Bookdown

Contents

Table 1: Table continues below

Kernel	Formula
Biweight	For $\left\|d\right\| < m$, kernel weight is $d(1 - \frac{d^2}{m^2})^2$. Otherwise, 0 when $\left\|d\right\| > m$. The value $m = d$ for the k^{th} neighbor.
Rectangular	$Pr(Y = j) = \frac{1}{k}\sum_{i=1}^{k} I(y_i = c)$
Inverse	$Pr(Y = j) = \sum_{i=1}^{k} w(d)(y_i = j)$ where $w(d) = \frac{1}{d_i \sum_{i=1}^{k}(\frac{1}{d_i})}$
Gaussian	

Interpretation
Calculate the proportion of j based on k nearest neighbors. This is the same of simple arithmetic mean.
Calculate the weighted proportion of j based on the inverse distance to k nearest neighbors.

Table 3: Measures used for attribute tests

Measure	Formula	Types of Target	Interpretation
Gini Impurity	Gini Impurity = $\sum_{i=1}^{k} p_i(1 - p_i) = 1 - \sum_{i=1}^{k} p_i^2$	Discrete	Commonly use in CART. The Gini Impurity ranges from 0 to $1 - \frac{1}{k}$.
Entropy	Entropy = $\sum_{i=1}^{k} -p_i log_2(p_i)$	Discrete	Relied on in other decision tree learning algorithms. Values range from 0 to 1
Variance	$Variance = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n}$	Continuous	Focus on reducing the spread in the target within partitions.

These impurity measures greatly simplify the task of identifying useful information. If we think back to the downed trees, the modeling objective is to learn what distinguishes affected areas from unaffected areas. Suppose emergency responders will only be assigned to areas north of a single parallel, which greatly simplifies the targeting problem into one focused on partitioning. *Which degree of latitude do we choose?* In the example table below, two candidate splits have been identified: Split A is somewhere north and Split B is farther north. To evaluate which split is better, we calculate the IG for each split, applying Gini Impurity as I:

$$IG_A = I(D_{\text{region}}) - \frac{n_{\text{A,north}}}{N} I(D_{\text{north}}) - \frac{n_{\text{A,remainder}}}{N} I(D_{\text{A,remainder}})$$

Text goes here @ref(tab:ginitab)

Table 4: Common measures used for decision tree learning attribute tests.

Impurity	Candidate.A	Candidate.B
Regional	$I(D_{\text{regional}}) = 1 - (p^2_{\text{trees}} + p^2_{\text{untouched}})$ $= 1 - ((\frac{13}{39})^2 + (\frac{26}{39})^2$ $= 0.4\bar{4}$	$I(D_{\text{regional}}) = 1 - (p^2_{\text{trees}} + p^2_{\text{untouched}})$ $= 1 - ((\frac{13}{39})^2 + (\frac{26}{39})^2$ $= 0.4\bar{4}$
Northern	$I(D_{\text{A,north}}) = 1 - (p^2_{\text{trees}} + p^2_{\text{untouched}})$ $= 1 - ((\frac{10}{10+5})^2 + \frac{5}{10+5})^2$ $= 0.4\bar{4}$	$I(D_{\text{B,north+}}) = 1 - (p^2_{\text{trees}} + p^2_{\text{untouched}})$ $= 1 - ((\frac{9}{9+0})^2 + \frac{0}{9})^2$ $= 0$
Remainder	$I(D_{\text{A,remainder}}) = 1 - (p^2_{\text{trees}} + p^2_{\text{untouched}})$ $= 1 - ((\frac{3}{3+21})^2 + \frac{21}{21+3})^2$ $= 0.21875$	$I(D_{\text{B,remainder}}) = 1 - (p^2_{\text{trees}} + p^2_{\text{untouched}})$ $= 1 - ((\frac{4}{4+26})^2 + \frac{26}{26+4})^2$ $= 0.23\bar{1}$
Information Gain	$IG_A = 0.4\bar{4} - \frac{15}{39}0.4\bar{4} - \frac{24}{39}0.21875$ $= 0.13\bar{8}$	$IG_A = 0.4\bar{4} - \frac{9}{39}0 - \frac{30}{39}0.23\bar{1}$ $= 0.2\bar{6}$