AML - Home Work 03

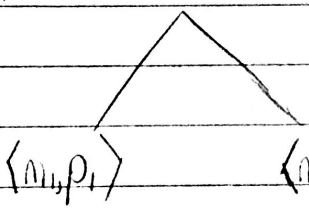
Written Exercises

(1)

a) $I(p) = \min(p, 1-p)$

$$(p_1 + n_1) \cdot I\left(\frac{p_1}{p_1 + n_1}\right) + (p_2 + n_2)\, I\left(\frac{p_2}{p_2 + n_2}\right)$$

Split

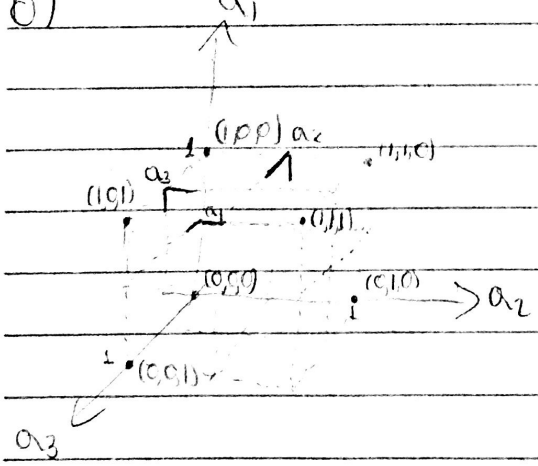


$(p_1+n_1) \cdot \text{minerr}\left(\frac{p_1}{p_1+n_1}, \frac{n_1}{p_1+n_1}\right) + (p_2+n_2) \cdot \text{minerr}\left(\frac{p_2}{p_2+n_2}, \frac{n_2}{p_2+n_2}\right)$

Assuming $n_1 > p_1$ and $n_2 > p_2$

$$\text{Imp} = (p_1 + n_1)\, I\left(\frac{p_1}{n_1+p_1}\right) + (p_2 + n_2)\, I\left(\frac{p_2}{p_2+n_2}\right)$$

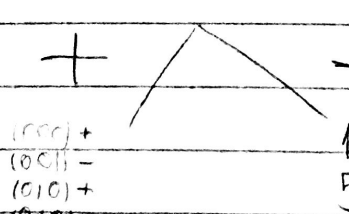
b)



3 possible splits
↳ along a_1 axis
↳ // a_2 //
↳ // a_3 //

Gini:

a_1

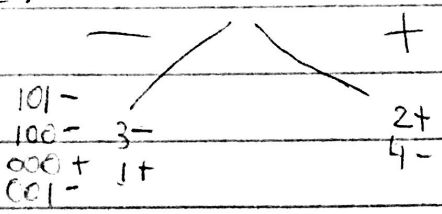


$$\text{Gini} \quad \sum_k p_{mk}(1 - p_{mk}) = \frac{2}{4}\left(1 - \frac{2}{4}\right) + \frac{5}{6}\left(1 - \frac{5}{6}\right)$$

$$= 0.25 + \frac{5}{36} = 0.388$$

$$\text{min error} : \frac{3}{10} = 0.3$$

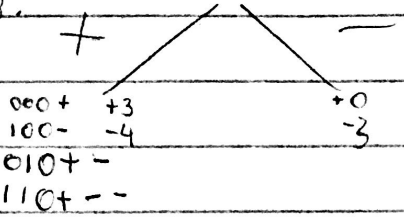
a_2:



$Gini = \frac{3}{4}\left(1-\frac{3}{4}\right) + \frac{4}{6}\left(1-\frac{4}{6}\right) = \frac{3}{16} + \frac{8}{36} = 0.409$

$minerror: \frac{5}{10} = 0.5$

a_3:



$Gini: \frac{4}{7}\left(1-\frac{4}{7}\right) + \frac{3}{2}\left(1-\frac{3}{2}\right) = \frac{12}{49} = 0.241$ (the $\frac{3}{2}$ term → 0)

$minerror: \frac{4}{10} = 0.4$

$min(Gini)$ is on split along the a_3 axis, therefore this is the selected split for gini impurity. Similarly, for minerror the selected split is along a_1, which minimizes minerror.

c) before: $me = min\left(\frac{p_1+p_2}{p_1+p_2+n_1+n_2}, \frac{n_1+n_2}{n_1+n_2+p_1+p_2}\right)$

After: $me_1 = min\left(\frac{p_1}{p_1+n_1}, \frac{n_1}{p_1+n_1}\right)$

$me_2 = min\left(\frac{p_2}{p_2+n_2}, \frac{n_2}{p_2+n_2}\right)$

If $me_1, me_2 < me$, then $\frac{p_1}{p_1+n_1}, \frac{p_2}{p_2+n_2} < \frac{p_1+p_2}{p_1+p_2+n_1+n_2}$

or

$\frac{n_1}{p_1+n_1}, \frac{n_2}{p_2+n_2} < \frac{n_1+n_2}{p_1+p_2+n_1+n_2}$

/ /

d) The answers to "b" and "c" indicate that the minimum error might not be the most suitable impurity measure for growing decision trees because it creates more complex decision models than gini or cross-entropy.

<Written Exercises>

2. Bootstrap aggregation ("bagging")

- Expected fraction of the training set

$$= 1 - (1 - \frac{1}{n})^n \approx 1 - e^{-1} = \underline{0.632}$$

- Limit of the expectation as $N \to \infty$?

$$\lim_{n \to \infty} (1 - (1 - \frac{1}{n})^n)$$

$$= \lim_{n \to \infty} 1 - \lim_{n \to \infty} (1 - \frac{1}{n})^n$$

$$= 1 - \frac{1}{e}$$

$$= \underline{0.632}$$

* $1 - \frac{1}{n}$ is probability for not being selected at a specific drawing.

Thus, bootstrap will contain 63.2% of unique training set and the rest are replicate.