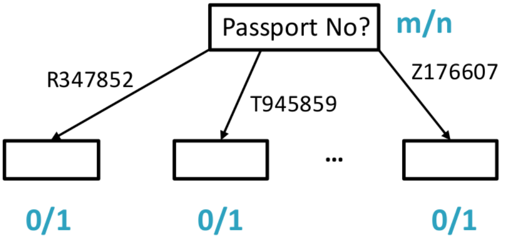
Attribute Impurity

- Attribute takes values $\{v_1, v_2, \ldots, v_k\}$
- v_i appears n_i times across n rows
- $p_i = n_i / n$
- Entropy across k values

$$-\sum_{i=1}^{k} p_i \log_2 p_i$$

- Gini index across k values

$$1 - \sum_{i=1}^{k} p_i^2$$



Attribute Impurity

- Extreme case, each $p_i = 1/n$
- Entropy

$$-\sum_{i=1}^{n} \frac{1}{n} \log_2 \frac{1}{n} = -n \cdot \frac{1}{n}(-\log_2 n) = \log_2 n$$

- Gini index

$$1 - \sum_{i=1}^{n} \left(\frac{1}{n}\right)^2 = 1 - \frac{n}{n^2} = \frac{n-1}{n}$$

- Both increase as n increases

Penalizing scattered attributes

- Divide information gain by attribute impurity
- Information gain ratio(A)

$$\frac{\text{Information-Gain(A)}}{\text{Impurity}(A)}$$

- Scattered attributes have high denominator, counteracting high numerator

Summary

- Can find better measures of impurity than misclassification rate
  - Non linear impurity function works better in practice
  - Entropy, Gini index
  - Gini index is used in most decision tree libraries
- Blindly using information gain can be problematic
  - Attributes that are unique identifiers for rows produces maximum information gain, with little utility
  - Divide information gain by impurity of attribute
  - Information gain ratio